Search for vector-like leptons at CMS and prospects at the HL-LHC

विद्या वाचस्पति की
उपाधि की अपेक्षाओं की आंशिक पूर्ति में प्रस्तुत शोध प्रबंध

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

द्वारा / By

अर्णब लाहा / Arnab Laha

पंजीकरण सं. / Registration No.: 20183622

शोध प्रबंध पर्यवेक्षक / Thesis Supervisor:

डॉ. सौरभ दूबे / Dr. Sourabh Dube

भारतीय विज्ञान शिक्षा एवं अनुसंधान संस्थान पुणे

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH PUNE

2025

Certificate

Certified that the work incorporated in the thesis entitled Search for vector-like leptons at CMS and prospects at the HL-LHC submitted by **Arnab Laha** 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: May 6, 2025

(Dr. Sourabh DUBE)
Supervisor

Declaration

I declare that this written submission represents my research work in my own words and where others' ideas or works 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, fabricated, or falsified any idea/data/fact/source in my submission. I understand that the violation of the above will be a 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: May 6, 2025

(Arnab Laha)

Reg. No: 20183622

Abstract

The standard model of particle physics (SM) describes fundamental particles and their interactions. The predictions of the SM have been extensively tested in the experiments. However, it does not explain key phenomena such as the presence of dark matter, small non-zero neutrino masses, or gravity, suggesting the existence of beyond standard model (BSM) physics. Identifying rare BSM signals from the SM backgrounds with the increasing complexity arising from high-granularity detector information and growing volume of data taken or proposed (High luminosity LHC) by the current experiments sets a unique challenge. Advanced techniques such as Machine learning (ML) algorithms are well-suited for this task, as they can efficiently process large datasets and uncover hidden patterns in high-dimensional data.

Even if the LHC experiments excluded a large BSM parameter space at the TeV scale, minimal extension to SM, such as vector-like leptons (which appear in SUSY, GUT, etc) at the electroweak scale can evade detection primarily due to the overlapping signatures with the dominant SM processes (control region in usual searches), or the analysis choices that are tuned to probe massive particles. A targeted search for vector-like leptons at the electroweak scale is conducted in the final state with one muon and at least two jets using proton-proton collision data of integrated luminosity of 138 fb^{-1} collected by the CMS experiment at the LHC from 2016–2018 at center-of-mass energy 13 TeV. Deep neural network (DNN) based classifiers are used in a unique combination to suppress each background. No significant deviations from the SM expectations were observed. While this search did not have enough sensitivity to exclude phase space for the considered models, it exercised the power of good analysis to improve the signal-to-background significantly.

The heavy fermion search program will benefit from the detector enhancements and the increased center-of-mass energy ($\sqrt{s} = 14$ TeV) with unprecedented integrated luminosity of 3000 fb^{-1} at HL-LHC. The discovery potential of vector-like leptons coupling to first-, second-, and third-generation SM leptons is reported using multiple leptons in the final state for a high-luminosity LHC scenario (HL-LHC). Detector granularity and increased luminosity expected at HL-LHC require a significantly faster yet accurate event simulation pipeline to be developed to exploit the full potential of the HL-LHC physics reach. Variational Auto Encoder and Generative Adversarial Network-based generative models with dimensionality reduction techniques are discussed to build such a faster simulation chain and as an aid to understand the performance of deep learning networks.

Acknowledgments

Many people, both actively and behind the scenes, have helped me reach this point. A few pages could never capture my gratitude to them. Nonetheless, here is my attempt to Taylor-expand my time here– and for once, I think I need to include plenty of higher-order terms to be close to my feelings for them!

First and foremost, I would like to thank my supervisor, Prof. Sourabh Dube, for his continuous support and encouragement. Sourabh, you are an incredible person. I have learned innumerable things from observing you and discussing everything with you. I enjoy our conversations. How often do we meet people with a thoughtfully different and unique perspective on various things of shared interest? I am thankful that I met you. We may disagree on many topics and agree on more–I appreciate the openness of our exchanges. Your kindness, positive outlook, passion for the work, and clarity of thought inspire me. I feel very fortunate to have worked with you for my thesis and learn about the field with a little prior knowledge. Thank you for giving me the liberty to explore other topics that excite me from time to time. Not a single day working with you was ever boring, and it will never be.

I want to thank my research advisory committee, Dr. Arun Thalapillil and Dr. Halil Saka, for their valuable insights and suggestions throughout the course of my doctoral studies. Halil has also been a collaborator, and I have learned many things from our multilepton meetings and discussions during our work on the review paper. I thank Prof. Seema Sharma for her insightful conversations and questions on various EHEP topics whenever we chatted. Special thanks to Dr. Sam Bein for debugging the frustrating exception errors in the FastSim tasks with me in the online meeting. A big thank you to the staff of the physics and academic offices for ensuring a smooth and hassle-free experience in all administrative matters. I feel lucky to be a part of the IISER Pune physics department.

I am also thankful to all my past and present lab members: Angira, Anshul, Shubhanshu, Vipul, Steenu, Chitrakshee, Prachurjya, Yash, Riya, Kiruteeka, Bhumika, Aparna, Parijat, Soumya, Shreyas, Vaidehi, Shriyansh, Alpana, Uttsavi, Maer, Shailee, and Raj. We have shared this journey at various stages, and I am grateful to each of you for spending a part of your life with me. I will remember the sheer passion and enthusiasm during last year's Science Day preparations. In fact, I am convinced that moments like these inspire me more deeply to pursue science than thinking about the Higgs or vector-like leptons. A special thanks goes to Angira, whom I worked with during my early years here at IISER, and she

patiently guided me through numerous CMessy-things. I am sure that you will go on to achieve many great things in the future, Angira! I also consider myself lucky to have had the opportunity to mentor so many fantastic guys during my time here. I enjoyed every conversation with you, Chitrakshee: from our GAN-things, clustering hits to the 2 AM memes. I feel very fortunate to work with Aparna for her master's thesis on generative networks. All the best to each one of you guys. I could write another thesis describing all of you, so I will stop myself here!

I will forever cherish the time I spent with the Més que un friend group at IISER– Ratheejit, Tamaghna, Abhijit, Dipta, and Bhutu bhai (Sagnik). We have had so many interesting discussions: from our research topics to the political landscape. I will always remember our time at Yeble, the monsoon bike trips to the Western Ghats, and our support for our beloved teams, Argentina and Barcelona. Those sleepless nights during the Copa America and the unforgettable FIFA World Cup were truly special–I feel blessed to have experienced them with you all. They were crucial for me to appreciate life more than the crab jobs or limits. I am also grateful to Abhishek, Keerti, Sandipan, Panda, Pahan, Sumit, Pratim, Pradyut, Baccha, Subhayan bro, Rahul Mama, Krishnendu da, Prajjal da, Aslam da, and Joyeeta di for the wonderful times we spent in occasional cooking sessions and celebrating the small joys of life. I have been lucky to have amazing friends from my master's days– Amarnath, Sayantan, Soumen, Ranita, Angira, and Shampita. Though we have only met a few times since our days at HCU, you have always been in my thoughts and have supported me throughout this long journey. I thank my friends–Swagata, Olivia, Roney, Nirmalya, and Aniruddha– for listening to the occasional academic rant and offering their support.

Finally, I would like to thank my parents, Madhumita Laha and Kushal Laha, for their unconditional love and care. No words can truly describe their sacrifice for my well-being and education. Stepping out of the comfort zone was not easy, but I feel incredibly fortunate to have had the freedom to make my own choices at every stage of my career and pursue my dream of becoming a scientist. In hindsight, I thank myself for choosing to do a PhD. I often wish the principle of least action could be applied to life and planning. It would be a lie to say there were no frustrating moments, but the experience has profoundly shaped my perspective. It feels more special to pursue it in a field I enjoy.

Dedicated to my parents...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

What we observe is not nature in itself but nature exposed to our method of questioning.

— Werner Heisenberg

One of humanity's most profound pursuits is the attempt to understand the workings of nature. In some sense, nature is an overwhelming stage where phenomena of different length scales take place. From the cosmic evolution of the universe–the formation of galaxies, stars, and planetary systems– to the microscopic realm where atoms emerge from fundamental constituents–our method of questioning has revealed nature at different scales; what it is made of, and how they interact to form the universe as we observe it today. From telescopes and microscopes to powerful modern accelerators, the availability of tools is crucial for examining nature at different scales. As a result, our understanding of the fundamental building blocks of nature has evolved significantly, from atoms to quarks and leptons.

Our understanding of nature at the smallest scale is encoded in the language of quantum field theory called the standard model (SM). The SM describes fundamental particles and their interactions. The predictions of the SM have been extensively tested in experiments, the latest being the discovery of the Higgs boson in 2012 by the CMS and ATLAS collaborations. However, the SM does not explain key phenomena such as dark matter, neutrino masses, or gravity, suggesting the existence of physics beyond the standard model (BSM). There are models beyond the SM, such as supersymmetry, extra dimensions, and vector-like fermions, that attempt to explain these shortcomings.

Our tools have evolved from the tabletop experiments of emulsion plates to identify particles to collide high-energy particles in a massive accelerator facility to examine the signature of the particles produced in such collisions. The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator at CERN near Geneva, Switzerland. Two counter-rotating proton beams traveling at the speed of light are made to collide head-on with each other at four different points around the machine. At these collision points, four different particle detectors - CMS, ATLAS, ALICE, and LHCb- are situated to detect the particles created in such high-energy collisions. Identifying rare BSM signals

from the SM backgrounds with the increasing complexity and growing volume of data taken or proposed (High luminosity LHC or HL-LHC) by the current experiments sets a unique challenge. Advanced techniques such as Machine learning (ML) algorithms are well-suited for this task, as they can efficiently process large datasets and uncover hidden patterns in high-dimensional data.

This thesis focuses on the following key areas to tackle the challenges mentioned above,

- Search for vector-like leptons (VLLs) using data collected by CMS between 2016–2018 and the discovery potential of such heavy fermions at HL-LHC.
- Improving lepton identification using supervised and unsupervised ML algorithms.
- Integrating the phase-2 tracker geometry into the CMS Fast Simulation software and developing fast simulation techniques for the HL-LHC.

Vector-like leptons could address the Higgs naturalness problem, the muon g-2 anomaly, and be a suitable candidate for dark matter. A brief description of the standard model and physics beyond is given in Chapter 2 with a detailed description of vector-like leptons. The details of the LHC and the CMS detector used to search for such BSM particles are described in Chapter 3.

Earlier vector-like lepton searches using multiple leptons in the final state ruled out a large model parameter space. Still, some model scenarios, especially singlet VLLs, remain weakly constrained due to their decay topology favoring low charged lepton yield. This thesis presents an analysis with one muon and at least two jets in the final state to enhance sensitivity to singlet VLLs at the electroweak scale. This is the first CMS search that probed the VLLs coupling to second-generation SM leptons. The reconstruction and identification techniques of the objects used in this search are described in Chapter 4. ML-based identification techniques and their performance in data to distinguish prompt leptons (coming from W/Z/H/τ_h or VLLs) from non-prompt leptons originating from the semileptonic b-decays are also described.

An overview of the VLL search strategy, different SM backgrounds, preliminary event selections with background estimation and validation are described in Chapter 5.

Deep neural network (DNN) based classifiers discussed in Chapter 6 are used in a unique combination to suppress each background and enhance signal sensitivity. Unfortunately, no significant deviations from the SM expectations were observed. While this search did not have enough sensitivity to exclude phase space for the considered models, it exercised the power of good analysis to improve the signal-to-background significantly. The result of this search is presented in Chapter 7.

The heavy fermion search program will benefit from the detector enhancements and the increased center-of-mass energy (14 TeV) with unprecedented integrated luminosity (3000

fb^{-1}) at HL-LHC. Model-independent signal regions of a published CMS analysis that performed a search for new BSM phenomena in multilepton final states are utilized to extrapolate sensitivities for these three generations of vector-like lepton models (both singlet and doublet scenarios) at the HL-LHC scenario. This projection study is discussed in Chapter 8.

On the other hand, a significantly faster, accurate, and efficient event simulation pipeline needs to be developed with the increased luminosity and detector granularity expected at HL-LHC. They are crucial to increasing the discovery potential of our physics searches. We will discuss a few deep learning based generative techniques to establish a faster event simulation chain in Chapter 9. We will also delve into a few dimension reduction techniques to discover the hidden patterns in multidimensional data and as an aid to understand the performance of deep learning networks.

Chapter 2

The Standard Model and Beyond

The standard model is the fundamental theory of nature that describes all known fundamental particles and their interactions except gravity. In this chapter, the standard model is described in a top-down approach, not from a historical perspective and development of the working principles, but rather the different pieces of matter particles and their interactions that make up the SM. The description follows Ref. [1, 2, 3]. Various limitations of the SM are discussed. Next, models such as vector-like leptons are introduced that can potentially solve some of the outstanding questions of nature. An experimental overview and the motivation to search for these particles are discussed at the end.

2.1 The standard model of particle physics

The standard model is a mathematical framework based on quantum field theory that describes all known fundamental particles and their interactions. A fundamental particle is not made up of anything smaller- it is considered a building block of nature. The SM is a gauge theory based on $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ symmetry group, where each symmetry corresponds to one of the three fundamental interactions included in the theory: the strong interaction (governed by quantum chromodynamics), and the electroweak interaction, which unifies the weak and electromagnetic forces. Gravity, the fourth fundamental interaction, is not described within SM, and including gravity in the same framework is a very active area of particle physics research.

The particle content of the SM could be broadly classified into two categories: fermions, which constitute matter, and bosons, which mediate the fundamental forces. Fermions include six quarks and six leptons, organized into three generations of increasing mass. The quark family is composed of the "up type" up (u), charm (c) and top (t) with $+\frac{2}{3}$ electromagnetic charge and "down type" down (d), strange (s), and bottom (b) quark with $-\frac{1}{3}$ electromagnetic charge. These quarks participate in the strong, weak, and electromagnetic interactions. Similarly, the leptons are also composed of charged leptons, electron (e), muon

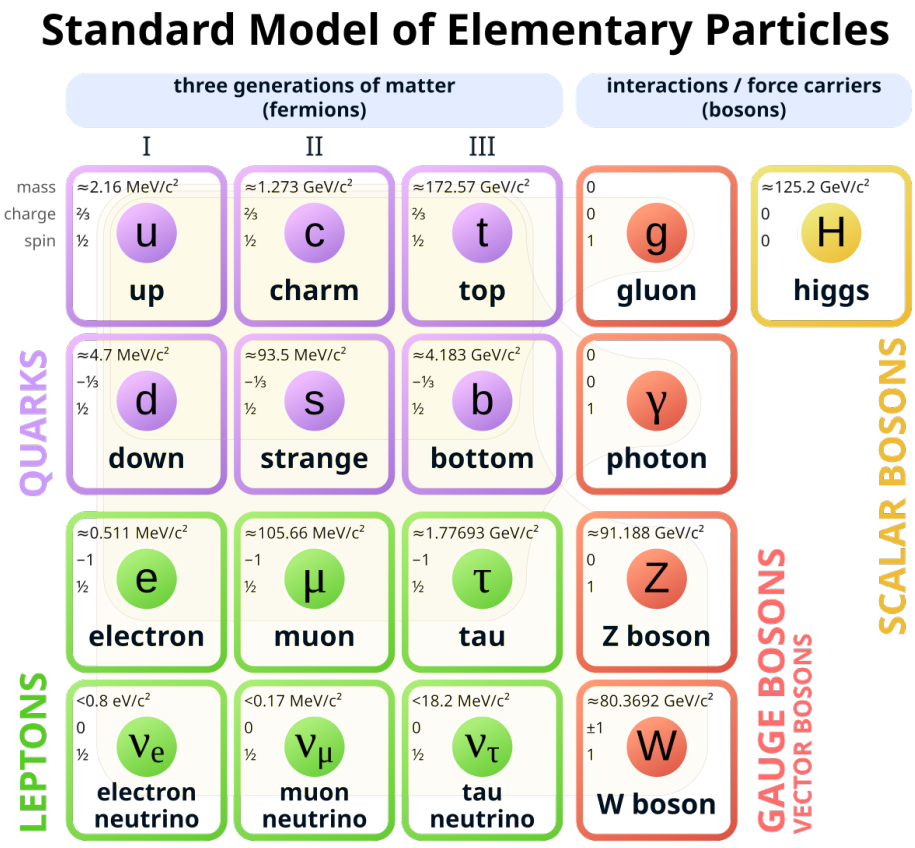


FIGURE 2.1: The standard model of particle physics. (Image courtesy: Wikipedia)

(μ), and tau lepton (τ) having 1 unit of negative electromagnetic charge and neutral leptons, electron-neutrino (ν_e), muon-neutrino (ν_μ), and tau-neutrino (ν_τ). The charged leptons participate in the weak and electromagnetic interactions, but the neutral leptons only feel the weak force and do not participate in the electromagnetic interactions, as they have no charge. Both the charged and neutral leptons do not take part in strong interactions. Each fermion has a corresponding antiparticle in the SM. The bosons are the force carriers of such interactions. For example, photon (γ) mediates the electromagnetic interaction, $W^{\pm}$ and Z boson mediate the weak interaction, and the strong interaction is mediated via gluons. The final and important piece of the model is a scalar particle, the Higgs boson, which arises as a consequence of the spontaneous symmetry breaking of the electroweak theory (at higher energy, electromagnetism and weak interactions are indistinguishable). The Higgs field is responsible for the masses of elementary particles, as we see later in this chapter. In other words, particles interact with the Higgs field in proportion to their mass. The existence of the Higgs boson was confirmed experimentally at the LHC in 2012. A summary of all the SM particles is shown in Figure 2.1.

The interactions between the particles are key aspects to understand nature. The particles can transform into other particles through interactions. Some interactions, like strong or electromagnetic, can not change the flavor of the interacting particles. However, it is a key feature of the weak interaction that allows particles to convert into other types of the same family, playing a vital role in particle decays. In quantum field theory, the spin-$\frac{1}{2}$ fermions are described by the Dirac equation.

$$(i\gamma^\mu \partial_\mu - m)\psi = 0 \tag{2.1}$$

where ψ is a four-component Dirac spinor, m is the mass of the particle, and γ^μ are the Dirac matrices. Here, the fermion field is not interacting with any gauge fields. The corresponding Lagrangian density of this free fermion field is given by,

$$\mathcal{L} = i\bar{\psi}\gamma^\mu \partial_\mu \psi - m\bar{\psi}\psi \tag{2.2}$$

where the field $\bar{\psi}$, is the conjugate field of ψ following $\bar{\psi} = \psi^\dagger \gamma^0$. In the next few sections, we will discuss how the fundamental interactions are incorporated in the SM through a simple constraint: the Lagrangian must be invariant under the local gauge transformation.

2.1.1 Electromagnetic interaction

The particles carrying the electric charge have electromagnetic interaction via an exchange of photons. In the SM framework, the electromagnetic interactions are described by an

abelian gauge theory based on the symmetry group $U(1)_{em}$. This theory is known as Quantum Electrodynamics (QED). The Lagrangian density in the Equation. 2.2 remains invariant under a global $U(1)_{em}$ phase transformations of the fermion field:

$$\psi \rightarrow e^{i\alpha}\psi \tag{2.3}$$

where α is a constant phase. The invariance implies that the equation of motion that describes the fermions wouldn't change as long as the α is any constant and independent of the space-time co-ordinates. This symmetry corresponds to the conservation of electric charge via Noether's theorem. The local gauge transformation where α is a function of the space-time coordinates is given by:

$$\psi \rightarrow e^{i\alpha(x)}\psi \tag{2.4}$$

The Lagrangian is not invariant under this transformation as the partial derivative of $\alpha(x)$ with respect to the space-time coordinates in the kinetic term survives and breaks the invariance. However, suppose we now introduce "local U (1) symmetry" as a requirement. Is it possible to modify the Lagrangian so that it obeys this symmetry? The answer is "yes", provided we introduce a new field, the gauge field. The theory can be made local gauge invariant by introducing a covariant derivative that takes into account a new vector gauge field (A_μ),

$$D_\mu = \partial_\mu - ieA_\mu \tag{2.5}$$

where A_μ transforms as,

$$A_\mu \rightarrow A_\mu + \frac{1}{e}\partial_\mu\alpha(x) \tag{2.6}$$

Invariance of the Lagrangian is then achieved by replacing ∂_μ by D_μ:

$$\begin{aligned} \mathcal{L} &= i\bar{\psi}\gamma^\mu D_\mu\psi - m\bar{\psi}\psi \\ &= \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi + e\bar{\psi}\gamma^\mu\psi A_\mu \\ &= \mathcal{L}_{free} + \mathcal{L}_{int} \end{aligned} \tag{2.7}$$

The interaction term ($\mathcal{L}_{int}$) describes the interaction between the fermion and the gauge field, A_μ. This is the exact form of the electromagnetic interaction, where the interaction term can be written in terms of the current (J^μ), e is the EM charge, and A_μ is the photon field. Now, the associated kinetic term of the photon field must be added to the Lagrangian to get the QED Lagrangian.

$$\begin{aligned} \mathcal{L}_{QED} &= \mathcal{L}_{free} + \mathcal{L}_{int} + \mathcal{L}^A_{free} \\ &= \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi + e\bar{\psi}\gamma^\mu\psi A_\mu - \tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} \end{aligned} \tag{2.8}$$

where $F_{\mu\nu}$ can only take the following form to respect the gauge invariance,

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{2.9}$$

One important aspect is that if we add any mass term corresponding to the gauge field (like the fermion field, ψ) to the Lagrangian as:

$$\mathcal{L}^{\gamma}_{mass} = \frac{1}{2} m^2_\gamma A^\mu A_\mu \tag{2.10}$$

This violates the gauge invariance of the Lagrangian. This means that the photon of the associated gauge field must be massless to respect the local phase invariance of the Lagrangian. Thus, the simple requirement of local gauge invariance (here a local U(1) phase) on the free fermion Lagrangian led to the interacting field theory of QED. The reason that local gauge symmetries are so important is because of what is called "renormalizability". Lagrangians with interaction terms generated by local gauge symmetries are renormalizable. In other words, if we want to have a theory in which we can compute something, then we cannot have any other interactions than those derived from internal symmetries.

One important thing to note is that the family of phase transformation $U(\alpha) \equiv e^{i\alpha}$ forms a unitary Abelian group, where the group multiplication is commutative:

$$U(\alpha_1)U(\alpha_2) = U(\alpha_2)U(\alpha_1) \tag{2.11}$$

Analogously, we can hope to infer the structure of weak interactions and QCD interactions by imposing local gauge invariance. However, we need to consider the mass of the gauge vector bosons, as we will see in the next few sections. Let's see how the local gauge invariance imposed on a non-Abelian group gives rise to quantum chromodynamics.

2.1.2 Non-Abelian gauge invariance and strong interaction

Similar to QED, local gauge invariance can be applied, but with the U(1) gauge group replaced by the SU(3) group of phase transformation on the free quark color field. The free Lagrangian is:

$$\mathcal{L}^{q}_{free} = \bar{\psi}_{q_j}(i\gamma^\mu \partial_\mu - m)\psi_{q_j} \tag{2.12}$$

where the ψ_{q_j} denotes the three color fields for $j = 1, 2,$ and 3, respectively. Each quark has a color charge under SU(3), labeled as red, green, and blue. The Lagrangian is invariant under the global phase transformation, but not under the local gauge transformations:

$$\psi_q \rightarrow U\psi_q \equiv e^{i\alpha_a(x)T_a}\psi_q \tag{2.13}$$

where T_a are the 8 generators of the SU(3) group with $a = 1, 2, 3.., 8$ and can be written as 8 linearly independent traceless 3×3 Gell-Mann matrices ($\lambda_a/2$). Similarly, we need to have eight vector gauge fields (G^a_μ) to write down the covariant derivative as:

$$D_\mu = \partial_\mu + igT_aG^a_\mu \tag{2.14}$$

where the vector gauge fields must transform as:

$$G^a_\mu \to G^a_\mu - \frac{1}{g}\partial_\mu\alpha_a - f_{abc}\alpha_b G^c_\mu \tag{2.15}$$

The last term is needed to cancel the extra terms that arise from the non-commutativity of the generators. f_{abc} are called the structure constants of the group. The final QCD Lagrangian can be written, including the kinetic term of the free gauge fields, as:

$$\mathcal{L}_{QCD} = \bar{\psi}_{q_j}(i\gamma^\mu\partial_\mu - m)\psi_{q_j} - g(\bar{\psi}_{q_j}\gamma^\mu T_a\psi_{q_j})G^a_\mu - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a \tag{2.16}$$

where the field tensor, $G^a_{\mu\nu}$ is:

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - gf_{abc}G^b_\mu G^c_\nu \tag{2.17}$$

As for the photon, local gauge invariance requires the gluons to be massless. Imposing the gauge symmetry leads to the form of the field strength tensor, $G^a_{\mu\nu}$. Unlike the photon field, the gluon fields have self-interactions due to the last term. It reflects the fact that gluons themselves carry color charge. This arises due to the non-Abelian character of the gauge group SU(3), and the gauge symmetry uniquely determines the structure of these gluon self-coupling terms. There is only one strong coupling as denoted by g.

2.1.3 Electroweak unification and symmetry breaking

The next task is: can the weak interactions be described similarly to QED or QCD with one or more mediating fields playing the role of the photons? In the early 1900s, many physical processes had been observed that helped to formulate a proper description of a consistent theory of the weak interactions. One typical interaction was the β-decay: $n \to p+ \nu_e + \bar{\nu_e}$. Fermi proposed the famous Fermi 4-point interaction with an effective Lagrangian description:

$$\mathcal{L}_{fermi} = G_F\psi_1\psi_2\psi_3\psi_4 \tag{2.18}$$

where the ψ_i represent each of the particles in the interaction and G_F determines the coupling. This description is only good at energies $E << G_F^{-1/2}$, but the theory becomes

non-renormalizable at higher energies. Given the absence of an underlying theory at the time, one could write the most general Lorentz invariant 4-fermion interactions, with the current J_i in form of an operator $\mathcal{O}^i$:

$$J_i = \bar{\psi}\mathcal{O}^i\psi,\ \mathcal{L} = \sum_i g_i J^i J_i,\ g_i = G_F \tag{2.19}$$

The operator could be in the form of a vector with γ^μ, an axial vector ($\gamma^\mu\gamma^5$), a tensor ($\gamma^{\mu\nu}$), a pseudo-scalar (γ^5), or a scalar (1). In the 1950s, Marshak and Sudarshan identified the correct combination that describes all the weak interaction processes as V-A from the results of many experiments that led to the conclusion that weak interaction violates parity maximally. This included interactions, e.g., for the $\beta-$decay

$$g\bar{\psi}_p\gamma^\mu(1-\gamma^5)\psi_n\bar{\psi}_e\gamma^\mu(1-\gamma^5)\psi_\nu + h.c.. \tag{2.20}$$

The V-A theory describes an important observation of the weak interactions, namely chirality (parity violation). To appreciate this important point, we have to define the chirality operator that acts on a Dirac spinor:

$$P_L = \frac{1}{2}(1-\gamma^5), P_R = \frac{1}{2}(1+\gamma^5); P_L\psi = \psi_L; P_R\psi = \psi_R \tag{2.21}$$

Naturally, these operators are called chirality operators as they project a spinor field ψ onto its left-handed and right-handed components. Chirality refers to the handedness of a particle's wavefunction. A left-handed spinor transforms differently under the Lorentz group than a right-handed one. It is easier to understand for massless fermions, where the chirality coincides with helicity: left-handed means spin opposite to momentum, and right-handed means spin along momentum. However, for massive fermions, chirality and helicity are distinct. Chirality is a Lorentz-invariant property, while helicity is not. The direction of a massive particle would change if we choose another reference frame; hence, its helicity would also change, but it is a Lorentz invariant quantity for massless particles as they travel at the speed of light (maximum allowed speed in vacuum). As a result, only left-handed neutrinos (ν_L) and right-handed antineutrinos ($\bar{\nu}_R$) are involved in weak interactions. Looking at the Equation 2.20, we can see the presence of only $1-\gamma^5$ on the operators. This difference between left- and right-handed fermions is a very important property of the weak interactions that comes from observation. Therefore, from the dependence on $1-\gamma^5$, it is usually said that weak interactions are left-handed and so chiral. For this reason, the standard model is also called a chiral theory, although QED or QCD cannot differentiate between left and right-handed particles. This was the first successful description of weak interactions at low energies. The theory gives diverging results for higher energies, which are clearly against

experiments. This suggested that the four-fermion vertex with dimensionful coupling G_F vertex should be replaced by a three-point interaction and propagator for mediator particles as in QED.

With this discussion, we can write the first part of the Dirac equation as:

$$\bar{\psi}\gamma^{\mu}\psi = \bar{\psi}_L\gamma^{\mu}\psi_L + \bar{\psi}_R\gamma^{\mu}\psi_R \tag{2.22}$$

where the chiral projections of the adjoint spinors are given by:

$$\bar{\psi}_L = \frac{1}{2}\bar{\psi}(1+\gamma^5), \bar{\psi}_R = \frac{1}{2}\bar{\psi}(1-\gamma^5) \tag{2.23}$$

We can rewrite the Dirac Lagrangian in terms of the left- and right-handed spinor projections as:

$$\mathcal{L}_{free} = \bar{\psi}_L(i\gamma^{\mu}\partial_{\mu})\psi_L + \bar{\psi}_R(i\gamma^{\mu}\partial_{\mu})\psi_R - \bar{\psi}_R m\psi_L - \bar{\psi}_L m\psi_R \tag{2.24}$$

The mass terms of the free fermion Lagrangian mix the left- and right-handed components. However, since left-handed and right-handed fermions transform differently under the electroweak gauge group, this mass term is not gauge invariant and thus is not allowed in the symmetric phase of the theory. Therefore, we consider only massless fields now and come back to the non-zero mass later.

To identify the gauge group, Lorentz invariance requires putting those fields with the same Lorentz transformation properties into a single representation. Hence, we split the fields into left- and right-handed content

$$\Psi_L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix}; e_R; \tag{2.25}$$

Here Ψ_L is called the weak isospin doublet and e_R is the singlet. Note that Ψ is not a Dirac spinor (ψ), but a doublet of Dirac spinors. We just reformulated the representation. Now we impose the SU(2) gauge symmetry on the left-handed doublets only. That is, we require that the Lagrangian be invariant for local rotations of the doublet. To do this, we need to ignore that the two components of a doublet have different charges, a problem that needs to be resolved later, along with the massless fermion field assumption.

Let's construct the weak $SU(2)_L$ theory imposing the $SU(2)_L$ symmetry on the weak isospin doublets. That is, we require that the Lagrangian be invariant for local phase transformation,

$$\mathcal{L}_{free} = \bar{\Psi}_L i\gamma^{\mu}\partial_{\mu}\Psi_L \tag{2.26}$$

Now we can follow the same procedure of non-Abelian gauge theory to make the Lagrangian invariant under local phase transformation,

$$\partial_\mu \rightarrow D_\mu = \partial_\mu + igw_\mu; w_\mu = T_a W_\mu^a \tag{2.27}$$

where W_μ^a are the three SU(2) gauge fields for a =1, 2, and 3, and the Lagrangian with the interaction term is:

$$\mathcal{L}_{free} \rightarrow \mathcal{L}_{free} - g\bar{\Psi}_L \gamma^\mu T_a \Psi_L W_\mu^a \tag{2.28}$$

where T_a are the group generators. They can be represented by the Pauli spin matrices, $\tau_1, \tau_2, \; and \; \tau_3$. The interaction term can be written as current (J_{weak}^μ), which can be further split into the charged-current and neutral current components. It can be seen that the generators τ_1 and τ_2 mix the components of the doublet due to their non-diagonal terms in the matrices. The gauge fields associated with the charged current part can be rewritten as, with a slight change in the basis ($\tau^\pm = \frac{1}{2}(\tau_1 \pm i\tau_2)$) as,

$$W^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2) \tag{2.29}$$

and the neutral current part,

$$\mathcal{L}_{NC} = -g\bar{\Psi}_L \gamma^\mu \frac{\tau_3}{2} \Psi W_\mu^3 \tag{2.30}$$

The third gauge field may look like a Z^0 boson, and we can add the U(1) term where Q is the generator. However, it is evident that the left-handed SU(2) doublets we constructed are not eigenfunctions of Q since they mix fields with different charges. Therefore, our $SU(2)_L$ invariant Lagrangian cannot be symmetric under a transformation with Q as a generator.

The idea is to start from another U(1) gauge symmetry called weak hypercharge. The generator of this group is Y, and we require that Y commutes with $SU(2)_L$ generators. The combined symmetry is denoted by $SU(2)_L \otimes U(1)_Y$. The Lagrangian can be written as,

$$\mathcal{L}_{EW} = \mathcal{L}_{free} - g\bar{\Psi}_L \gamma^\mu T_a \Psi_L W_\mu^a - \frac{g'}{2} \bar{\Psi}_L \gamma^\mu Y \Psi_L B_\mu \tag{2.31}$$

The transformations corresponding to T_3 and Y both lead to neutral current interactions. As a result, the gauge boson fields can mix.

But what about the masses of the fermions and these gauge fields we have ignored so far? Here comes the Higgs mechanism. Their masses and the masses of all fermions can be generated in a mechanism called spontaneous symmetry breaking, which involves a new scalar field, the Higgs field. It also mixes the two neutral fields associated with the T_3 and Y gauge fields to create the physical neutral fields.

To break the $SU(2)_L \otimes U(1)_Y$ symmetry, a complex scalar field is added which is also an isospin doublet,

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \tag{2.32}$$

The corresponding Lagrangian is:

$$\mathcal{L}_{scalar} = (D^\mu \phi)^\dagger (D_\mu \phi) - V(\phi) \tag{2.33}$$

where the D_μ is the covariant derivative that we discussed earlier as:

$$D\mu = \partial_\mu + ig\frac{1}{2}\tau_a W_\mu^a + ig'\frac{1}{2}Y B_\mu \tag{2.34}$$

With the potential:

$$V(\phi) = \mu^2(\phi^\dagger\phi) + \lambda(\phi^\dagger\phi)^2 \tag{2.35}$$

It is evident from the potential that it has a trivial minimum at $\phi = 0$ for $\mu^2 > 0$, while for $\mu^2 < 0$ the minimum of the potential is at:

$$\phi_{min} = \pm\sqrt{-\mu^2/2\lambda} \tag{2.36}$$

The gauge symmetry is broken spontaneously for the non-trivial VEV (v) $< \phi > \neq 0$.

$$SU(2)_L \otimes U(1)_Y \rightarrow U(1)_{em} \tag{2.37}$$

The unbroken $U(1)_{em}$ corresponds to electromagnetism, with the generator:

$$Q = T^3 + \frac{Y}{2} \tag{2.38}$$

By expanding the kinetic term $(D^\mu\phi)^\dagger(D_\mu\phi)$ around the VEV, mass terms for the weak gauge bosons naturally emerge, and also allow the mixing of the W_μ^3 and B_μ fields.

$$\begin{aligned} Z_\mu &= W_\mu^3 cos(\theta_W) - B_\mu sin(\theta_W) \\ A_\mu &= W_\mu^3 sin(\theta_W) + B_\mu cos(\theta_W) \\ cos(\theta_W) &= \tfrac{g}{\sqrt{g^2+g'^2}}; sin(\theta_W) = \tfrac{g'}{\sqrt{g^2+g'^2}} \end{aligned} \tag{2.39}$$

where θ_W is called the weak mixing angle or Weinberg angle. The mass terms could be written as:

$$\begin{aligned} m_h &= \sqrt{2\lambda v} \\ m_W &= \tfrac{gv}{2} \\ m_Z &= \tfrac{v}{2}\sqrt{g^2 + g'^2} \\ m_{A_\mu} &= 0 \\ m_W &= m_Z cos(\theta_W) \end{aligned} \tag{2.40}$$

Describing the physics of weak interactions leads not only to a consistent theory for the weak interactions but also to a theory that includes the electromagnetic interactions in a unified way. Both interactions, mediated either by A_μ giving rise to QED or by $W^{\pm}_{\mu}$,Z_μ giving rise to the weak interactions, come from the same underlying theory, a spontaneously broken $SU(2)_L \otimes U(1)_Y$ gauge theory.

The standard model gauge symmetry also forbids explicit mass terms for the fermionic degrees of freedom of the Lagrangian. The fermion mass terms are then generated via gauge-invariant, renormalizable Yukawa couplings to the scalar field. The Higgs mechanism allows for fermion masses through Yukawa interactions:

$$L_{Yukawa} = y_f \bar{\psi}_L \phi \psi_R + h.c \tag{2.41}$$

After the symmetry breaking, this gives,

$$m_f = \frac{y_f v}{\sqrt{2}} \tag{2.42}$$

connecting the mass of fermions directly to their coupling with the Higgs field.

2.1.4 Inadequacies of the standard model

SM is the most successful theory of fundamental interactions. Although SM can successfully describe fundamental physics phenomena, it is not a complete theory of nature. SM fails to explain various experimental observations. Some theoretical questions that have no answer so far. Some of these inadequacies of the SM are discussed here.

Neutrinos are massless in the SM framework. However, neutrino oscillation experiments [4, 5] have demonstrated that neutrinos change flavor as they propagate. This is possible if neutrinos are described by a superposition of mass eigenstates, enabling flavor transitions as they propagate. The SNO experiment first observed this conversion ($\nu_e \rightarrow \nu_\mu \; or \; \nu_\tau$) in the solar neutrinos, which can be explained if at least two species of neutrinos have finite non-zero mass [6, 7]. In the SM, there is no explanation for these finite non-zero masses.

One of the prominent indications of BSM physics is the existence of dark matter (DM) [8]. There is (are) no DM candidate(s) in the SM. There are various astrophysical and cosmological evidences that DM exists. Their presence can be inferred from the galaxy rotation curves [9], where the velocity of the stars is measured as a function of the radial distance from the galaxy center. As most of the mass ("luminous") is concentrated in the galaxy center, the stars should rotate slowly as the mass density falls outside a critical distance from the galaxy core. However, observing the higher velocity of these stars indicates that there must be some "invisible" mass that acts only gravitationally, as if the entire galaxy is sitting in this invisible matter halo. It is called dark matter, as they don't interact electromagnetically and only interact gravitationally, so we can't see them. There is other strong evidence in favor of dark matter, like gravitational lensing. Observations show that the lensing mass is significantly greater than the visible mass in stars and hot gas. Another profound cosmological evidence is the anisotropies in the Cosmic Microwave Background (CMB) spectrum. The CMB temperature fluctuations measured with high precision by satellites like WMAP and Planck reveal the primordial density perturbations of the early universe. The angular power spectrum of the CMB, particularly the heights and positions of the acoustic peaks, depends sensitively on the matter content of the universe. The data are best fit by a $\Lambda-$CDM model (Lambda Cold Dark Matter), which includes about 26% of the total energy density as dark matter [10], consistent with other cosmological probes. All this evidence suggests the presence of a non-luminous and non-baryonic form of matter that interacts predominantly through gravity. The standard model does not have any explanation for these observations.

There are a few "why" theoretical questions that are not necessarily a shortcoming, but have no first-principle answers. It includes the existence of additional fermion families, the number of families, and the mass hierarchies present in the three generations of families in the SM. Matter we know is made only of up and down quarks and the electron. Why are there two more families of identical particles, differing only in mass, with the first family (and decaying to them by different interactions)? On the other hand, there are only three families of fermions (in the flavor sector), no less or more. The large mass differences in the families, from top quark being 173 GeV to 0.5 MeV for the electron, and extremely light neutrino masses, there is no explanation of why they have to take the values they do.

The CMS and ATLAS collaborations have observed and measured the mass of the Higgs boson to be around 125 GeV, at the electroweak scale [11, 12]. The Higgs boson mass gets radiative corrections via loop diagrams involving virtual particles (such as fermions, gauge bosons, and the Higgs itself). In the SM, higher-order corrections to the Higgs-boson mass parameter square contain quadratic ultraviolet divergences with respect to the cutoff scale of the theory. If the cut-off scale (Λ) is extremely large (at the Planck scale), the physical Higgs mass should be large–unless there is an extremely fine-tuned cancellation between the

bare Higgs mass and the radiative corrections. This requirement for precise cancellations to achieve the observed Higgs mass of $\sim$125 GeV is described as the naturalness problem or the hierarchy problem in the SM.

2.2 Beyond standard model

There are various theories beyond the standard model (BSM) that attempt to solve the issues with the SM. For example, a well-known extension of the SM is Supersymmetry (SUSY). SUSY predicts the existence of superpartners for each SM particle. For all fermions, there are bosonic superpartners, and for each boson, there are fermionic superpartners. SUSY is a popular theory as it solves several issues at once. It can solve the fine-tuning of the Higgs mass as the radiative corrections get cancelled by the fermionic and bosonic degrees of freedom of the theory, and it can give a dark matter candidate (lightest supersymmetric particle) if a certain quantum number in SUSY is respected. It also unifies the three SM gauge couplings at a single Grand Unified Theory (GUT) scale. Since we didn't observe any superpartners of the SM particles at their corresponding mass, SUSY has to be broken at some energy scale. With some constraints on this cut-off scale, it can still solve the quadratic loop divergence of the Higgs mass and stabilize it at the electroweak scale. There are other models like seesaw, extra dimensions, additional vector-like fermions, and compositeness models. While seesaw could potentially explain the origin of extremely small neutrino mass, models with extra dimensions attempt to explain the apparent weakness of gravity by proposing the existence of additional spatial dimensions beyond the familiar 3+1 spacetime.

Theory with extra dimensions, grand unified theories (GUTs), SUSY, little Higgs models, and composite Higgs models often have additional vector-like fermions. One of their primary motivations is to help stabilize the Higgs boson mass by cancelling the large radiative corrections from the top quark loop, thus addressing the hierarchy or naturalness problem. In some models, vector-like quarks (VLQs) play a role similar to the top partners in supersymmetry or composite Higgs theories. Furthermore, because vector-like fermions are not tightly constrained by electroweak precision tests (as they are not chiral), they offer a phenomenologically viable way to introduce new heavy states that couple to SM particles. The leptonic sector of such vector-like fermions and the search strategy for such particles are described next.

2.2.1 Vector-like leptons

Vector-like fermions are new particles whose left- and right-handed components transform similarly under the standard model (SM) gauge symmetries, unlike SM fermions, which are

chiral (i.e., left- and right-handed components transform differently under $SU(2)_L \times U(1)_Y$). This "vector-like" nature allows them to obtain masses through gauge-invariant mass terms, independent of electroweak symmetry breaking, and hence they are less constrained by the Higgs boson properties. In particular, vector-like leptons (VLLs) are color-singlets produced via electroweak interactions at the LHC and yield rich signatures with multiple leptons or jets. Their production cross-section depends on their mass and electroweak charges but is generally lower than that of colored particles like vector-like quarks (VLQs). The VLLs may account for the mass hierarchy between the different generations of particles in the SM via their mixings with the SM leptons. However, the mixing is generally constrained by precision electroweak and flavor observables. The strength of the mixing impacts the branching ratios of their decays and can lead to lepton-flavor-violating signatures in certain models. Apart from the minimal extension to the SM framework, VLLs appear in a wide variety of models ranging from supersymmetry to extra dimensions and grand unification [13, 14, 15, 16, 17, 18].

The VLL models can be broadly classified into two categories, $\mathrm{SU}(2)$ doublets $\mathrm{L_i} = (\mathrm{E_i}, \mathrm{N_i})$ or singlets $\mathrm{E_i}$, where E and N denote the electrically charged and neutral states, respectively. The subscript i denotes the VLLs coupling to the first, second, and third generation SM leptons for $\mathrm{i} = 1, 2, 3$, respectively. First, second, or third generation VLLs are also termed as vector-like electrons, muons, or taus in literature. From an experimental perspective, the difference between these models is that the singlet VLL model allows only charged VLLs, while the doublet VLL model allows charged and neutral VLLs. For example, the singlet model adds only one charged vector-like electron (or muon or tau) and its antiparticle to the SM particle content. On the other hand, the doublet model also adds one neutral vector-like neutrino along with its antiparticle. Therefore, the singlet model has two new particles, whereas the doublet model has four.

2.2.2 Production and decay of VLLs

In the singlet VLL model, the production channel at the LHC proceeds through the s-channel Z, γ^* pair production as:

$$pp \to Z/\gamma^* \to \mathrm{E}\overline{\mathrm{E}} \tag{2.43}$$

Since the doublet VLL model has the heavy vector-like Dirac neutrino, in addition to the previous charged VLL pair production, the associated production modes of E and N are also possible, as illustrated in Figure 2.2.

$$\begin{aligned} pp &\to Z \to \mathrm{N}\overline{\mathrm{N}} \\ pp &\to W^- \to \mathrm{E}\overline{\mathrm{N}} \\ pp &\to W^+ \to \overline{\mathrm{E}}\mathrm{N} \end{aligned} \tag{2.44}$$

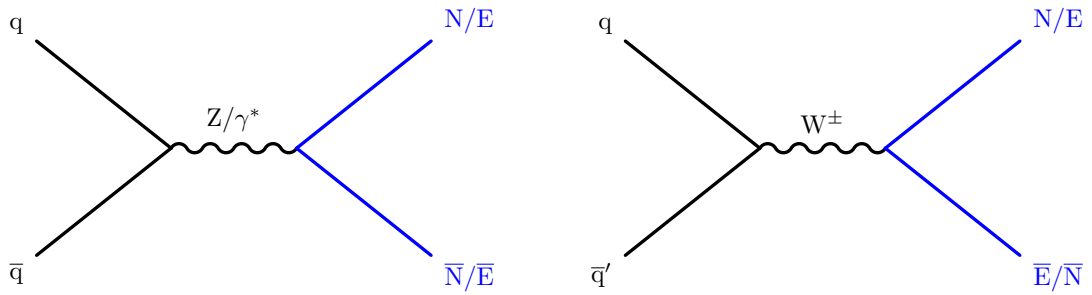


FIGURE 2.2: Pair (left) and associated (right) production modes of vector-like leptons at LHC. The associated production mode is only available for the VLL doublet model.

In both cases, the production rates are a function of only one free parameter, the mass of the VLLs. In the doublet VLL model, the neutral VLL is assumed to be mass degenerate with the charged VLL at tree level. Negligible mass splitting may arise due to one-loop radiative corrections, but they are assumed to be mass degenerate in this model for kinematic purposes. It is evident that the doublet VLL production cross section is much larger than the singlet model, as illustrated in Figure 2.3. This is partly because of the larger couplings and the associated production of charged VLL with neutral VLL. W boson-mediated s-channel EN production mode dominates the total doublet production cross-section. Table 2.1 shows the production cross-section for vector-like leptons in the singlet and doublet models.

Mass (GeV)	Singlet (pb)	Doublet (pb)	Mass (GeV)	Singlet (pb)	Doublet (pb)
100	1.17	1.69×10^{1}	550	1.78×10^{-3}	2.24×10^{-2}
125	5.45×10^{-1}	–	600	1.19×10^{-3}	1.49×10^{-2}
150	2.90×10^{-1}	3.88	650	8.05×10^{-4}	1.01×10^{-2}
200	1.05×10^{-1}	1.36	700	5.56×10^{-4}	6.97×10^{-3}
250	4.60×10^{-2}	5.89×10^{-1}	750	3.90×10^{-4}	4.89×10^{-3}
300	2.29×10^{-2}	2.91×10^{-1}	800	2.77×10^{-4}	3.47×10^{-3}
350	1.25×10^{-2}	1.57×10^{-1}	850	2.00×10^{-4}	2.49×10^{-3}
400	7.20×10^{-3}	9.07×10^{-2}	900	1.45×10^{-4}	1.81×10^{-3}
450	4.36×10^{-3}	5.49×10^{-2}	950	1.06×10^{-4}	1.32×10^{-3}
500	2.74×10^{-3}	3.45×10^{-2}	1000	7.80×10^{-5}	9.71×10^{-4}

TABLE 2.1: Production cross-section at NLO (in pb) of vector-like leptons in the doublet and singlet model at the LHC.

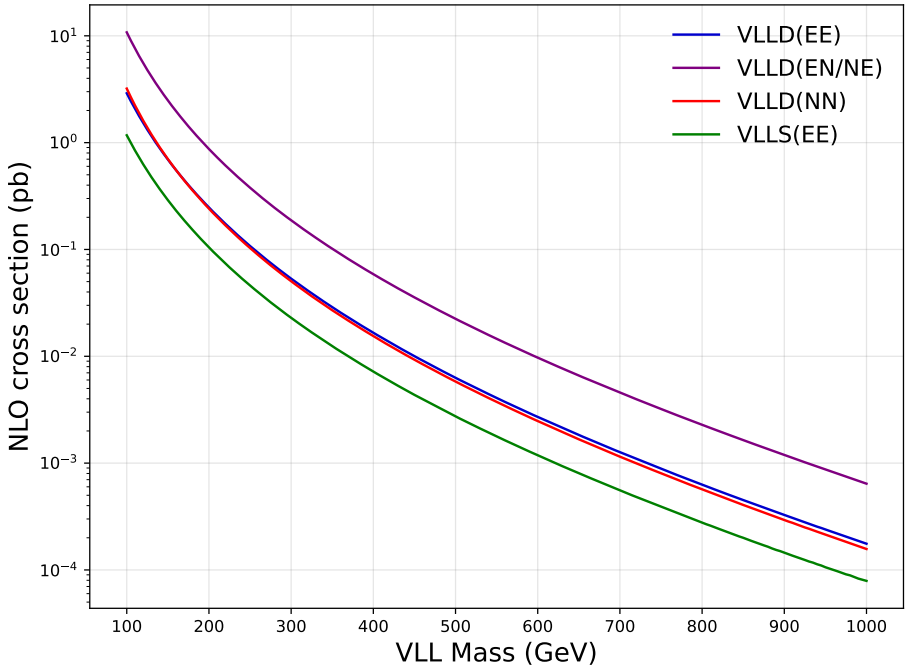


FIGURE 2.3: Production cross-section of the singlet and doublet VLL model. VLL doublet has pair and associated production modes, and the singlet model has only pair production.

After these VLLs are produced, their decays to SM leptons and the decay width of VLLs depend on the mixing parameter. It can be established from the interaction Lagrangian of this minimal model of VLLs that the branching ratios only depend on the single parameter VLL mass, as all of the widths have quadratic dependence on the mixing parameter. The following decay modes are available for doublet VLLs,

$$\begin{aligned} \mathrm{E_i} &\to Z\ell_\mathrm{i} \\ \mathrm{E_i} &\to H\ell_\mathrm{i} \\ \mathrm{N_i} &\to W\ell_\mathrm{i} \end{aligned} \tag{2.45}$$

and for singlet VLLs,

$$\begin{aligned} \mathrm{E_i} &\to W\nu_{\ell_\mathrm{i}} \\ \mathrm{E_i} &\to Z\ell_\mathrm{i} \\ \mathrm{E_i} &\to H\ell_\mathrm{i} \end{aligned} \tag{2.46}$$

Here we purposefully bring back the subscript to emphasize the three coupling scenarios of VLLs to the SM leptons family. For $\mathrm{i} = 1, 2,$ or 3, VLLs can only decay to gauge bosons and electrons, muons, or taus (and corresponding neutrinos), respectively, with their respective branching ratios. It is important to note that simultaneous decay to electrons, muons, or taus is not allowed due to the Yukawa-type mixing structure. Figure 2.4 demonstrates a few complete decay chains of VLLs producing a rich signature at the collider. For VLL mass

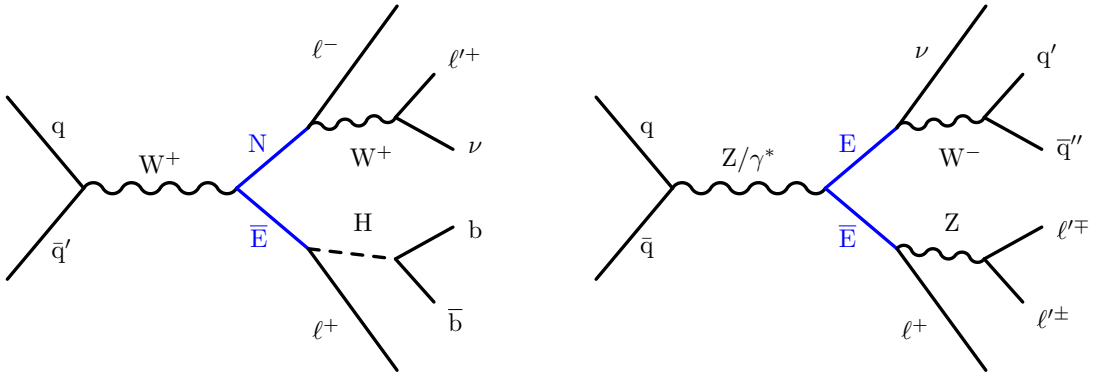


FIGURE 2.4: Examples of production and decay of vector-like leptons at the LHC.

much higher than Higgs, Z, or W, the branching ratios asymptotically approach,

$$\begin{aligned} &\text{BR}(\text{E} \to W\nu_\ell) : \text{BR}(\text{E} \to Z\ell) : \text{BR}(\text{E} \to H\ell) = 2:1:1 \text{ (Singlet VLL)} \\ &\text{BR}(\text{E} \to W\nu_\ell) : \text{BR}(\text{E} \to Z\ell) : \text{BR}(\text{E} \to H\ell) = 0:1:1 \text{ (Doublet VLL)} \end{aligned} \tag{2.47}$$

The doublet model has,

$$\text{BR}(\text{N} \to W\ell) = 1 \text{ (Only for doublet VLL)} \tag{2.48}$$

It is assumed that there is no mass mixing between this heavy neutral state and the SM neutrinos. Also, the decay of charged VLL to neutral VLL ($\text{E} \to \text{N}$) is highly kinematically suppressed in the assumed model parameter space. This is also the condition needed for the decays of these VLLs to have a decay length ($c\tau$) less than the centimeter scale (with some mass dependence). This allows the prompt decays of VLLs, where prompt refers to the decay close to the interaction point of collisions.

In this thesis, both models with all possible coupling scenarios are used for different studies.

- A search for singlet vector-like taus and vector-like muons using a dataset collected during 2016, 2017, and 2018 is discussed first. Electroweak precision data allow couplings between vectorlike leptons and SM leptons $V' \lesssim 10^{-2}$, allowing prompt decays of VLLs for the mass values around the electroweak scale [19, 20]. We assume prompt decays of VLLs, and our analysis is insensitive to the precise values of the actual mixing angles.

- The discovery potential of VLLs in the context of high luminosity LHC with proposed 3000 fb^{-1} of data at $\sqrt{s} = 14$ TeV is discussed later (Chapter 8) by recasting an earlier analysis performed with 138 fb^{-1} at center-of-mass energy 13 TeV. This projection

study presents a comprehensive picture of both the VLL models and their coupling to the first-, second-, or third-generation SM leptons.

2.3 Prior experimental constraints

Prior to direct searches at the LHC, a lower bound of about 100 GeV was placed by the L3 Collaboration at the CERN LEP collider on such additional heavy lepton states [21]; the limits are similar for both singlet and doublet scenarios.

The CMS Collaboration has carried out two direct searches targeting minimal extensions of the SM with VLLs in the $\sqrt{s} = 13$ TeV pp collision dataset. In the first of these efforts, multilepton final states with electrons and muons were probed using a dataset collected during 2016 and 2017, and the first direct constraints were set on doublet models with vector-like leptons couple to the third generation SM leptons ($\mathrm{E}_3, \mathrm{N}_3$) in the mass range of 120–790GeV [22]. This result has been superseded by a second search targeting both doublet and singlet third-generation vector-like lepton models, conducted with the larger full Run-2 dataset with additional multilepton final states, including hadronically decaying tau leptons (τ_h) [23]. In the third-generation doublet model, vector-like leptons with masses up to 1040 GeV are excluded, while the expected mass exclusion is at 970 GeV. The expected exclusion for the singlet model is only at a VLL mass ≈150 GeV, while the observed exclusion is in the VLL mass range of 125–170 GeV.

The ATLAS Collaboration excluded singlet type vector-like electrons (muons) in the mass range 129-176 GeV (114-168 GeV), except for the interval 144-163 GeV (153-160 GeV) in a trilepton resonance search using a data sample of 20.3 fb^{-1} at $\sqrt{s} = 8$ TeV p-p collisions [24]. Another ATLAS search is performed using the full Run 2 dataset at $\sqrt{s} = 13$ TeV with integrated luminosity of 140 fb^{-1} excluded the doublet third generation VLLs in the mass range of 130-900 GeV [25]. A recent ATLAS search has been reported on the VLLs coupling to first- and second-generation SM leptons using the full Run 2 dataset of 140 fb^{-1} at $\sqrt{s}$ =13 TeV [26]. The resulting mass lower limits are 1220 GeV (1270 GeV) and 320 GeV (400 GeV) for vector-like electrons (muons) in the doublet and singlet scenarios, respectively. However, vector-like taus remain a challenge to experimental searches due to their difficult parameter space.

The less stringent constraints observed in the singlet model arise from the notably lower cross section of VLL pair production, which proceeds exclusively through the $pp \to Z/\gamma^* \to \mathrm{E}\overline{\mathrm{E}}$ and involves a weaker gauge coupling strength compared to the doublet scenario. Additionally, the prevalent $\mathrm{E} \to W\nu_\ell$ decay mode in the singlet model might not result in multiple energetic charged leptons in the final state.

It is quite interesting that even if a large BSM parameter space at the TeV scale is excluded, such minimal VLL models (which appear in SUSY, GUT, etc) can evade detection primarily due to overlapping signatures with the dominant SM processes (control regions in usual searches), or the analysis choices that are tuned to probe massive particles. It is easy to miss these new particles at the electroweak scale not only in the direct searches, but also in standard model precision measurements, since the BSM effects are proportional to the mass of such new particle states. Collider searches for VLLs are limited compared to their strong sector counterparts, vector-like quarks, the so-called large radius jet (fat jet) signature generator at the LHC.

2.4 Strategy for VLLs

The analysis presented in this thesis specifically targets low-mass singlet VLLs at the electroweak scale. As discussed earlier, the signal acceptance is small in the multilepton analysis for such a signal mainly due to the following reasons:

- At low mass: VLLs predominantly decay through $\mathrm{E} \to W\nu_\ell$, and not enough charged leptons are produced in the final state. Even at high mass, this decay mode contributes to 50% of the total decay modes.
- For VLLs coupling to the third generation SM leptons (vector-like taus), the additional suppression of tau leptonic decay, or the limited experimental hadronic tau reconstruction and identification efficiency, resulted in poor acceptance for such a signal. For other coupling scenarios, the possibility of getting well-reconstructed and identified leptons from the VLL increases the signal acceptance, and different kinematic distributions from the SM processes.
- At high mass, other decay modes ($\mathrm{E} \to Z\ell/H\ell$) contribute to the final state with multiple charged leptons. However, due to the extremely small production cross-section, the collected amount of data is not sufficient to be sensitive to the current analysis strategy.
- Boosted Decision Trees (BDTs) were used in earlier multilepton analysis for the vector-like tau singlet model, but the training performance was suboptimal at low mass due to the low signal yield.

Beyond multilepton and third-generation coupling

This requires devising a strategy to look beyond the multiple charged leptons in the final state, considering other decay modes, such as hadronic decay of the gauge bosons, to take

advantage of the higher branching ratios. A smaller lepton multiplicity is required to be sensitive to the dominant decay modes at low mass. On the other hand, there is no strong theoretical basis to assume the VLLs coupling to tau leptons is preferred over electrons or muons.

With this in mind, we will describe an analysis designed to search for singlet VLLs in muon and at least two jets in the final state. Feynman diagrams exemplifying the production and decay of VLL pairs in the final state with a lepton and jets are shown in Figure 2.5. It is important to note that this would allow huge SM backgrounds (mostly dominated by $W+$ jets production) and very similar signal kinematics, as the background at this mass range may limit the experimental sensitivity to such BSM particles. However, a dedicated machine learning strategy can be designed to optimally separate the signal events from the huge SM backgrounds. The new final state with one muon and two or more jets considered (μjj) in this analysis may yield enough events (unlike the multilepton final state) to train a ML algorithm for optimal performance, with a possibility to improve the sensitivity targeting the singlet model compared to the earlier iteration of the multilepton analysis. This analysis has also been extended to cover a large model parameter space by targeting the VLLs coupling to second-generation SM leptons (vector-like muons), which has never been probed in CMS.

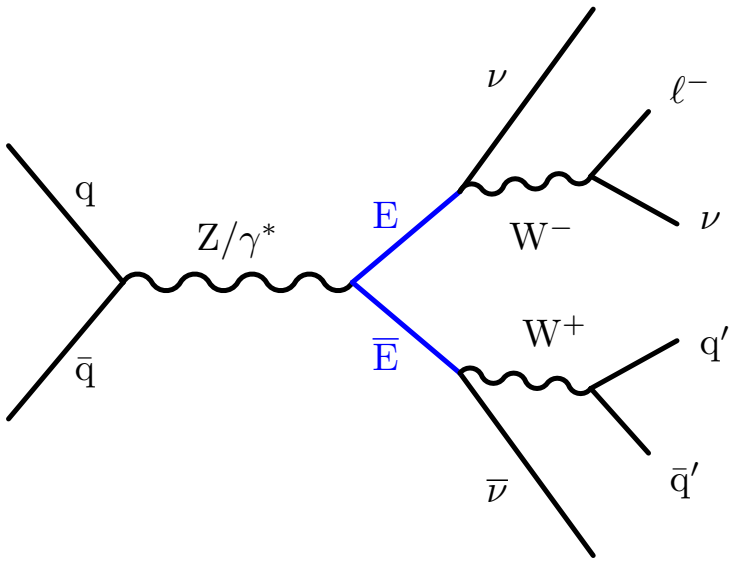


FIGURE 2.5: Example Feynman diagram illustrating the production and decay of singlet vector-like leptons at the LHC with lepton and jets in the final state.

Chapter 3

The Experimental Apparatus

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator located at CERN (the European Organization for Nuclear Research) near Geneva, Switzerland [27]. The LHC is a 27 km long circular tunnel built at an average depth of 100 meters beneath the France-Switzerland border near Geneva.

Inside the LHC, two proton (or ion) beams travel close to the speed of light in opposite directions in separate beam pipes kept at ultrahigh vacuum. These beams are guided around the accelerator ring by a strong magnetic field produced by superconducting electromagnets. The electromagnets in the LHC are cooled down to $-271.3\,\mathrm{deg}$ C (1.9K), a temperature colder than outer space, to achieve their superconducting state. These beams are then made to collide head-on with each other at four different points around the machine. At these collision points, four different particle detectors - CMS, ATLAS, ALICE, and LHCb- are situated to detect the particles created in such high-energy collisions. Out of these, CMS and ATLAS are general multipurpose detectors designed to be sensitive to a range of new physics, while ALICE focuses on studying heavy-ion collisions, and LHCb is designed to study the forward decays of the bottom and charm hadrons. Figure 3.1 shows a view of the accelerator complex at CERN. Operationally, LHC collides proton-proton (p-p) beams, proton-lead (p-Pb), and lead-lead (Pb-Pb) ion beams at different times of the year. The center-of-mass energy of the collision is denoted as $\sqrt{s}$, where s is one of the Mandelstam variables and varies based on the colliding particles. The LHC collisions are often categorized in different *Run* based on the collision energy and number of data-taking years.

Run-1 of the LHC is between 2010 and 2013, when the proton collisions happened at $\sqrt{s} = 7$ TeV in 2010/2011 and 8 TeV in 2012. The next run began in 2015 at $\sqrt{s} = 13$ TeV and concluded in 2018. This period is called Run-2, and this thesis is based on the data collected by the CMS detector in this period. Currently, the LHC is in the Run-3 phase with an increased collision energy of 13.6 TeV, which started around July 2022 and is scheduled to conclude in 2025. The intermediate periods between the Runs are called

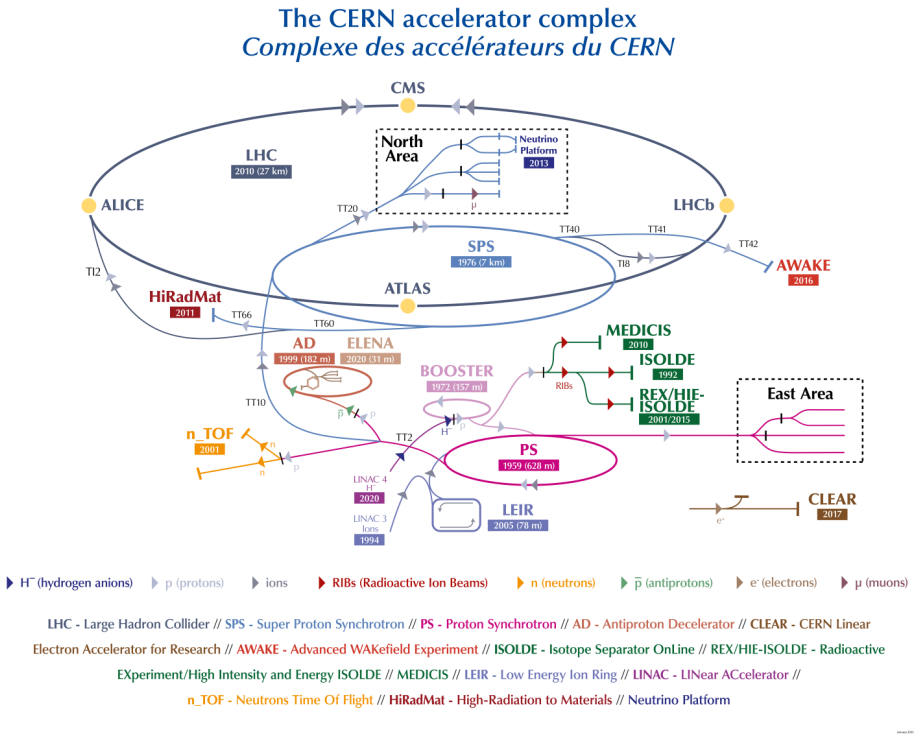


FIGURE 3.1: Schematic view of the CERN accelerator complex showing various steps of accelerating the proton beam and the four collision points where the detectors are built to study the particles produced in the collisions. [27]

the long shutdown (LS) period, when detector or accelerator improvements are carried out. One such crucial LS period will commence at the end of Run-3, and a series of detector and accelerator improvements will be carried out to prepare for the anticipated high-luminosity phase of LHC (HL-LHC), which is described in Chapter 8.

3.1.1 Quantifying collisions

Every bunch crossing is termed a *collision event*, or simply an *event*. The collision takes place at an interval of 25 ns or at a frequency of 40 MHz. Since the size of the colliding particles is very small, a higher density of particles in each bunch is needed to increase the chance of head-on collisions. A relevant quantity could be defined to measure the number of potential collisions per surface unit (or cross-section) per unit of time named instantaneous luminosity ($\mathcal{L}$). In particle physics, a cross-section measures the probability of some interaction happening and is measured in the units of area - barns, b ($1b = 10^{28} m^2$). A higher instantaneous luminosity means a greater likelihood of particles colliding and resulting in interactions, which is crucial in increasing the probability of interactions we are interested in. At the LHC, the instantaneous luminosity depends only on the beam parameters and can

be calculated for a Gaussian beam distribution as:

$$\mathcal{L} = \frac{N_p^2 n_B \gamma f_{rev}}{4\pi\epsilon_n \beta^*} F \tag{3.1}$$

where N_p is the number of particles per bunch, n_b the number of bunches per beam, f_{rev} the revolution frequency, γ the relativistic gamma factor, ϵ_n the normalized transverse beam emittance, β^* the betatron function at the collision point, and F the geometric luminosity reduction factor due to the crossing angle at the interaction point (IP):

$$\mathcal{F} = 1/\sqrt{(1 + (\frac{\theta_c \sigma_z}{2\sigma^*})^2)} \tag{3.2}$$

θ_c is the full crossing angle at the IP, σ_z the RMS bunch length, and σ^* the transverse RMS beam size at the IP. With these values, Eqn 3.1 yields an instantaneous luminosity of $\mathcal{L} \approx 2.1 \times 10^{34}$ cm^{-2}s^{-1}. This is the highest luminosity to be achieved by any hadron collider in the world. The integrated luminosity, L_{int}, is calculated by integrating the instantaneous luminosity over the total time for which the collisions were happening in the unit of inverse barns (b^{-1}). Since the inverse barn is a large unit, often data collected at high-energy experiments are reported in much smaller units such as inverse picobarn (pb^{-1}) or inverse femtobarn (fb^{-1}).

The amount of proton-proton collision data delivered by the LHC and collected by the CMS detector depends on the data collection efficiency of the detector. Efficiency should be as close to 100% to maximize the detector's physics capability. Figure 3.2 shows LHC delivered and CMS recorded data from 2016 to 2018. CMS collected 150 fb^{-1} of data in Run-2 for the data-taking year between 2016-2018, with 41.6 fb^{-1} , 49.8 fb^{-1} , 67.9 fb^{-1} of data in 2016, 2017, and 2018, respectively. In the analysis, we use the data after it has passed specific filters to ensure we do not use the events where the important detector part was not operating correctly. After applying these event filters, 138 fb^{-1} data is used in the analysis for the full Run-2 period, comprising approximately 19.5, 16.8, 41.5, and 59.8 fb^{-1} from 2016 preVFP, 2016 postVFP, 2017, and 2018, respectively. The 2016 dataset is split into two separate eras to take into account different detector conditions. In the 2016 preVFP era, the strip tracker had a lower signal-to-noise ratio and fewer hits on tracks due to saturation effects in the readout chip under high-luminosity conditions. This was mitigated in the 2016 postVFP era by changing the feedback preamplifier bias voltage (VFP) [28].

The number of events of a probable physics process in the p-p collision could be written as:

$$N_{process} = L\sigma_{process} \tag{3.3}$$

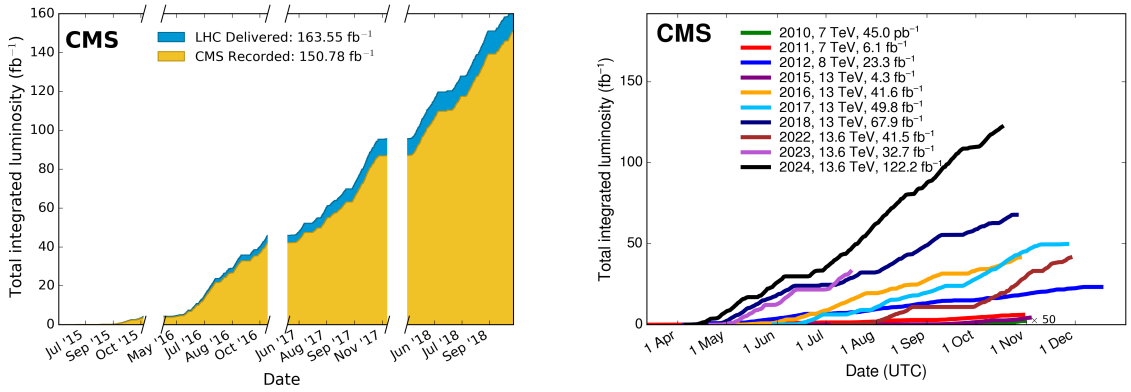


FIGURE 3.2: Delivered and recorded luminosity cumulative over 2015-2018 during stable beams (left) and for all years of data-taking, including the Run-3 periods (right) for pp collisions at nominal center-of-mass energy. Image Courtesy: CMS Luminosity measurement group

where L is the integrated luminosity, and $\sigma_{process}$ is the cross-section of the physics process under consideration.

3.2 The CMS Detector

The CMS detector is one of the two (the other being ATLAS) multipurpose detectors at LHC [29]. It is located 100 meters underground, close to the French village of Cessy, at point 5 of the LHC ring. CMS was designed to discover the Higgs boson and explore BSM physics at the TeV scale. The CMS detector is shaped like a cylindrical onion, with several concentric layers of components, and built around a huge solenoid magnet made of a cylindrical coil of superconducting fibres. It can generate a magnetic field of 3.8 tesla. The inner silicon tracker, electromagnetic calorimeter, and hadron calorimeter are inside the superconducting solenoid. The magnetic field is returned by a large steel yoke, which also holds the muon chambers and helps keep the CMS detector structurally stable [30]. The high magnetic field in the tracker volume provides an excellent measurement of the charged particle momentum and helps in distinguishing between energy deposits of neutral and charged particles in the calorimeters. A complete picture of the collision is built by combining information from various subsystems, with a precise and accurate measurement of the momentum and energy of electrons, photons, muons, or hadrons. Figure 3.3 shows the basic design of the CMS detector.

3.2.1 CMS coordinate system

The CMS experiment uses a right-handed coordinate system, with the origin centered at the nominal interaction point, the x-axis pointing towards the center of the LHC ring, the y-axis

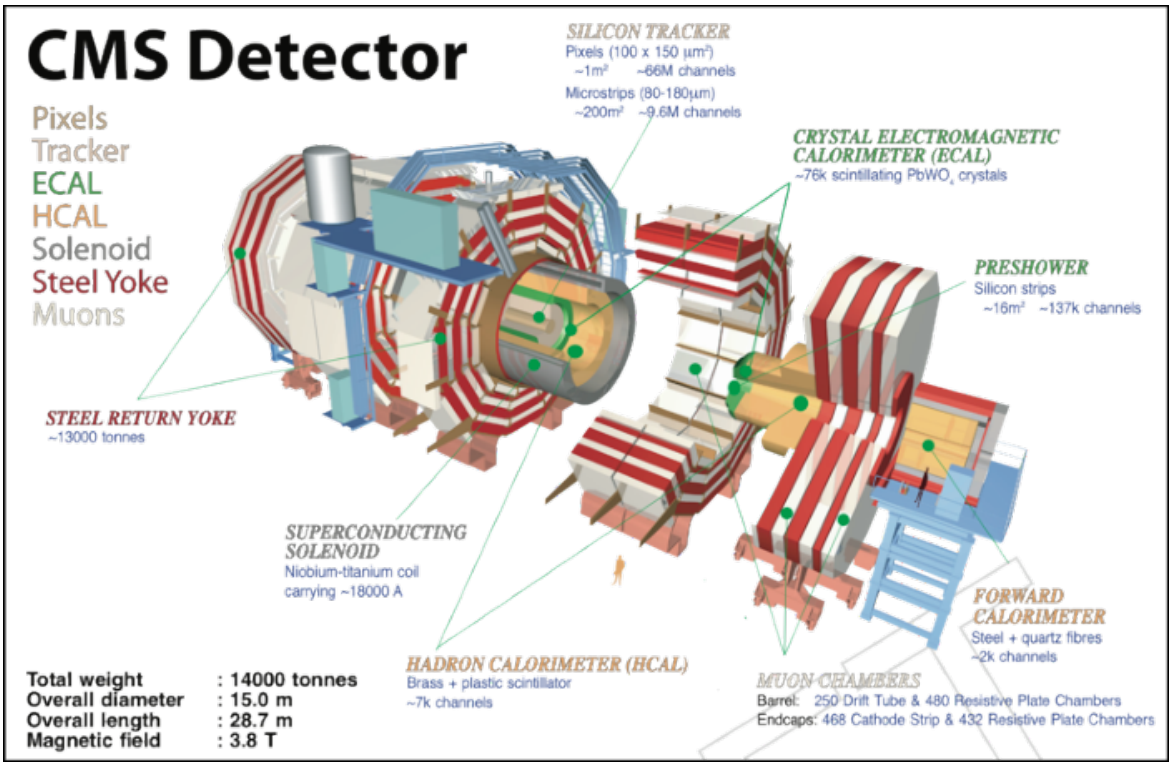


FIGURE 3.3: A view of the CMS detector at the LHC, CERN, with different subcomponents in an onion shell structure. Image courtesy: CMS

pointing up (perpendicular to the LHC plane), and the z-axis along the counterclockwise beam direction (towards Jura mountain). As it is a cylindrical geometry, the azimuthal angle ϕ is measured in the x-y plane from the positive x-axis towards the y-axis with $\phi = 0$ along the positive x-axis, and $\phi = \pi/2$ along the positive y-axis. The polar angle θ is the angle between the particle's 3D momentum vector and the positive z-axis (beam direction). The transverse momentum vector, $\vec{p_T}$ is the projection of the 3D momentum vector, $\vec{p}$ on the transverse (xy) plane. Figure 3.4 demonstrates the particle three momentum vector ($\vec{p}$), and transverse momentum vector ($\vec{p_T}$) in the CMS coordinate system.

In the hadron colliders, the energies and the momenta of the incoming hadrons are known, but the energies and momentum fractions of the participating partons in the collisions are not known a priori. We measure the produced particles in the detector to characterize the collisions. Therefore, using a modified form of a spherical coordinate system suited to the detector geometry is convenient. One such example is rapidity (y), which is preferred over the polar angle θ because the rapidity difference between particles is Lorentz invariant under boosts along the longitudinal axis (or z-axis), which corresponds to the transformation between the detector frame and the center-of-mass frame. The rapidity (y) is given as,

$$y = \frac{1}{2} ln(\frac{E + p_z}{E - p_z}) \tag{3.4}$$

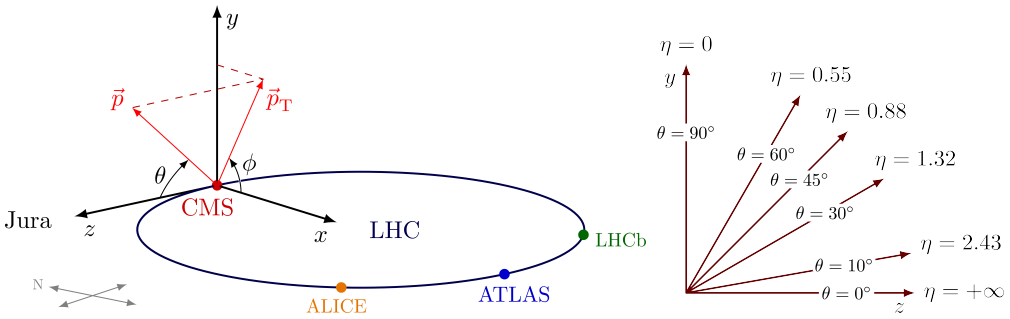


FIGURE 3.4: The CMS coordinate system against the backdrop of the LHC (left) and pseudo-rapidity coverage of the detector plane (right). Image courtesy: Izaak Neutelings.

where E and p_z are, respectively, the energy and z-component of the momentum of the particle. y-ϕ coordinates exploit the cylindrical symmetry of the detector better, and y can be directly mapped to the detector geometry. However, the rapidity can be hard to measure as the z-component of the momentum is usually not measured in the detector, as it is along the beam direction. For massless particles or particles with very high energy ($p >> m$), the Equation 3.4 may be reduced to pseudorapidity (η),

$$\eta = -ln[tan(\frac{\theta}{2})] \tag{3.5}$$

The pseudorapidity can be obtained by measuring the polar angle θ, and it is used with the azimuthal angle ϕ to indicate the position of a particle in the detector. The conversion of θ and η value is shown in Figure 3.4 right. The pseudorapidity value changes from minus to plus infinity for the polar angle between $-\pi/2$ to $\pi/2$. Typically, the central region of the detector is defined by $\eta = 0$, and as the pseudorapidity increases, the particles move closer to the beam axis or forward detector region. Rapidity and pseudorapidity do not coincide for massive particles, and rapidity is often used in those cases. The particles emerging from the collisions interact with different parts of the detector components as they traverse through the detector. In the following subsections, the different CMS detector subcomponents will be discussed.

3.2.2 Inner tracker

The inner tracker is located at the innermost part of the CMS detector, as close as 3 cm from the collision vertex. The inner tracking system measures the trajectories of charged particles emerging from the LHC collisions precisely and efficiently. Since it is located in the innermost part of the detector, it is exposed to high particle density and intense radiation. It should be able to distinguish between many overlapping p-p interactions at each bunch crossing (at a 40 MHz rate). Therefore, the inner tracker must be highly granular to identify close-by trajectories, radiation hard, and fast. These requirements led to a tracker design

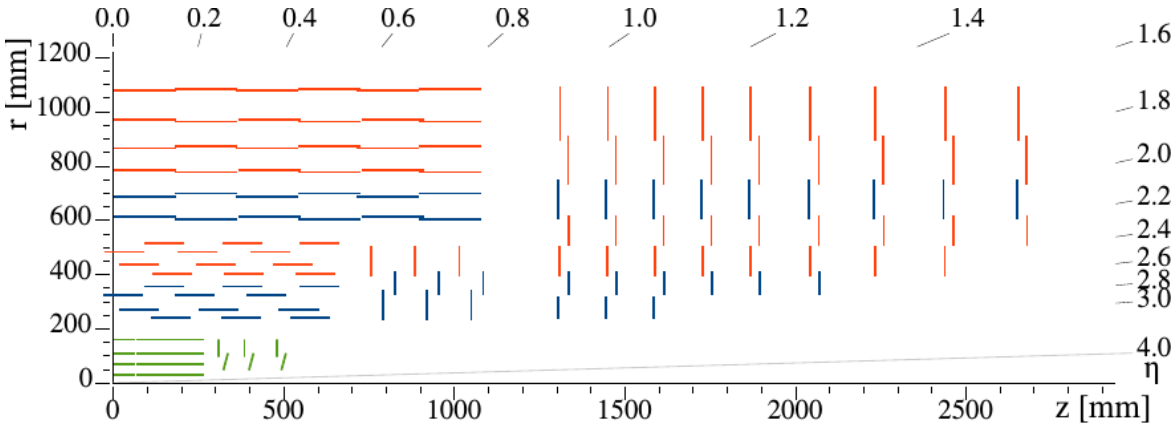


FIGURE 3.5: A schematic diagram of one-quarter of the CMS inner tracking system. The innermost barrel layer was installed in 2017 during the phase-1 upgrade of the CMS detector. *Image courtesy: CMS Tracker detector performance group*

.

entirely based on silicon detector technology. A schematic diagram of the CMS tracker layers is illustrated in Figure 3.5. The CMS tracker [31, 32] is composed of a pixel detector with four barrel layers at radii between 2.9 cm and 16.0 cm and a silicon strip tracker with 10 barrel detection layers extending outwards to a radius of 1.1 m. It has four subsystems: the Tracker Inner Barrel (TIB), Disks (TID), Tracker Outer Barrel (TOB), and Tracker EndCaps (TEC). Each system is completed by endcaps which consist of 2 disks in the pixel detector and 12 disks in the strip tracker on each side of the barrel, extending the acceptance of the tracker up to a pseudorapidity of $|\eta| < 2.5$. The silicon sensor modules are finely segmented into 66 million 150×100 μm pixels and 9.6 million 80×180 μm wide strips. With about 200 m^2 of active silicon area, the CMS tracker is the largest silicon tracker ever built. The tracker is submerged in the homogenous magnetic field provided by the superconducting solenoid that bends the charged particles from which their direction and momenta are measured. One important aspect is that the particles traversing the inner tracker should be exposed to a minimum amount of material before they deposit their energy in subsequent calorimeter systems. Figure 3.6 shows the total thickness of the tracker material traversed by a particle produced at the nominal interaction point, as a function of pseudorapidity. The thickness in terms of the radiation length (X_0) is 0.4 X_0 in the barrel region ($|\eta| < 1.1$) to a maximum material budget of 1.8 X_0 in the transition region ($|\eta| < 1.4$), and then 1.0 X_0 in the endcap region ($|\eta| \approx 2.5$). This is crucial to suppress nuclear interactions with the material for an efficient tracking algorithm.

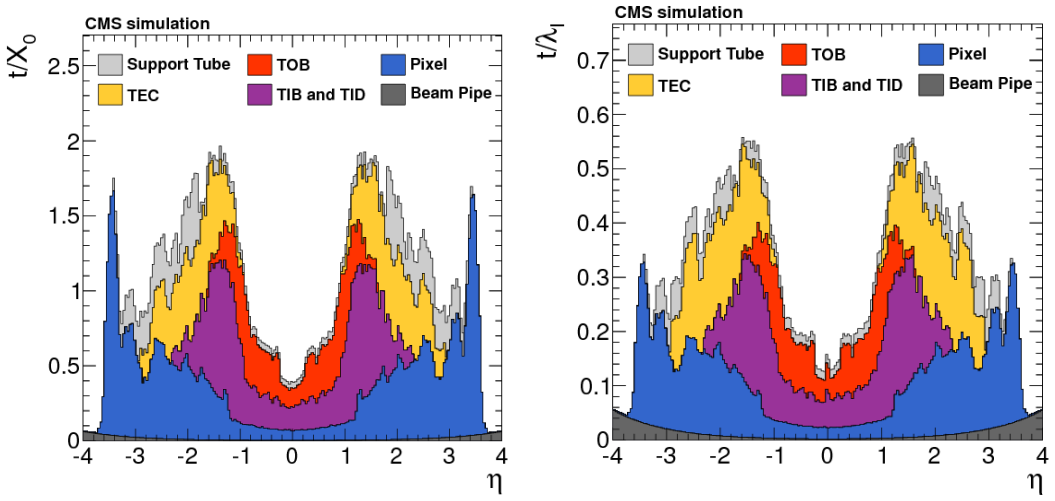


FIGURE 3.6: Total thickness of the tracker material traversed by a particle produced at the nominal interaction point, as a function of pseudorapidity, expressed in units of radiation length (left) and nuclear interaction length(right) [31]

.

3.2.3 Electromagnetic calorimeter

The CMS electromagnetic calorimeter (ECAL) is designed to measure the energy of particles that deposit energy predominantly via electromagnetic interactions. Given the extreme condition of LHC collisions, the ECAL must respond fast, compatible with the 25ns of bunch crossing, radiation hard with fine granularity. Crystals made of lead tungstate ($PbWO_4$) are chosen for their high density ($8.28\ g/cm^3$), transparency, short radiation length (0.89 cm) and small Moliere radius (2.2 cm). CMS ECAL is a compact, hermetic, and homogeneous calorimeter. The barrel section (EB) is made of 61200 lead tungstane crystals (with dimension of $2.2 \times 2.2 \times 23$ cm) covering $|\eta| < 1.47$ and two endcap disks (EE) are made of 7324 crystals ($2.86 \times 2.86 \times 22$ cm) each with a coverage of $|\eta| < 3.0$. The layout of the CMS ECAL is shown in Figure 3.7. ECAL is constructed out of single lead tungstate crystals, which can contain the full longitudinal EM shower and the transverse shower profile due to the small Moliere radius, which provides excellent energy resolution. The crystals emit scintillation light proportional to deposited energy, read out by Avalanche Photodiodes (APDs) in the barrel and Vacuum Phototriodes (VPTs) in the endcaps [29].

A preshower (PS) detector is also installed in the endcap region covering a fiducial region $1.653 < |\eta| < 2.6$ to enhance photon and neutral pion (π^0) identification by distinguishing between single high-energy photons and overlapping photon pairs from $\pi^0 \rightarrow \gamma\gamma$ decays. It also improves the position determination of electrons and photons with high granularity. The preshower is a sampling calorimeter with a total thickness of 20 cm. It is made of two active silicon strip layers and two lead absorber plates placed before each silicon layer. Lead radiators initiate the EM showers, and high granularity of silicon strip layers (137,000

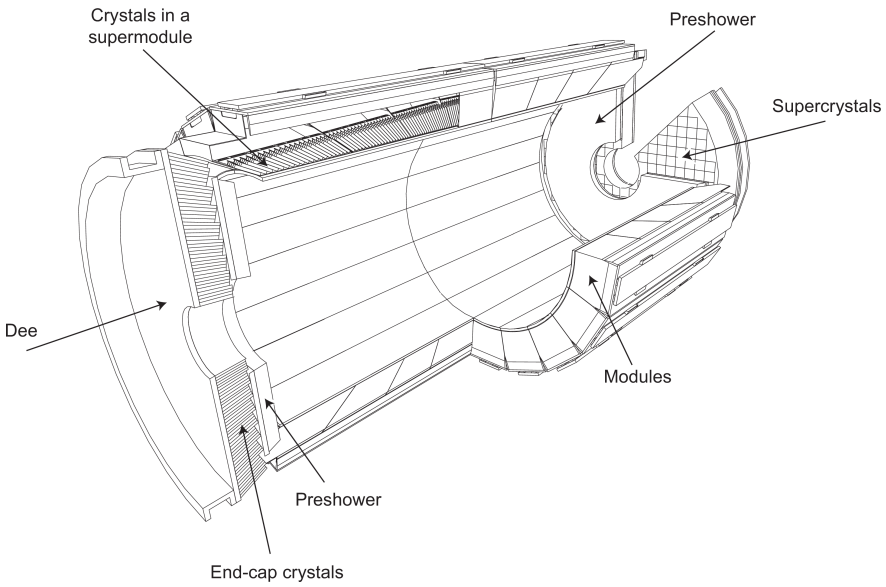


FIGURE 3.7: Layout of the CMS electromagnetic calorimeter, showing the barrel supermodules, the two endcaps, and the preshower detectors [33].

channels) provide precise shower shape measurements of the incoming particles.

3.2.4 Hadron calorimeter

The hadron calorimeters are particularly important for the measurement of hadron jets and neutrinos or exotic particles resulting in apparent missing transverse energy. The CMS hadron calorimeter (HCAL) measures the energy and direction of neutral or charged hadrons (made of quarks or gluons) via nuclear interaction of such particles with the detector material. HCAL is a sampling calorimeter that uses alternating layers of dense absorber material and active scintillator layers to detect hadronic showers. HCAL is positioned between the ECAL and the CMS solenoid magnet and is composed of four main subsystems: HCAL barrel (HB), HCAL endcaps (HE), HCAL forward detector (HF), and HCAL outer detector (HO). HB covers $|\eta| < 1.3$, surrounding the ECAL barrel, and HE covers $1.3 < |\eta| < 3.0$, placed behind the ECAL endcaps. HF helps to detect highly forward jets ($3.0 < |\eta| < 5.2$), and HO is a single-layer detector outside the solenoid to catch late-developing hadronic showers. A schematic diagram of the HCAL subcomponents can be seen in Figure 3.8.

The HB and HE calorimeters are composed of layers of brass absorbers (5 cm thick) interleaved with plastic scintillator tiles as active material. Hybrid photodiodes (HPDs) and silicon photomultipliers (SiPMs) convert scintillation light into an electrical signal. The granularity of the HB and HE calorimeters depending on the η coverage are $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ for for $|\eta| < 1.6$ and $\Delta\eta \times \Delta\phi = 0.17 \times 0.17$ for $|\eta| \geq 1.6$. The thickness of the calorimeter is 5.8 interaction lengths (λ_I) in HB and 10 interaction lengths in HE, ensuring full hadronic shower containment. HF uses quartz fibers embedded in steel. Cherenkov light

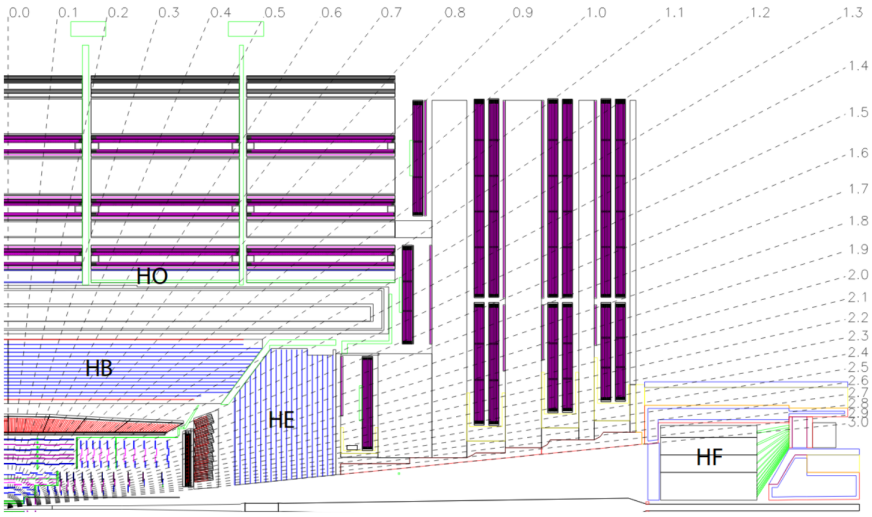


FIGURE 3.8: Schematic diagram of the CMS HCAL in the r-z plane with the pseudorapidity ranges. [29]

produced by charged particles in the fibers is detected by photomultiplier tubes (PMTs). It is designed to withstand high radiation levels due to the high occupancy in the forward region.

3.2.5 Muon system

Precise and robust detection of muons is central to the CMS physics goals. Muons are charged particles like electrons or positrons, but 200 times heavier. Muons don't lose energy in the tracker via bremsstrahlung, and being a minimizing ionization particle (MIP), they are not stopped by any of the CMS calorimeters. Muons are effectively long-lived in the dimension of the CMS detector. Therefore, the CMS muon spectrometer (sometimes called muon tracker) is located outside the superconducting solenoid and is embedded within the iron return yoke, which acts as a magnetic field guide and a hadron absorber. Muon chambers in the CMS detector rely on gaseous ionization and avalanche multiplication to detect charged particles. When a muon traverses a chamber, it ionizes the gas, creating free electron-ion pairs. Under the influence of an applied electric field, these electrons drift toward an anode, amplified through avalanche multiplication, generating a measurable electrical signal. Three types of gaseous particle detectors are used to measure muon momentum and for identification purposes. The barrel drift tube (DT) chambers, arranged in four stations, interleaved with iron return yoke layers, cover the pseudorapidity region $|\eta| < 1.2$. Cathode strip chambers (CSC) are used for their fast response time, fine segmentation, and radiation resistance in the endcap region ($0.9 < |\eta| < 2.4$) where the muon rates and background levels are high with the non-uniform magnetic field. DT and CSC subsystems are crucial to trigger an event based on the p_T of muons with good efficiency and momentum resolution. To remain fully efficient when the LHC reaches full luminosity, resistive plate chambers (RPC) was

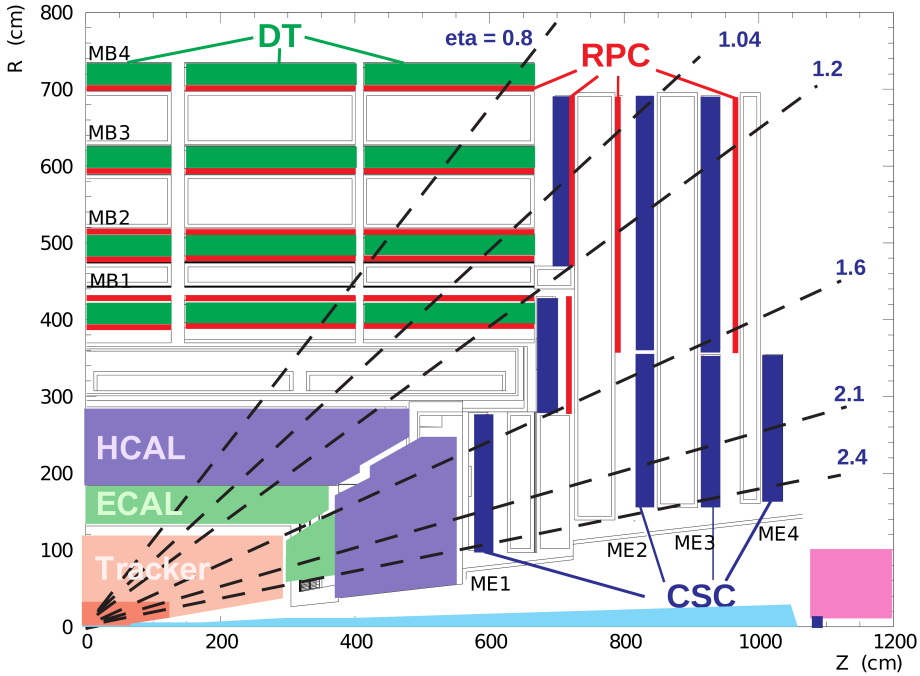


FIGURE 3.9: Schematic of one quadrant of the CMS muon systems [34].

added in both the barrel and endcap region to leverage them as a dedicated trigger system for their fast, independent, and highly segmented trigger with a sharp p_T threshold covering a large pseudorapidity region ($|\eta| < 1.6$) of the muon system. It provides sub-nanosecond timing resolution (1 ns), making them ideal for triggering muons at the LHC bunch-crossing rate. Gas electron multipliers (GEMs) are a new addition to the CMS muon system. They complement existing detectors in the forward regions close to the beam pipe, where large radiation doses and high event rates will increase during Phase 2 of the LHC (HL-LHC). 144 GEM chambers were installed during Long Shutdown 2 on the first disk of the two endcaps. These chambers are contributing to the Run-3 data-taking of the LHC. Two more disks of GEM chambers will be installed in each endcap during 2024-2026, before Phase 2 of the LHC. Figure 3.9 illustrates the different muon subsystems of the CMS detector.

There are total 1400 muon chambers, out of which 250 drift tubes (DTs) and 540 cathode strip chambers (CSCs) track the particle positions and provide a trigger, while 610 resistive plate chambers (RPCs) and 72 gas electron multiplier chambers (GEMs) form a redundant trigger system used to keep or discard the acquired muon data quickly.

3.2.6 Trigger and Data Acquisition System

The LHC collides proton-proton bunches 40 million times per second, leaving only 25 ns to keep or discard an event before the next event arrives. Ideally, data from all events should be kept to analyze later in offline (not real-time) mode. However, it is impossible to store events

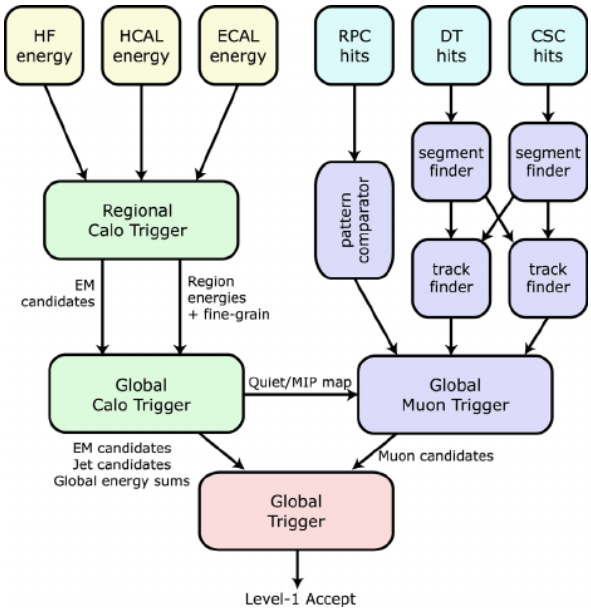


FIGURE 3.10: Schematic of the L1 global trigger system at CMS [29].

at such a high rate with the existing storage capacity. Moreover, most interactions might be low-energy glancing collisions, for instance, rather than head-on energetic collisions, and are unlikely to have interesting new phenomena. CMS employs a two-level trigger system to filter out interesting events and reduce the number of events from 1 billion to 1000 events every second. These two steps are Level-1 (L1) trigger and High-Level Trigger (HLT). The detectors require excellent time resolution, and the signals from millions of electronic channels must be precisely synchronized to distinguish the particles from different bunch crossings.

L1 trigger relies on custom electronics, Field-Programmable Gate Arrays (FPGAs), and Application-Specific Integrated Circuits (ASICs) to make ultra-fast decisions. L1 reduces the collision rate from 40 MHz to 100 KHz with a latency of 3.2 μs and has to analyze every bunch crossing. Since the decision-making time is limited, L1 uses coarsely segmented data from calorimeters and muon systems to construct objects such as electrons, muons, jets, and missing energy. These are called trigger primitive objects. These objects are then fed into the global muon and calorimeter trigger, which sorts them based on the quality of reconstruction, momentum, or energy of these trigger objects, and passes them to the global trigger system. The global trigger can either reject or accept the event based on algorithm calculations and the readiness of the sub-detectors and the Data Acquisition System (DAQ). Figure 3.10 illustrate the L1 trigger system. A schematic view of the components of the CMS DAQ system is shown in Figure 3.11.

The L1 Trigger electronics are partly installed on the detectors and partly in the underground control room. The L1 accepted events are then transferred to a computer farm via the

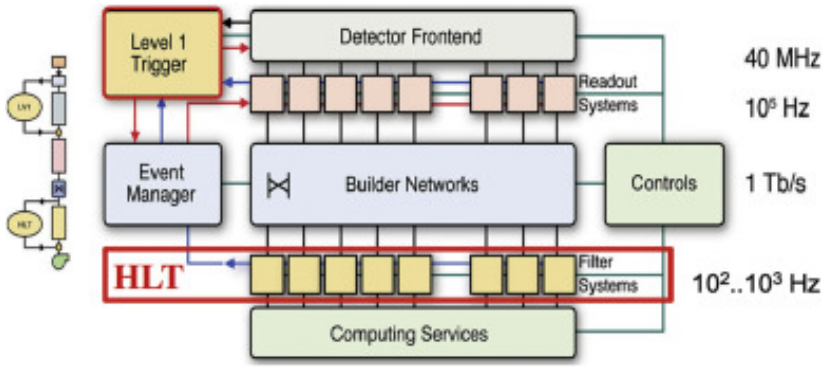


FIGURE 3.11: Schematic of the CMS DAQ system [29].

DAQ that performs a more complex, sophisticated, but faster version of offline reconstruction algorithms, physics selections to reduce the rate of stored events by a factor of 1000. The filtering process at HLT uses the full granularity of the detector data for L1 accepted events. It is structured around the concept of HLT paths, a set of algorithmic processing steps run in a predefined order that reconstructs and makes selections of the physics objects. For example, electron and photon triggers reconstruct the ECAL shower and apply cluster shape and isolation criteria with information from the tracking system. Similarly, muon, jet, and missing energy triggers use the relevant subsystem information to select events, keeping the trigger bandwidth at the accepted level. The recent addition of a machine learning based anomaly trigger (trained on minimum bias data) in the HLT path enhances the chance of storing anomalous interesting events that may not be kept otherwise in the usual high-level trigger strategy. This may increase the discovery potential of the LHC by analyzing anomalous events in a model-agnostic fashion [35].

Analysis trigger

The list of trigger paths used in this analysis to select events with at least one isolated muon is unprescaled isolated single muon trigger paths across the three years of data-taking. Offline transverse momentum (p_{T}) thresholds of 26, 29, and 26 GeV are chosen to remain fully efficient in 2016, 2017, and 2018, respectively. Figure 3.12 shows the trigger efficiency for 2018 in barrel (left) and endcap (right), demonstrating high trigger efficiency compatible with the choice of offline threshold. Table 3.1 describes the trigger paths and the offline threshold used in this analysis for all the eras. The trigger efficiency measurement study in the three years of data-taking is demonstrated in the Appendix 11.1.1.

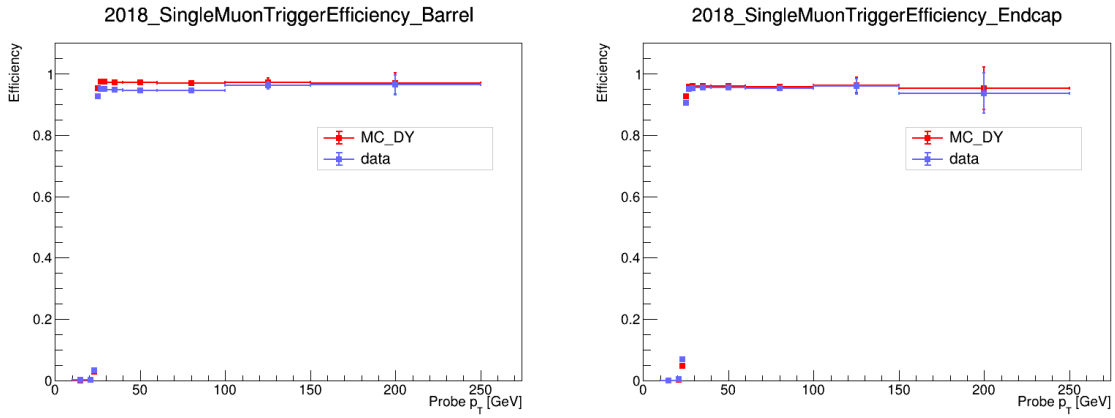


FIGURE 3.12: Single muon trigger efficiency in data (blue) and simulation samples (red) as a function of muon p_{T} for 2018.

Era	Triggers	Offline p_{T} threshold
2016	HLT_IsoMu24 OR HLT_IsoTkMu24	26 GeV
2017	HLT_IsoMu27	29 GeV
2018	HLT_IsoMu24	26 GeV

TABLE 3.1: Isolated single muon trigger paths and offline p_{T} threshold used in the analysis for different data-taking years.

Chapter 4

Simulation and Event Reconstruction

In the previous chapter, we discussed the complex CMS detector and the design of various subdetector components driven by the interaction of produced particles in the p-p collision with the detector materials. It is important to understand each part of the detector performance, particle interactions with the detector material, and even to understand the feasibility of adding new detector components to the existing ones. Often, our knowledge about the quantum field theory based calculation of a physics process initiated by the colliding particle beam, underlying physics principles for the particle-matter interaction are parceled in computer programs to study the response of the detector, model particle interactions, identify potential biases, optimize detector's performance, and design data analysis strategy to identify unknown potential signals, and extract meaningful information about fundamental physics. In short, they are used to simulate virtual particle collisions on our computers at a much cheaper cost, and we can test our understanding by performing the experiments. At the same time, results from the experiments are fed back to the simulation software to improve its robustness and reliability of prediction.

In the next few sections, we will explore how collision events are simulated in CMS. After the simulation, the reconstruction and identification of different particles or physics objects to reconstruct the full collision event are described.

4.1 Monte Carlo simulation

Simulation techniques are necessary to mimic the detector environment to understand the experimental conditions and model many physics processes. SM background processes in a BSM search could also be modeled using an appropriate simulation. We can interpret the results of the proton-proton collisions by comparing the simulation to the experimental measurements. Monte Carlo (MC) simulation techniques are used to replicate the randomness of nature in the simulations.

Data begins with proton beams, while the simulation chain starts from a theorist writing down the Lagrangian of an interaction or theory. Various couplings of SM particles, BSM

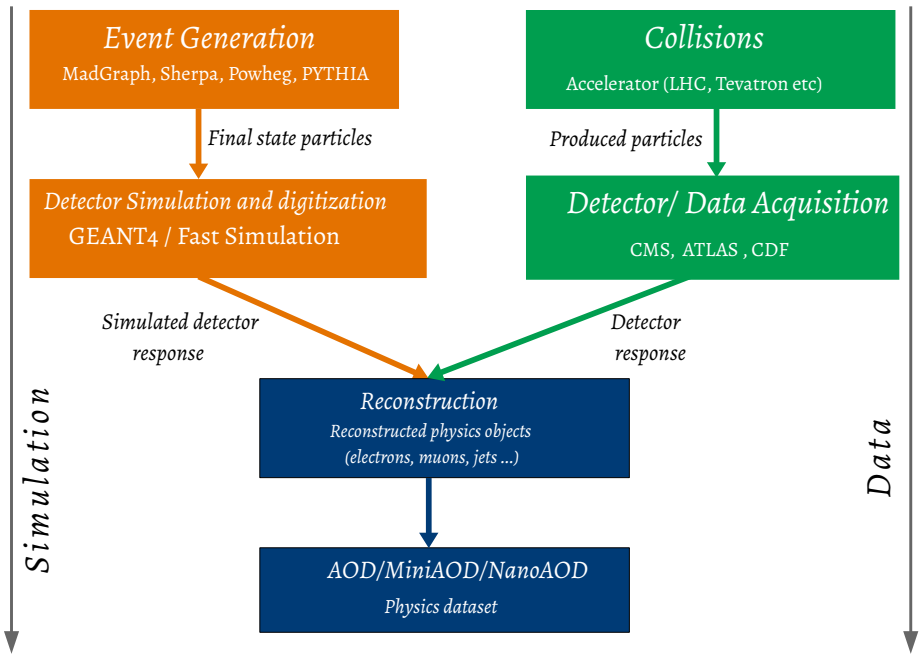


FIGURE 4.1: Schematic diagram of the simulation and data workflow.

particles, decay widths, and allowed interaction vertices encoded in a Lagrangian are generally provided in a universal file format called the UFO model. We use the UFO files in the event generators to produce the events. There are four main parts of connecting the simulation to the experimental measurements: event generation, simulation, digitization, and reconstruction. Figure 4.1 illustrates schematically the simulation steps and the actual data workflow.

4.1.1 Event generation

Event generators are designed to simulate (could be called generate) the final state particles of collisions. Event generators like MADGRAPH use the parton distribution function to pick a parton for interaction from each proton. The hard interactions between the partons can be represented in terms of Feynman diagrams, and the scattering amplitudes can be calculated by integrating the amplitudes coming from Feynman diagrams of all orders. This calculation depends on the interacting particle's color charge, four momentum, spin, etc, and a multi-dimensional phase-space integration is needed. The MC method is an integration tool to estimate the area bounded by any function employed for such calculations. The event generator like MADGRAPH produces a list of final state particles, and may consist of

quarks (has strong charge). PYTHIA 8 takes these particles provided by the MADGRAPH and produces color singlet states through hadronization and fragmentation. The generator models also take into account not only the hard scattering, but the initial state radiation or final state radiation, multi-parton interactions (MPI), and avoiding double counting of jets initiated by matrix level elements (ME) or parton showering (PS) by taking different measures. Even the generator can be tuned to set values to parameters that can not be determined from first principles in the event generator programs.

VLL Singlet samples are generated using MADGRAPH5_AMC@NLO [36] at LO precision. The production cross-section for the VLL signal model is calculated at next-to-leading order (NLO) precision [37]. A single lepton requirement was imposed to restrict the phase space of event generation aligned with the analysis trigger requirement using the PYTHIA filter while producing the VLL-tau like samples. Table 4.1 shows the cross-section and single lepton filtering efficiency for all data-taking eras, used in this analysis for different VLL mass hypotheses. No such filter was applied in the central production of VLL-muon like samples.

VLL Mass	2018	2017	2016preVFP	2016postVFP
100	0.4369	0.4369	0.4385	0.4385
125	0.5023	0.4877	0.4995	0.4995
150	0.5325	0.5383	0.5315	0.5295
200	0.5634	0.5665	5643	0.5673
250	0.5836	0.5857	5839	0.5836
300	0.5928	0.5937	0.6003	0.5972
350	0.5996	0.6122	0.5985	0.6101
400	0.6167	0.6175	0.6159	0.6156

TABLE 4.1: Single lepton filtering (in PYTHIA) efficiency for different VLL-tau mass hypotheses and years used in this analysis.

All background and signal samples in 2016 are generated with the NNPDF3.0 NLO or LO parton distribution functions (PDFs), with the order matching that in the matrix element calculations. In 2017 and 2018, the NNPDF3.1 next-to-next-to-leading order PDFs [38, 39] are used. Parton showering, fragmentation, hadronization, and the decay of unstable particles for all samples are performed using PYTHIA 8.230 [40] with the underlying event tune CP5 [41]. Double counted partons generated with PYTHIA and MADGRAPH5_AMC@NLO are removed using the FxFx [36, 42] or MLM [43] jet matching schemes.

4.1.2 Simulation

The list of particles produced at the last step of event generation models is now propagated through the detector. The particles interact with the sensitive detector layers, interact with

the detector material, lose their energy through electromagnetic interactions or interacting with the atomic nuclei, or can decay into other particles. The interaction of the particles with the detector material is modeled using the GEANT4 toolkit [44]. The presence of multiple proton-proton interactions in the same or adjacent bunch crossing (pileup) is incorporated by simulating additional interactions that are both in-time and out-of-time with the hard collision according to the pileup in the data samples. This detailed first principle-based simulation with intricate details of the detector geometry, shape, material (type of material) budget, etc, is commonly called **Full Simulation**. They are used to generate events for almost all background events and most signal samples in BSM searches. All MC samples in this thesis used Full Simulation software (in CMSSW) to simulate the detector response.

Full Simulation takes around 100 seconds to produce a full $t\bar{t}$ event. It is computationally costly to simulate billions of events needed to ensure good statistical precision of the simulated samples to compare with the huge volume of data taken at LHC. Future operations like high luminosity HL-LHC will take 20 times more data than what has been collected. Naturally, the need for a faster yet accurate version of the simulation chain is increasing [45]. There exists detector simulation software like Delphes, which uses parameterized detector response, but it is far from being accurate. Competitive machine learning models are being developed to accomplish this task to achieve accuracy as compared to the Full Simulation, but 1000 times faster using generative AI, which is a hot area of current HEP research. Another competitive version of a faster event production chain exists in CMS called Fast Simulation, which is described next.

4.2 Fast Simulation

The Fast Simulation (FastSim) is built around some simplified assumptions on the detector geometry, particle propagation in the tracker layers, and particle-matter interactions in calorimeters or tracker materials [46].

The input to Fast Simulation consists of a collection of particles, typically produced by an event generator. Each particle is defined by its momentum and production vertex and includes information about its parent and daughter particles to trace the complete decay chain within the event. As the quasi-stable particles are propagated through the detector, they can decay based on their known branching ratios and decay kinematics. Any new particles produced from interactions with detector material or from decays during flight are added to the original particle list and undergo the same propagation and decay procedures.

The key modifications in simulating detector response compared to the Full Simulation are [47, 48]:

- The **tracker geometry** is reduced to simplified cylindrical layers and planes, interleaved with non-instrumented cylinders with dead material (cables, support, etc.). The material is assumed to be uniformly distributed over each cylindrical barrel and endcap disk part. The complete magnetic field map is used for the track propagation between two surfaces. Charged particles interact with the tracker materials and lose their energy through multiple scattering and ionization. The number of Bremsstrahlung photons produced by the charged particles is parameterized with respect to the Full Simulation, and the thickness of each tracker layer is tuned in terms of radiation lengths. The crossing point between a particle and tracker layer is defined as a *simulated hit*. Reconstructed hits are produced by introducing smearing to these simhits. For simulated hits in pixels, a full simulation-based smearing is used that takes into account the pixel cluster size and incident angle of the particles. Simulated hits in the strip are smeared using a Gaussian smearing profile (in transverse and longitudinal direction to the beam) obtained from the Full Simulation.

- **Calorimeter response** of the particles is parameterized using analytical functions. For example, showers of electrons or photons in the ECAL are simulated using the Grindhammer parameterization [49]. HCAL energy response is simulated using the Full Simulation information of the response of charged pions to HCAL. This smeared energy is then distributed in the calorimeters using parameterized longitudinal and shower profiles, with shower-to-shower variations, following an approach similar to that of *GFLASH* [50].

- Similarly, for muons, they are propagated as a charged particle in the inner tracker as described before. For muons, only the multiple scattering and energy loss by ionization are taken into account. Full geometric descriptions of the muon chambers are taken from the actual geometry database, and simulation hits are positioned in the detector whenever the muon trajectory crosses an active layer of those chambers. Then, it follows the full simulation chain from digitization to the final physics object step.

- **Fast Tracking**: These simplifications in the detector response and geometry provide a significant gain in event reconstruction time using Fast Simulation. Another crucial aspect of its faster event reconstruction is in the tracking modification. FastSim tracking is simplified using generator-level truth information about a particle trajectory and using the reconstructed hits compatible with simulated tracks to make track candidates. This truth information-aid tracking drastically reduces the tracking time, avoiding the computationally heavy hit combinatorics and standard pattern recognition algorithms. Each reconstructed track corresponds to a simulated track in Fast

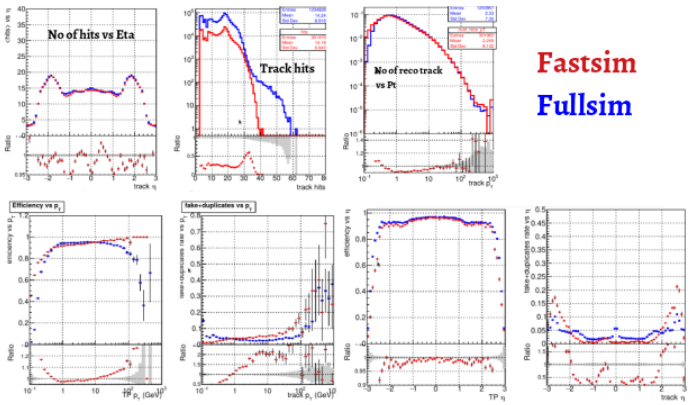


FIGURE 4.2: Comparison between Fast Simulation and Full Simulation for Run-2 tracking validation in simulated $t\bar{t}$ events.

Simulation. The downside of this is that there are no fake tracks in FastSim. For example, high-occupancy events may have fake tracks, which the Fast Simulation does not produce. Hit sharing between different tracks is also not included in the Fast Simulation.

Using all the described simplifications, FastSim is faster than FullSim by a factor of 100 in the simulation step, and overall 20 times faster considering the complete event production chain. Figure 4.2 shows the tracking validation between FullSim and FastSim using $t\bar{t}$ events in the zero pile-up case for the Run 2 tracker geometry. However, some complex physics observables such as discriminants associated with flavor tagging, jet substructure, etc, are not modeled well in FastSim. Some recent efforts of using regression ML techniques [51] to refine such high-level analysis observables resulted in improved agreement with the FullSim, and correlation amongst the variables is also retained, which is crucial for full event description. Details of the Phase 2 tracker implementation and relevant tracking techniques will be discussed next.

4.3 Phase 2 tracker geometry implementation in FastSim

The Phase-2 tracker upgrade [52, 53] of the CMS detector is designed to handle the extreme conditions expected at the High-Luminosity Large Hadron Collider (HL-LHC). With the HL-LHC projected to begin operation around 2029, the CMS detector must withstand a significantly increased instantaneous luminosity, leading to high pileup scenarios (about 200 collisions per bunch crossing). This upgrade ensures precise tracking performance and robust reconstruction of physics objects under these conditions.

4.3.1 Phase 1 and Phase 2 tracker geometry

The current tracker system, including the silicon pixel and strip detectors, will be replaced with a new, highly granular, radiation-hard design. The upgraded tracker will provide extended geometrical coverage up to $|\eta| < 4$. This will significantly enhance the capability to reconstruct particles in the forward region, which is crucial for searches involving boosted objects and forward jets, such as vector boson fusion or t-channel single production processes. The Phase 2 tracker can be divided into two main parts:

- **Inner Tracker:** In the central region, the inner tracker has four cylindrical layers made of pixel modules which are named as Tracker Barrel Pixel Modules (**TBPX**) and eight small plus four large disc-like structures in each forward direction named as Tracker Forward Pixel Modules (**TFPX**) and Tracker Endcap Pixel Modules (**TEPX**), respectively. High granularity of macro-pixel modules in the inner tracker ensures better detector resolution, occupancy, two-track separation, and the pixel-based track seeding, which is an important part of the iterative tracking, ensures a good track finding performance even in low transverse momentum and in a high pile-up environment($\approx$ 140).

- **Outer Tracker:** Outer tracker is made of 5 endcap modules which are rings on disc like structures (Tracker Endcap Double Disc or **TEDD**) and outer barrel part which are further divided into three layers made of pixel strip modules (**TBPS**) in the radii region of 200-600 mm and three layers in the radial region above 600mm made of double strip modules (**TB2S**).

Figure 4.3 illustrates the new phase-2 tracker layers compared to the current ones. To reduce multiple scattering and energy loss, the tracker is designed with lightweight and low-density materials, minimizing the material budget within the tracking volume.

I implemented the phase 2 tracker geometry in the Fast Simulation event production chain and phase 2 tracking steps are also accommodated in FastSim tracking to take advantage of the new geometry.

4.3.2 Phase 2 tracker implementation in FastSim

Extracting layer parameters from Full Simulation

Before implementing the simplified phase 2 geometry model in Fast Simulation, a detailed study on the existing Phase 2 Full Simulation was needed. The motivation behind the detailed investigation is as follows:

- Technical Design Report of phase 2 tracker upgrade [52] had outdated tracker geometry at the time of implementation. As the phase 2 geometry development in Full

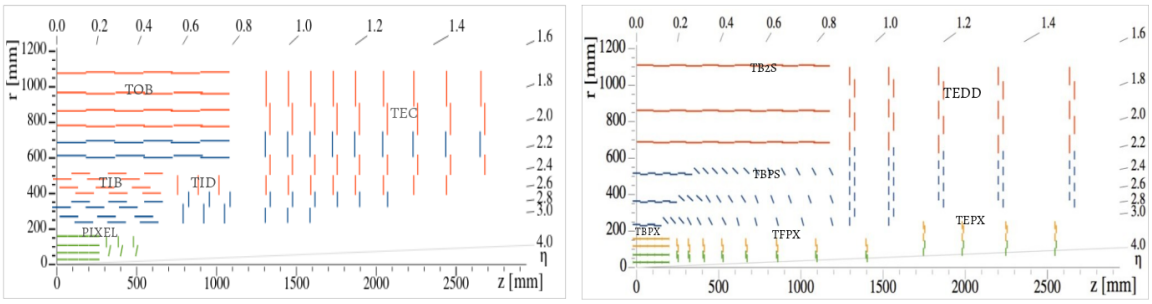


FIGURE 4.3: R-z view of one quadrant of the CMS phase 1 (left) and phase 2 (right) tracker geometry [52].

Simulation is ongoing, the latest and greatest layer parameters are required to correctly implement the geometry in Fast Simulation without any compatibility issue with the Full Simulation geometry (detector IDs).

- Find the appropriate Full Simulation C++ classes that should be used to create a simplified geometry in the Fast Simulation framework. The implementation can be generalized in this way and can be adopted quickly to any new changes to the FullSim tracker geometry.

Implementing the geometry

After extracting the properties of the layers from Full Simulation, the geometry was defined in the Fast Simulation Tracker Material Configuration File. During the implementation of geometry, a few things were taken care of.

- Layers were divided into sensitive layers and insensitive layers or dead materials (Cable wires or gaps in the detector), and the order of tracker layers is maintained.
- Particle interaction with tracker materials is defined by assigning interaction lengths in each layer. The same interaction length is used to simulate particle interaction with tracker materials for newly added layers.
- Only the tracker part is simulated, leaving other sub-detector volumes untouched.
- Order of the layers was crucial, as the software implementation should be exact as a particle traverses through tracker layers in the real detector.

The r-z plot of phase-2 tracker using simulated hits in Figure 4.4 shows that the geometry is correctly defined in Fast Simulation with the correct order of layers and no missing layers. The simulated hits are now used to produce reconstructed hits using the smearing profile as described earlier and fed into the track reconstruction after the digitization step is completed.

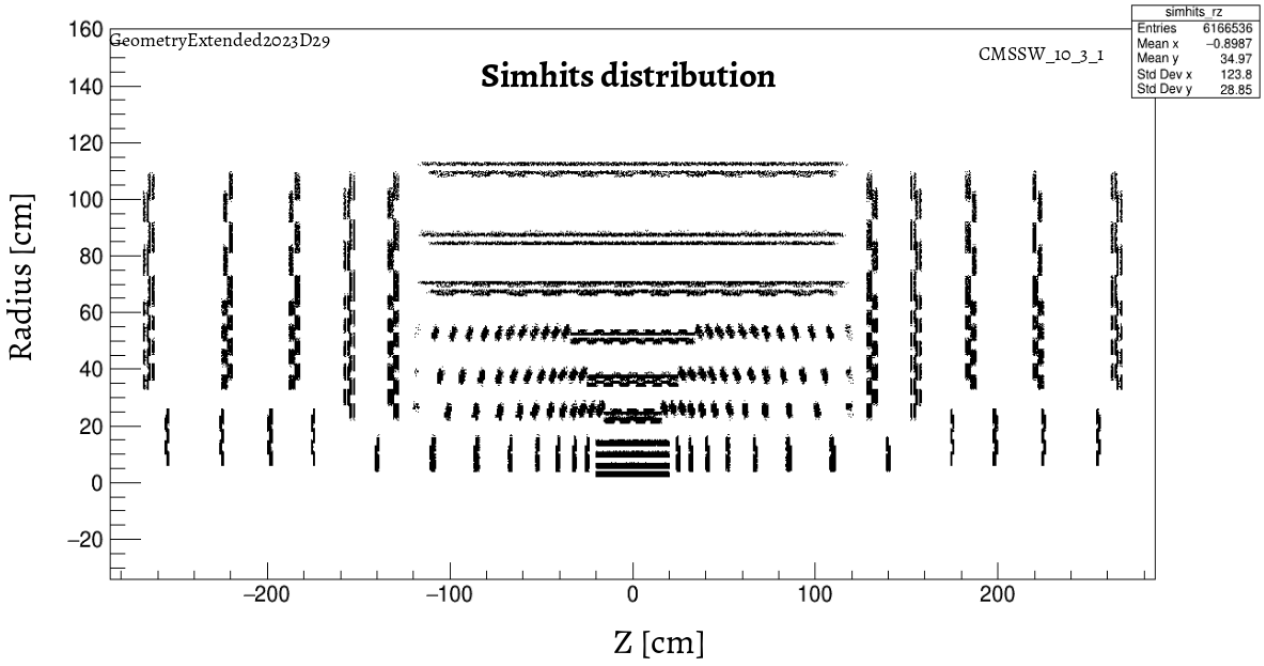


FIGURE 4.4: R-z view of the CMS phase-2 tracker using simulated hits produced with 2000 $t\bar{t}$ events using Fast Simulation chain with phase-2 geometry: GeometryExtended2023D29 and detector condition: phase2 realistic, implemented using CMSSW package.

Figure 4.5 illustrates the key differences in the tracking steps between phase-1 and phase-2 iterative tracking algorithms. The iterative tracking is described in Section 4.4.1.

FastSim phase-2 tracking performance is compared in the track-only validation settings (as other subdetector implementations are not ready) with Full Simulation phase-2 tracking and demonstrated in Figure 4.6. FastSim tracking efficiency at the forward region is poor due to the improper hit resolution histograms used to smear the hits in pixels, as the pixel smearing profile (derived from FullSim) is not available and currently being extracted using a detailed PIXELAV simulation that requires the latest tracker front-end electronics details and sensor description.

4.4 Physics objects reconstruction and identification

The CMS detector is designed to identify various particles passing through it. Charged particles passing through the silicon tracker layers produce a signal (hit) and bend in the magnetic field as they traverse the tracker layers. This allows us to measure their electric charge and momentum, and reconstruct the trajectories (tracks) and origins (vertices). Electrons and photons deposit their energy in the ECAL via electromagnetic interactions. Charged and neutral hadrons may initiate a hadronic shower in the ECAL and deposit energy in the HCAL. For both calorimeters, the energy deposition (showers) is measured to estimate their

Iteration	Name	Seeding	Targeted Tracks
1	InitialStep	pixel triplets	prompt, high p_T
2	DetachedTriplet	pixel triplets	from b hadron decays, $R \lesssim 5$ cm
3	LowPtTriplet	pixel triplets	prompt, low p_T
4	PixelPair	pixel pairs	recover high p_T
5	MixedTriplet	pixel+strip triplets	displaced, $R \lesssim 7$ cm
6	PixelLess	strip triplets/pairs	very displaced, $R \lesssim 25$ cm
7	TobTec	strip triplets/pairs	very displaced, $R \lesssim 60$ cm
8	JetCoreRegional	pixel+strip pairs	inside high p_T jets
9	MuonSeededInOut	muon-tagged tracks	muons
10	MuonSeededOutIn	muon detectors	muons

Iteration	Step name	Seeding configuration	Target track
0	HighPtQuadruplet	pixel quadruplets	prompt, high p_T
1	HighPtTriplet	pixel triplets	prompt, high p_T
2	LowPtQuadruplet	pixel quadruplets	prompt, low p_T
3	LowPtTriplet	pixel triplets	prompt, low p_T
4	DetachedQuadruplet	pixel quadruplets	displaced
5	PixelPair	pixel pairs	high p_T recovery
6	Muon inside-out	muon-tagged tracks	muon tracks
7	Muon outside-in	muon-tagged tracks	muon tracks

FIGURE 4.5: Key differences in phase-1 (upper) and phase-2 (lower) iterative tracking steps. [52, 29]

energies and directions. Muons produce a signal in the inner silicon tracking system and leave hits in the muon system. Neutrinos escape the detector without interacting or producing a signal. Conceptually, various subdetectors of the CMS experiment can individually reconstruct particles interacting with it. Muons can be reconstructed and identified using the information from the muon chambers, photons or electrons can be identified from their ECAL shower profile, or jets consist of hadrons and photons can be identified from ECAL and HCAL energy deposit, and a complete description of the event can be built. Figure 4.7 shows schematically the signatures of different particles in the various subdetectors of CMS.

CMS uses a holistic approach by combining information from various subdetectors to build a significantly improved global event description. This is called the particle-flow (PF) reconstruction algorithm [54]. The success of the PF approach is tied to the fine spatial granularity of the detectors and the large magnetic field. Combining the different subdetector measurements improves energy determination and resolution of the reconstructed particles. Basic elements of the PF algorithm are tracks, vertices, and calorimeter clusters. They can be utilized to reconstruct complex objects as follows,

- **Electrons**: linking track and ECAL cluster with no connection to HCAL cluster.
- **Muons**: link between inner track and muon chamber track
- **Photons and neutral hadrons** (eg, π^0 or K_L^0): ECAL and HCAL cluster with no track link.
- **Charged hadrons** (eg, $\pi^{\pm}$): ECAL and HCAL cluster with a track link.

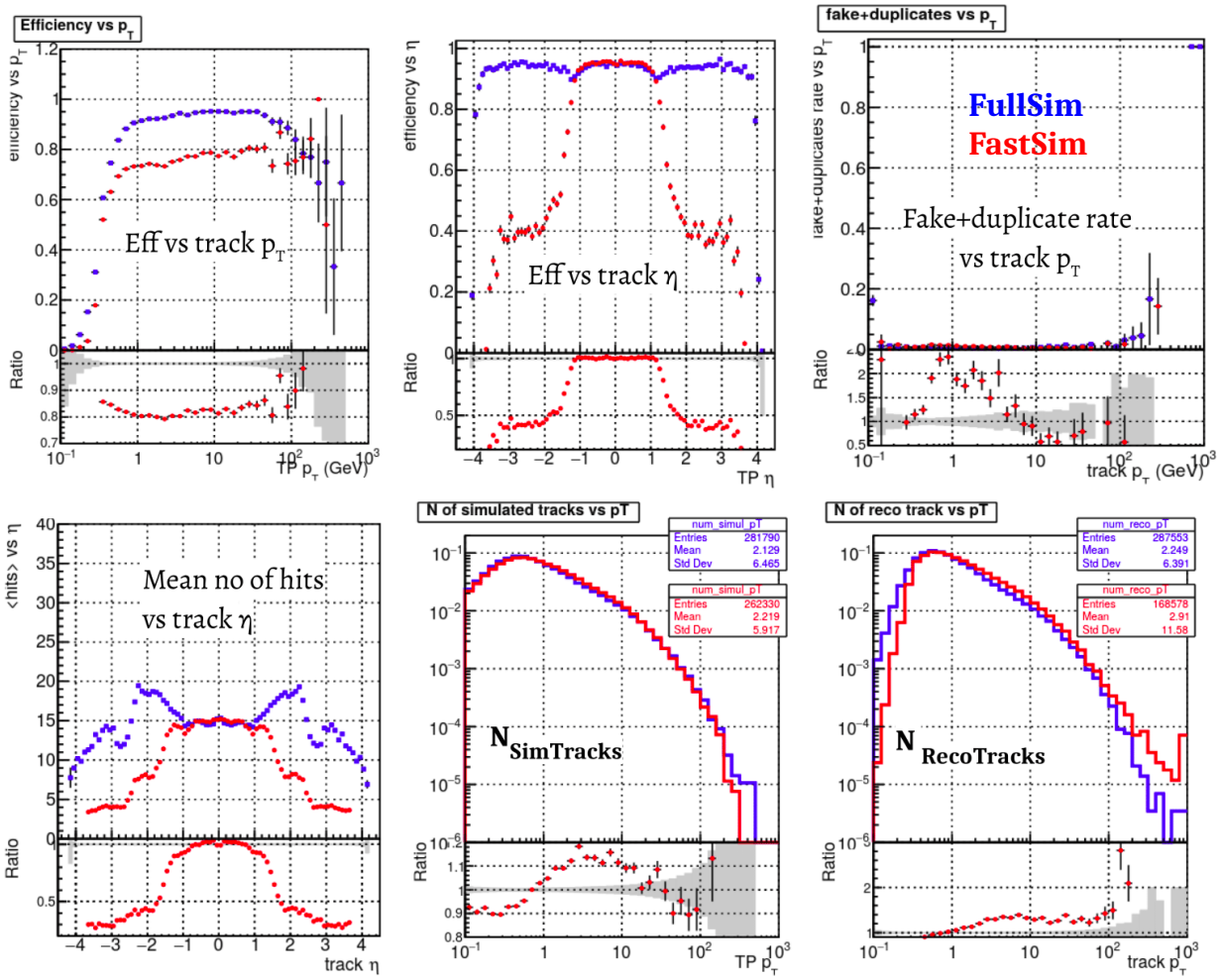


FIGURE 4.6: Phase-2 FastSim (red) vs phase-2 FullSim (blue) track validation performance in $t\bar{t}$ events.

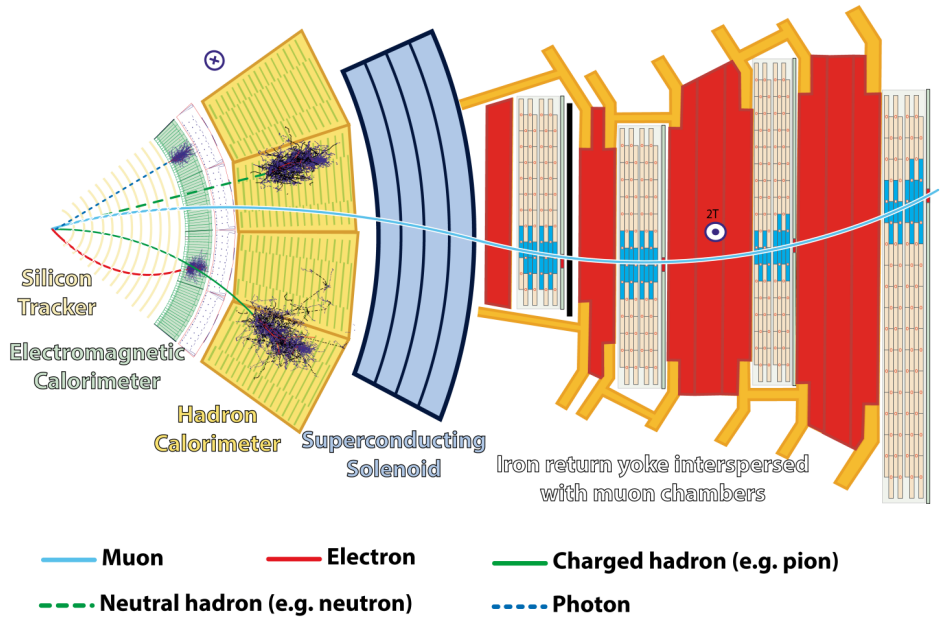


FIGURE 4.7: A transverse slice of the CMS detector and the particles detected by each subdetector.

Putting additional constraints on the basic PF building blocks, for example, a momentum-to-energy ratio close to unity for the electron, can reduce unwanted candidates for reconstruction purposes.

4.4.1 Tracks and vertices

All the charged particles produce an electronic signal in the silicon sensors of the tracker layers, which are called *hits*. Tracking algorithms use these hits to produce particle tracks in three steps:

- *Seeding:*Initial seed generation with compatible hits.
- *Track building:*Trajectory building or pattern recognition to extend the track using more hits from other layers.
- *Track fitting:*Final fitting to measure the charge, origin, momentum, or direction of the track.

Tracking algorithms are designed to correctly identify charged particle tracks with a low misidentification rate (high purity). A traditional single-step approach is hard to achieve high purity and reconstruction efficiency simultaneously, especially in a hadron collider due to the high multiplicity of charged particles. This may result in the degraded energy resolution of jets in downstream algorithms.

The CMS tracking algorithm thus takes an iterative strategy to reconstruct tracks based on their complexity. Each iteration of the iterative tracking algorithms proceeds in the three-step approach of seed generation, pattern recognition, and track fitting, with a few changes in the quality of seeds or tracks. This algorithm is also known as Combinatorial Track Finding or CTF. The basic idea is first to reconstruct the tracks easier to find (e.g., relatively large p_{T}, produced near the interaction point), remove the used hits for the next iteration to reduce combinatorial complexity, and simplify subsequent iterations that search for low p_{T}, largely displaced tracks. The tracks from the first three iterations are seeded with triplets of pixel hits with an overall track efficiency of 80%. The fourth and fifth iterations aim at recovering tracks with one or two missing hits in the pixel detector. The rest of the iterations are designed to reconstruct large displaced tracks, high-p_{T} jets with merged hits, and muon tracks reconstruction. The iterative tracking approach enables the reconstruction of tracks of particles with p_{T} 200 MeV to 1 TeV. After the reconstruction of tracks, some selected tracks with specific quality are clustered together based on their z-coordinates at their point of closest approach to the centre of the beam spot. This clustering allows for the reconstruction of the possible vertices (pile-up and hard interaction) in the same event and then feeds them into the adaptive vertex fitter [55] algorithm to reconstruct the primary vertex. The quality criteria of tracks and details of this multi-step process are described in [31].

4.4.2 Muons

Muon produces hits in the inner tracker and muon chambers. Both detectors can reconstruct muons with high efficiency and purity, with precise momentum measurement [56]. In the DT and CSC system, straight line track segments are built from the reconstructed hits in multiple layers. RPC hits are reconstructed by clustering the hit strips. These are used in the next step of the process. Three types of muon candidates are defined based on the detector system used to reconstruct them,

- standalone muon track is seeded with the line segment from the DT or CSC detector, and the trajectory is built further with all DT, CSC, and RPC chamber hits with fitting.
- global muons are constructed in *outside-in* approach by matching each standalone-muon track to a track in the inner tracker (propagated on a common surface), and fitted with the Kalman filter using information from both the tracker track and standalone-muon track.
- tracker muons are built *inside-out* by extrapolating an inner track to the muon system that matches with at least one DT or CSC muon segment.

These reconstructed muons (standalone, tracker, or global muons) candidates are fed into the PF algorithm with some selections imposed based on the category, and identified accordingly. The PF elements that make up these identified muons are masked against further processing in the corresponding PF block, i.e., are not used as building elements for another particle.

4.4.3 Electrons

Electron reconstruction combines information from the inner tracker and the calorimeters. Electrons often emit bremsstrahlung photons, and photons often convert to e^+e^- pairs, which in turn emit bremsstrahlung photons. The electron EM shower in the ECAL may spread, particularly in the azimuthal direction (ϕ). The energy of the electron and possible bremsstrahlung photons is collected by grouping the ECAL clusters via clustering algorithms (such as the Hybrid Algorithm in the barrel and Multi-5x5 Algorithm in the endcap region) to superclusters, which represent the total electron energy. Further energy corrections are applied to these clusters to account for the energy loss, non-uniform response, and shower leakage. This ECAL-based electron seeding strategy is efficient for high-energy isolated electrons, but not efficient for electrons in the non-isolated environment (inside jets) and low p_{T} electrons. For electrons in jets, the position and energy of the associated supercluster are often biased due to the limitation in distinguishing overlapping energy depositions from other particles and multiple compatible hits in the inner tracker layers when the superclusters are backpropagated to the interaction region. Low p_{T} electrons significantly bend in the magnetic field, and the radiated energy is spread over such an extended region that the supercluster cannot include all deposits. A tracker-based electron seeding is developed to reconstruct the electrons missed by the ECAL-based strategy. To account for the energy loss due to bremsstrahlung photons, a Gaussian Sum filter (GSF) algorithm runs on the preselected tracks to produce GSF tracks. The GSF fitting allows for sudden and substantial energy losses along the trajectory that improve the accuracy of electron momentum estimation, which is needed further for establishing a link between tracks and ECAL supercluster. No such link is sought for photons between a GSF track and a supercluster. Additionally, a small HCAL energy requirement is imposed to distinguish electrons and photons from the hadronic backgrounds.

The final momentum of the electron candidates is estimated using information from the tracker and ECAL. The tracker measurements are more precise due to low bremsstrahlung radiation for the low momentum electrons ($p_{\mathrm{T}} < 15$ GeV). In contrast, the ECAL energy measurements are crucial for the high-momentum electrons [57, 58].

4.4.4 Jets

Jets are collimated sprays of color-singlet particles originating from the hadronization of high-energy quarks and gluons produced in proton-proton collisions. Since individual quarks and gluons cannot be directly observed due to QCD confinement, their energy and momentum are inferred from the reconstructed particles in the detector. The collimated spray of hadrons appears as a cluster of energy deposited in the localized area of the detector. Charged hadrons carry most of the energy in a typical jet ($\sim$60%), photons (coming from the decay of π^0) carry $\sim$ 25% while neutral hadrons carry about $\sim$10%. There are many techniques to associate the detector signatures of the particles by forming a jet.

A key parameter of any jet clustering algorithm is jet size. The choice of jet radius affects its sensitivity to soft radiation. A larger jet radius helps capture more hadronized particles, improving the accuracy of jet mass and energy measurements. In comparison, a smaller jet radius is useful to reduce the underlying event (UE) and pileup contributions. The jet clustering algorithms should be resilient to a soft (low energy) gluon emission or from high-energy parton splits into two nearly collinear partons (e.g., a quark emits a gluon in almost identical direction or a gluon producing a quark-antiquark pair). These infrared and collinear (IRC) safe algorithms are crucial to measure jet properties that are compatible with theoretical QCD calculations and robust against random fluctuations affecting the jet multiplicity and kinematics. CMS uses a number of IRC-safe algorithms to reconstruct jets of various radius, such as the SISCone [59], Cambridge/Aachen [60],k_T [61], and anti-k_T [62].

Jets clustered using the anti-k_T algorithm on the particle flow candidates (PF) with radius parameter $R = 0.4$ are used in this thesis. This algorithm combines two particles based on the notion of two distances. The first distance variable computes distance between two particles following Equation 4.1, p_{T} is the transverse momentum of the particles, $R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$ is the euclidean distance of two particles in $\eta - \phi$ space and R is the radius parameter. The second distance variable is the momentum space distance between the beam axis and the particle (d_{iB}) as written in Equation 4.2.

$$d_{ij} = min(\frac{1}{p_{\mathrm{T}i}^2}, \frac{1}{p_{\mathrm{T}j}^2}) \times \frac{R_{ij}^2}{R} \tag{4.1}$$

$$d_{iB} = \frac{1}{p_{\mathrm{T}i}^2} \tag{4.2}$$

Starting from a given i^{th} object, the combination algorithm searches for another object j, such that $d_{ij} < d_{iB}$. If such a j^{th} object is found, the algorithm combines these two objects and continues adding more particles. It stops if no additional object is found satisfying $d_{ij} < d_{iB}$, and the modified i^{th} object is called a jet, and the constituent particles

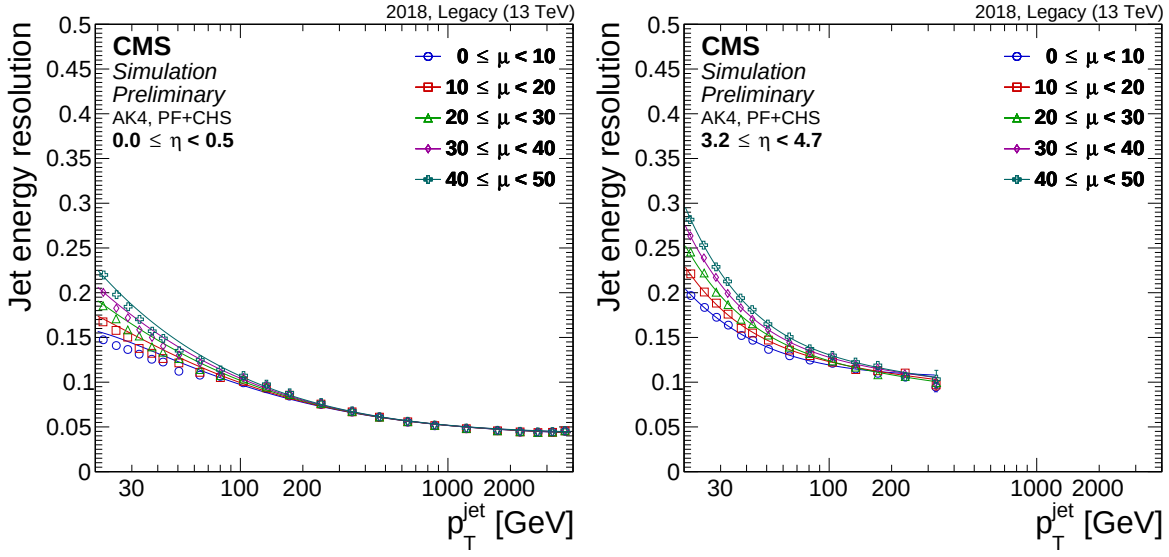


FIGURE 4.8: Jet energy resolution vs jet p_{T} for 2018 in two η ranges. Figures taken from [68].

are removed from the set for further iterations. The resulting jets are then corrected based on a detailed MC simulation of the detector, pile-up, detector response to hadrons, flavor of the originating quark, and residual corrections to take into account the differences between data and simulation as a function of jet p_{T} and η [63, 64, 65]. In particular, for the pile-up corrections, charged hadrons that originate from the pileup vertices are removed from the jet constituents using the Charged Hadron Subtraction method (CHS) before carrying out other pile-up mitigation techniques to remove neutral hadron contamination [66, 67]. We refer to the jets used in this analysis as AK4 CHS PF jets. Jet energy resolution as a function of jet p_{T} for two different η ranges is shown in Figure 4.8 for 2018.

4.4.5 Missing transverse energy

Neutrinos or other hypothetical weakly interacting particles (such as dark matter) do not leave any signature in the detector. However, their presence can be inferred from the momentum imbalance in the plane perpendicular to the beam axis (z-axis), known as Missing Transverse Momentum or Energy (MET). It is also labeled as E_T^{miss} or p_T^{miss} in the literature. MET is defined as the negative vector sum of the transverse momenta of all the particle flow candidates in an event, as shown in Eq. 4.3.

$$\vec{p}_T^{\,miss} = -\sum_{i}^{PF} \vec{p}_{Ti} \tag{4.3}$$

Another algorithm called *pileup per particle identification* (PUPPI) [66] method is developed (using PF candidates) to reduce the effect of pile-up in jets and $p_{\mathrm{T}}^{\mathrm{miss}}$ observables.

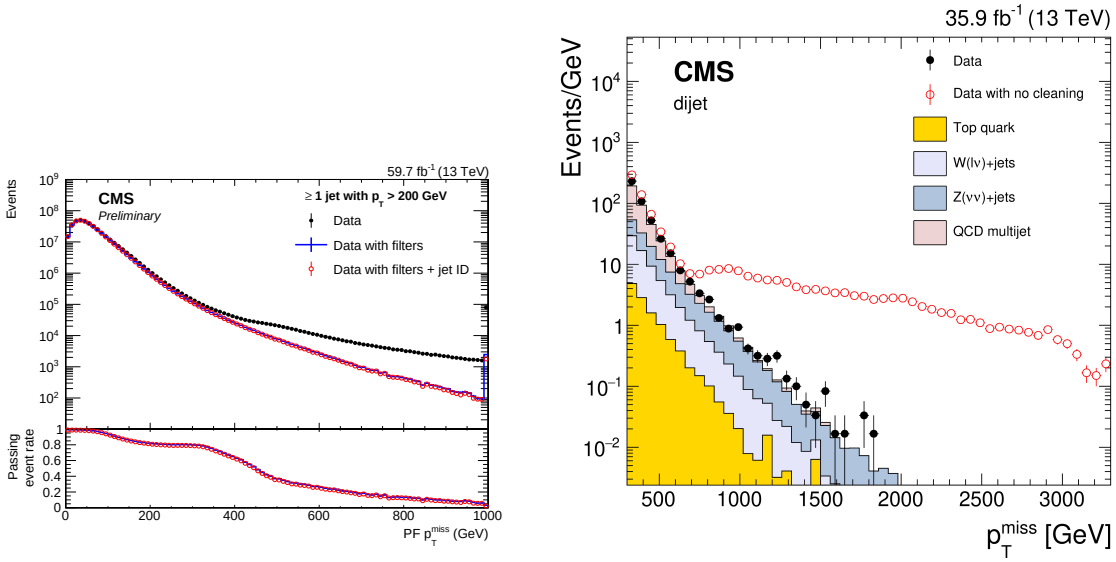


FIGURE 4.9: Particle flow $p_{\mathrm{T}}^{\mathrm{miss}}$ distribution in events with at least one jet with $p_{\mathrm{T}} > 200$ GeV for 2018 collision data (left) and in dijet events (right) before and after applying various $p_{\mathrm{T}}^{\mathrm{miss}}$ filters. Figures taken from [70] (left) and [69] (right).

Particle Flow MET is used in this thesis. MET reconstruction is sensitive to detector malfunctions and mis-measurement of visible particle momenta (due to the non-linear calorimeter response to hadrons, minimum energy threshold in calorimeters, etc). The estimation of $p_{\mathrm{T}}^{\mathrm{miss}}$ is improved by propagating the corrections of jets p_{T}, jet energy resolution, etc. Anomalous high p_{T} miss events due to reconstruction failure or malfunctioning detectors are also masked due to event-level filters as described in [69]. The improved modeling of $p_{\mathrm{T}}^{\mathrm{miss}}$ in data after masking such spurious events is illustrated in Figure 4.9.

4.4.6 Taus

Tau lepton with a mass of 1.77 GeV can decay into hadrons and a neutrino. About one-third of the time, tau leptons decay into an electron or a muon and two neutrinos. The neutrinos escape undetected, but the electrons and muons are reconstructed and identified through the techniques discussed earlier. These decay final states are denoted as τ_e and τ_μ, respectively. Most tau decays contain charged and neutral mesons, and a tau-neutrino. These hadronic decays are collectively referred to as τ_{h} decays. The most frequent τ_{h} decays involve either one or three charged hadrons, known respectively as 1-prong and 3-prong decay modes.

At CMS, the reconstruction of τ_{h} candidates is performed using the Hadrons-Plus-Strips (HPS) algorithm [71]. The HPS algorithm reconstructs the τ_{h} candidates by combining information from charged hadrons and neutral pions. The neutral pions decay into two photons, which can convert into electron-positron pairs in the tracker. These photon and electron candidates are then clustered in regions of the ECAL defined in terms of pseudorapidity (η)

and azimuthal angle (ϕ). These clustering regions are referred to as "strips". This characteristic energy deposition in narrow regions of (η, ϕ) is used to distinguish these τ_{h} candidates from the quark or gluon jets. Further improvement in the HPS algorithm is carried out by dynamically changing the strip size to contain the ECAL energy deposition. Electrons or muons could also be reconstructed as τ_{h} candidates. Genuine hadronic taus are identified using machine learning techniques against the jets, electrons, or muons that can mimic the signature of taus in the detector. For example one of the algorithms in Run-2, DEEPTAU, is a convolutional neural network that is used to identify hadronic taus by combining information from individual reconstructed particles (with PF objects) near the τ_{h} axis with information about the reconstructed τ_{h} candidate and other high-level variables [72]. The identification efficiency is roughly 50% for a jet misidentification probability of 1%.

4.5 Analysis level selection

Electrons or muons originating from SM gauge bosons (W, Z, or Higgs), leptonic decays of tau, or from the vector-like leptons used in this thesis are energetic, relatively isolated, and close to the primary vertex due to extremely small lifetimes of their mother (or source) particles. This feature is exploited in the light lepton identification criteria to reduce backgrounds that produce non-isolated (leptons inside a jet due to semileptonic decay of hadrons), displaced leptons (with sufficiently long lifetimes of a few hadrons).

Muons with $p_{\mathrm{T}} > 10$ GeV and $|\eta| < 2.4$ are chosen for this analysis. The η range ensures the muon is in the tracker acceptance. A delta-beta corrected relative PF isolation of a maximum 15% is allowed for the muons. The relative isolation is defined as the scalar p_{T} sum of neutral and charged hadrons or electromagnetic particles (after removing the target object) within a cone of radius $\Delta R = 0.4$ around the muon, normalized to the muon p_{T}. Additional corrections due to pile-up particles inside the cone are also accounted for in the isolation variable. Muons are required to satisfy the promptness criteria of $|d_{\mathrm{z}}| < 0.1$ 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 primary vertex of the event. In addition, muons must pass the *medium-ID* identification criteria for the analysis requirement. These IDs are typically a collection of variables satisfying some threshold value to suppress various kinds of backgrounds as described for the medium-ID in the Table 4.2. Muon reconstruction and identification efficiency is $> 96\%$ for the quality criteria imposed in this analysis. Although this medium identification criteria has similar efficiency to select muons that originate from the heavy flavor decay. Promptness criteria may reduce these muons as they are displaced due to slightly longer-lived b-hadrons. Additional selection on the significance of 3D impact parameter (SIP3D) and the probability of the closest jet (associated with muon) originates

Cuts	Requirements
PF Muon	Yes
Global or Tracker Muon	Yes
Fraction of valid inner tracker hits	< 20
If global muons	
Quality of track (normalized χ^2)	< 3
Tracker-standalone position χ^2	< 12
Number of kinks in the track	< 20
Compatibility of inner track and muon segment	> 0.303
If not global muons	
Compatibility of inner track and muon segment	> 0.451

TABLE 4.2: Variables and cuts used to make the medium identification criteria for muons.

from b-quark hadronization (LeptonDEEPJET score), which are used further to distinguish prompt isolated muons from non-prompt non-isolated muons. The SIP3D variable is the ratio of the 3D impact parameter and associated uncertainty. The closest jet to the muon is assigned by matching any jets (with p_{T} >10 GeV) that fall within the $\Delta R = 0.4$ of the muon, and tagged with the DEEPJET algorithm described later in this section. SIP3D $< 10, 12, 9$ and LeptonDeepJetScore $< 0.6, 0.4, 0.3$ are imposed on the muons for 2016, 2017, and 2018, respectively. In the rare occurrence that no or multiple such jets are found per lepton, DEEPJET score is set to be identically zero or the closest jet is utilized, respectively. Although for the μjj analysis, the impact of such cuts is less due to the low multiplicity of muons in the final state, these cuts are chosen to be consistent with a parallel multilepton analysis for combination purposes in the future.

Electrons are not the primary object in this analysis, but are used to veto the presence of any extra electrons in the event. Electrons are selected with $p_{\mathrm{T}} > 10$ GeV and $|\eta| < 2.4$ satisfying the medium identification criteria. Since a p_{T} parameterized isolation cut is already inbuilt in the identification criteria, we select electrons by relaxing the relative PF-isolation of max. 100% (with a cone size of 0.4 around the electron). A list of variables in the medium electron ID is summarized in Table 4.3. In addition, they must satisfy $|d_{\mathrm{z}}| < 0.1$ cm and $|d_{\mathrm{xy}}| < 0.05$ cm in the ECAL barrel ($|\eta| < 1.479$) and $|d_{\mathrm{z}}| < 0.2$ cm and $|d_{\mathrm{xy}}| < 0.1$ cm in the ECAL endcap. The same set of selection criteria on SIP3D and DEEPJET score is applied to the electrons. All selected electrons within a cone of $\Delta\mathrm{R} < 0.05$ of a selected muon are discarded to suppress contributions due to bremsstrahlung from muons, where $\Delta\mathrm{R}$ is the distance between a given pair of objects in the $\eta - \phi$ plane. Electrons satisfying the criteria mentioned above are part of the **Loose electron** collection.

Similarly, **Loose muons** collection is constructed with similar criteria as described before, but with looser relative PF-isolation of maximum 100% (rel.PFIsolation< 1.0). The combined collection of loose muons and loose electrons is called loose lepton collections,

TABLE 4.3: Variables and cuts for medium electron identification requirements.

Variable	Definition	$\lvert\eta_{SC}\rvert \leq 1.479$ (> 1.479)
full5x5_$\sigma_{i\eta i\eta}$	Shower shape	<0.0106 (<0.0387)
$\lvert\Delta\eta_{seed}\rvert$	$\lvert\Delta\eta_{SC,track}\rvert$at closest approach	<0.0032 (<0.00632)
$\lvert\Delta\phi_{In}\rvert$	$\Delta\phi_{SC,track}$at closest approach	<0.0547(<0.0394)
H/E	HCAL/ECAL energy	$< 0.046 + 1.16/E_{SC} + 0.0324 * \rho/E_{SC}$ ($< 0.0275 + 2.52/E_{SC} + 0.183 * \rho/E_{SC}$)
relIso with EA	EA Isolation with pileup corr.	$< 0.0478 + 0.506/p_T$ $(0.0658 + 0.963/p_T)$
abs(1/E - 1/p)	Energy and momentum matching	< 0.184
0.0721)		
expected missing inner hits	Missing hit in inner tracker	≤ 1 (≤ 1)
pass conversion veto	Photon conversion	yes (yes)

and we veto the presence of an extra loose lepton in the event. Data and MC efficiency measurements of the custom SIP3D, lepton DEEPJET score, and promptness criteria (d_{xy}, d_z) on the light leptons are described in Appendix 11.1.2.

Jets used in this analysis are AK4 PF CHS jets as mentioned in Section 4.4.4, and must have $p_{\mathrm{T}} > 30$ GeV and be within the tracker acceptance ($|\eta| < 2.4$). Jets that overlap ($\Delta\mathrm{R} < 0.4$) with a final state selected lepton are removed from the event. Additionally, jets need to satisfy the tight working point of the pile-up identification criteria. Jets selected by this set of criteria are considered further for b-tagging purposes.

B-tagging is used to select jets likelier to originate from a b quark. The medium working point of the DEEPJET [73, 74, 75] algorithm is used to reject events with b jets efficiently. DEEPJET is a deep neural network based algorithm using convolutional neural network, recurrent neural network and feed forward layers exploiting 16 (6) properties of up to 25 charged (neutral) particle-flow jet constituents, as well as 12 properties of up to 4 secondary vertices associated with the jet to identify a b-quark initiated jet from light jets (u, d, s or g) or charm (c) jets. This working point chosen for this analysis corresponds to an efficiency of 73.3, 71.4, 79.1, and 80.7% for 2016preVFP, 2016postVFP, 2017, and 2018, respectively, each for a mistagging efficiency in light-flavor jets of order 1%. Vetoing events with a b-tagged jet drastically suppresses the top background. Additional details on the b-jet identification and mis-tagging efficiency maps of the background and signal processes are provided in Appendix 11.1.3.

We applied the dedicated $p_{\mathrm{T}}^{\mathrm{miss}}$ filter to mitigate the effect of possible bad $p_{\mathrm{T}}^{\mathrm{miss}}$ calculation, such as interactions of the beam halo (particle created by the interaction of the beam with non-beam objects), which can lead to $p_{\mathrm{T}}^{\mathrm{miss}}$ signature in the detector. In addition, detector noise in the calorimeter can also be reconstructed to give a fake $p_{\mathrm{T}}^{\mathrm{miss}}$ signature in the

detector, along with a fake $p_{\mathrm{T}}^{\mathrm{miss}}$ contribution from poorly reconstructed PF muons. Events are required to pass the following $p_{\mathrm{T}}^{\mathrm{miss}}$ filters:

- Primary vertex filter: Ensures the collision originated from the intended beam interaction point.
- Beam halo filter: Removes events caused by particles traveling along the beam pipe, not from the collision.
- HBHE noise filter: Discards events with spurious signals from the Hadron Barrel and Endcap calorimeters.
- HBHEiso noise filter: Identifies and removes isolated noise spikes in the Hadron Barrel and Endcap calorimeters.
- EEbadSC noise filter (only applied on Data): Excludes events affected by problematic superclusters in the Electromagnetic Endcap calorimeter.
- ECAL TP filter: Filters out events with unphysical energy deposits in the Electromagnetic Calorimeter Trigger Primitives.
- BadMuon filter: Removes events where reconstructed muons are likely just detector noise.
- ECAL bad calibration filter (2017 and 2018): Excludes data from periods with known calibration issues in the Electromagnetic Calorimeter.

4.5.1 Corrections and scale factors

The $p_{\mathrm{T}}^{\mathrm{miss}}$ ϕ distribution has roughly a sinusoidal curve with the period of 2π. The possible causes of the modulation include anisotropic detector responses, inactive calorimeter cells or tracking regions, the detector misalignment, and the displacement of the beam spot. The distribution of true $p_{\mathrm{T}}^{\mathrm{miss}}$ is independent of phi because of the rotational symmetry of the collisions around the beam axis. The xy-shift correction reduces the $p_{\mathrm{T}}^{\mathrm{miss}}$ phi modulation. It is applied to $p_{\mathrm{T}}^{\mathrm{miss}}$ and propagated in the analysis.

In 2018 (from run 319077), two HCAL endcap (HE) sectors, HEM15 and HEM16, became unresponsive and could not be operated for the remainder of the 2018 run. The endcap calorimeter (HEM) lost power during the data-taking period. These modules correspond to the region of $-3.0 < \eta < -1.3$, $-1.57 < \phi < -0.87$, and the loss of HCAL information from this sector affects lepton, photon, and jet reconstruction in that region, as well as $p_{\mathrm{T}}^{\mathrm{miss}}$. We veto events from data and mc, if any jets passing the selection criteria mentioned above fall in this affected region.

The efficiency of different object identification and tagging algorithms may not be the same in data and simulation. These residual differences are often parameterized as kinematic properties (p_{T} or η) of the object. They are called scale factors. These scale factors are then applied to the simulation to correct the estimated yield. Scale factors related to the muon reconstruction, identification algorithm, b-tagging algorithm, pile-up, L1-prefiring, and trigger were taken into account in the analysis.

4.6 Refining lepton identification using Machine Learning

Identifying prompt and isolated leptons is crucial in many prompt BSM physics searches (such as vector-like leptons considered in this thesis) that use leptons ($e, \mu,$ or τ_h) as final state objects. It becomes particularly important for searches using multiple leptons where any of the leptons can come from the unwanted source, such as, heavy flavor decays (b or c), light quark or gluons (u, d, s or g), or the jets faking as a lepton (e.g., wrongly assigned track with ECAL cluster with minimal HCAL deposits misidentified as an electron). The leptons that pass our predefined identification quality criteria but come from unwanted sources are classified as misidentified, non-prompt, or fake. This section focuses only on the light leptons (electrons or muons) that may originate from unwanted sources. The light leptons that come from W, Z, Higgs, or leptonic decay of tau leptons are labeled as prompt leptons. Many properties of such non-prompt and prompt leptons can be exploited to distinguish between them. For example, isolation requirements can suppress the non-prompt leptons generally produced inside a jet (from the semi-leptonic decay of hadrons) as they are poorly isolated. Another useful property is the impact parameter, which is usually small for prompt leptons as the source particles are very short-lived, but can be large for b-hadron or charm hadron decayed leptons. The usual identification criteria may not be efficient or not designed (finely tuned) to reduce these unwanted leptons maximally. As seen in the muon and electron identification criteria described earlier, additional d_{xy}, d_{z}, or SIP3D parameters are useful to suppress such misidentified sources of leptons.

In this section, a deep neural network (DNN) based identification criterion is described to distinguish prompt leptons from non-prompt leptons coming from b-decays. We only consider b-decays as a source for the proof of concept and their relative importance for the multilepton final state searches, where the top-antitop process can produce such leptons, poses a significant amount of background in some phase-space of the search. It is important to note that, although for a reasonable identification criteria (IDs), the misidentification rate is low, the dominant production cross-section of top-antitop or even Drell-Yan (DY) processes can be a dominant source of background in the phase-space where we may search for new BSM signal. This study is particularly geared to reduce the amount of such

backgrounds by combining all the features of such light leptons using a binary deep neural network classifier.

Dataset of prompt and non-prompt leptons

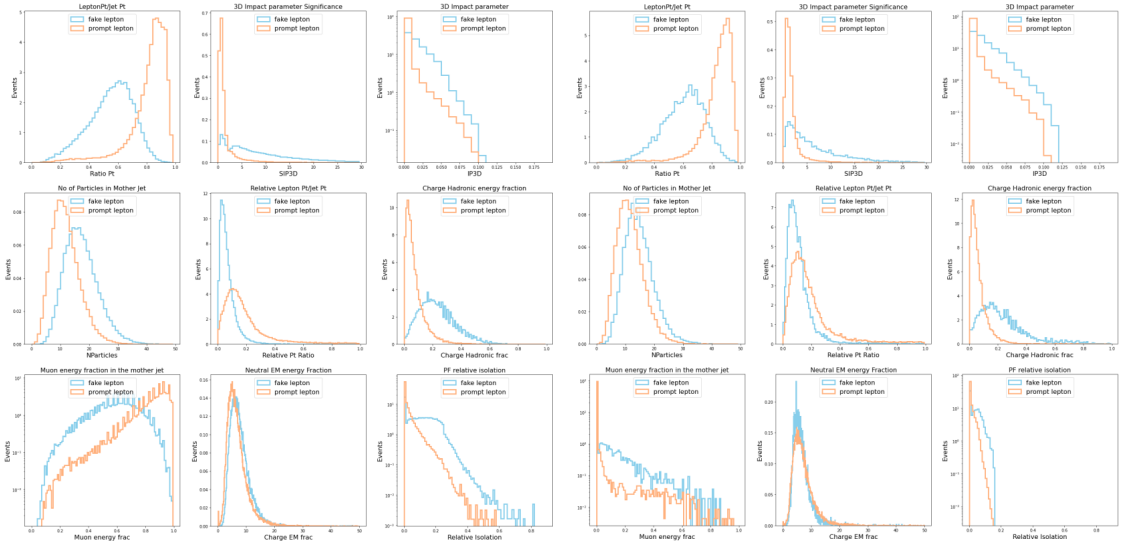
To create a dataset of prompt and fake leptons, $\mathrm{t\bar{t}}$ dileptonic process is exploited where W bosons from the top (or antitop) decays leptonically to produce leptons (in HEP experimental language, electrons or muons are treated as leptons, and hadronic taus are specified explicitly). The following selections are made,

- Three leptons satisfying the analysis level identification criteria described in section 4.5 are selected.
- Any two of the three leptons are matched to a generator-level leptons that come from a W boson using $\Delta\mathrm{R} < 0.2$ matching criteria. These are prompt leptons.
- The lepton that is left is required to be unmatched with any lepton from the W boson ($\Delta\mathrm{R} > 0.2$). This constitutes the non-prompt (or fake) lepton.
- The Closest jet of the unmatched lepton is found by requiring $\Delta\mathrm{R}(jet, lepton) < 0.4$. We call this jet a mother jet of the lepton.
- To ensure the source of the non-prompt lepton, the Jet Hadron Flavor of the mother jet must be a truth-level b-quark.
- Once the procedure is followed, a dataset of prompt and non-prompt leptons is created with multiple properties.

The following properties are kept for each lepton in the dataset,

- *Lepton related*: Lepton SIP3D (significance of 3D impact parameter), lepton IP3D (3D impact parameter), relative isolation of the lepton
- *Associated mother jet related*: Number of particles in the mother jet, charge and neutral hadronic energy fraction of the mother jet, charge and neutral electromagnetic energy fraction of the mother jet, jet muon energy fraction of the mother jet, mother jet mass.
- *Combined*: Ratio of Lepton p_{T} and mother jet p_{T}, Lepton p_{T} relative to jet axis/mother jet p_{T}.

Figure 4.10 demonstrates the listed properties for prompt and non-prompt leptons in the dataset.



(A) Muons (B) Electrons

FIGURE 4.10: Lepton-related features, associated jet features, and features combining the lepton and associated jets for prompt (orange) and non-prompt (sky blue) leptons are demonstrated for muons (left) and electrons (right).

4.6.1 DNN classifier

Network architecture

A binary Deep Neural Network with an input layer of 10 variables, 3 dense layers of 128, 64, 16 neurons, and one output layer is used. Rectified linear unit (Relu) is used as an activation function in hidden layers, sigmoid in the output layer, and the network is initialized by *he_normal* weight initialization technique. Loss is calculated using binary cross-entropy, and the network uses the Adam optimizer. The accuracy is used as a metric during training to save the best trainable model. Two classifier models with similar architecture are implemented using tensorflow-keras library [76] and trained separately for the muon and electron cases. Hyperparameters are optimized using grid and randomized search (RandomizedSearchCV algorithm in scikit-learn [77]). $32k(12k)$ prompt muons(electrons) and $30k(2.6k)$ non-prompt muons(electrons) are used to train the muon and electron classifiers, respectively. Training and testing datasets are independent, and no re-weighting factor is applied during training, as the training procedure was found stable. The network is trained in 32 epochs with a batch size of 256. The training dataset is divided into two parts, with 80% of the data used for training, and 20% of the data used for validation purposes during training to assess the overfitting and monitor the training performance.

Training performance

The metric of the performance is the ROC curve (AUC value or area under the curve). Figure 4.11 shows the classifier output for prompt and non-promt leptons for the train and validation (sometimes called test loosely during training) dataset. Muons and electrons classifiers are shown separately. It also demonstrates the ROC curve for the electron and muon classifiers for the dataset prepared from the $t\bar{t}$ dileptonic sample. It can be seen that overall the muon classifier is performing better than the electron classifier. However, in very high-purity cases (when background efficiency is low), the electron classifier is better than the muon classifier in identifying the prompt sources (signal efficiency).

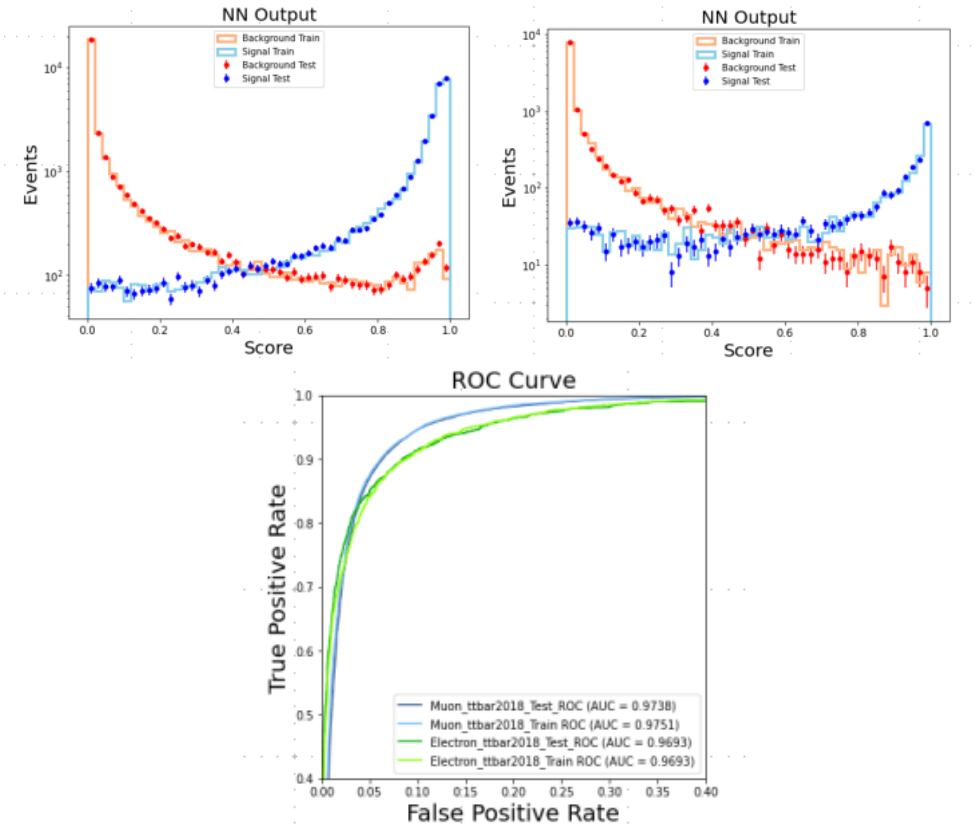


FIGURE 4.11: Classifier output score for training and validation (test) dataset during training demonstrated for muons (upper left) and electrons (upper right) case. The lower plot shows the ROC curve and AUC values for the muon and electron classifiers.

4.6.2 Performance in data

To evaluate the performance of the trained classifier on the data, events with exactly three leptons (3L) (satisfying the criteria described before) are selected from the data. To select the $t\bar{t}$ dominated events from the available 3L data events, as it gives a good mixture of prompt and fake leptons in data, events with an invariant mass of any two opposite sign

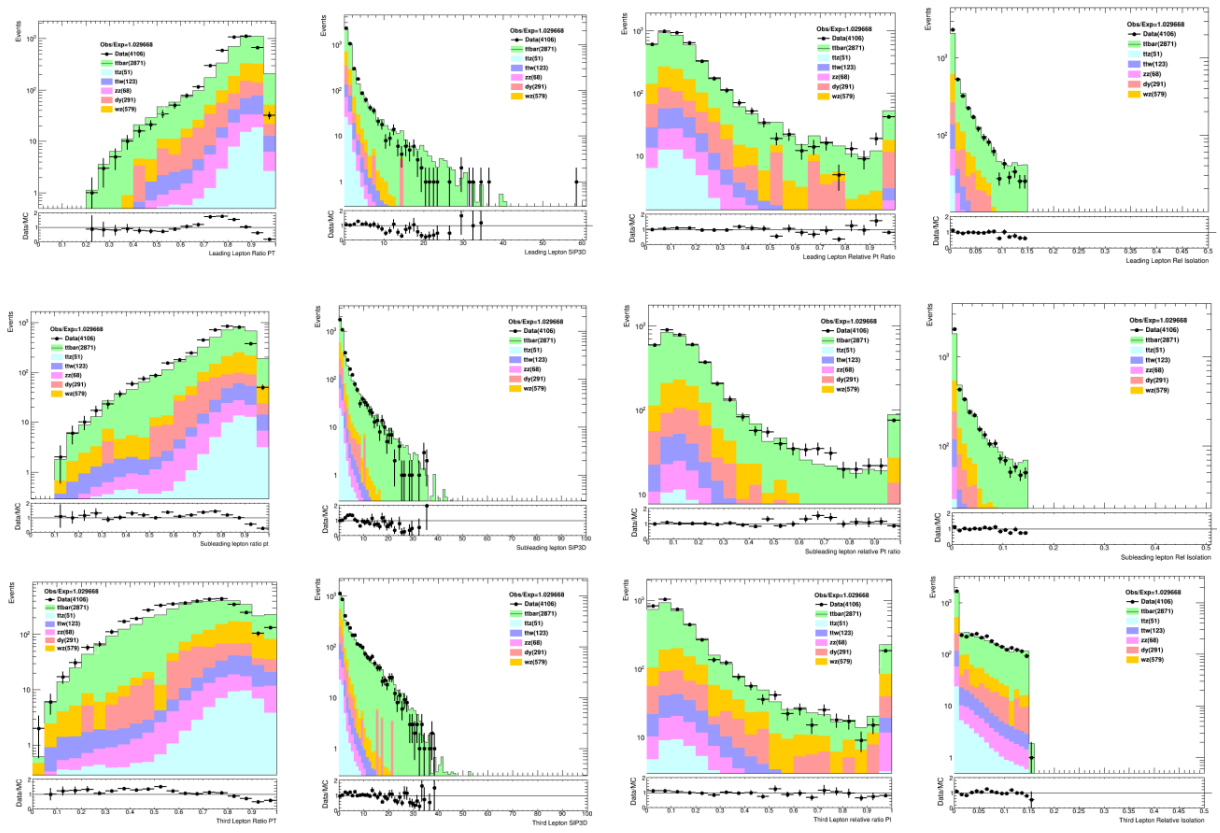


FIGURE 4.12: Illustration of input variables modeling in $t\bar{t}$ 3L CR events. Lepton p_{T}/mother jet p_{T}, SIP3D, lepton p_{T} relative to jet axis/mother jet p_{T}, and PF relative isolation distributions are shown from left to right for leading (top), sub-leading(middle), and third lepton(bottom). The first and last bins include the underflow or overflow.

same flavor leptons being outside the Z mass window(M_{OSSF} <76 GeV or M_{OSSF} >106 GeV) with MET> 60 GeV is required. We call this $t\bar{t}$ events dominated region in data as $t\bar{t}$ 3L control region (CR).

Input variables are well modeled in data, as illustrated for a few variables in the Figure 4.12. Muons and electrons are evaluated using their trained classifiers. Classifier score of all leptons (y-axis is number of leptons) and score of leading, sub-leading, and third lepton (p_{T} ordering) are shown in Figure 4.13. Backgrounds are predicted using dedicated MC samples, and the data are consistent with the predicted backgrounds. Note that non-prompt contributions are also estimated using MC samples, unlike the data-driven approaches usually employed in analysis.

Various event-level variables are studied to compare the performance of the developed lepton classifier against the existing selection criteria based on lepton SIP3D and DeepCSV score of the mother jet of the lepton. One such event variable is LT, the scalar sum of all lepton p_{T} in an event, is shown in Figure 4.14 for $t\bar{t}$ 3L CR for the following three cases,

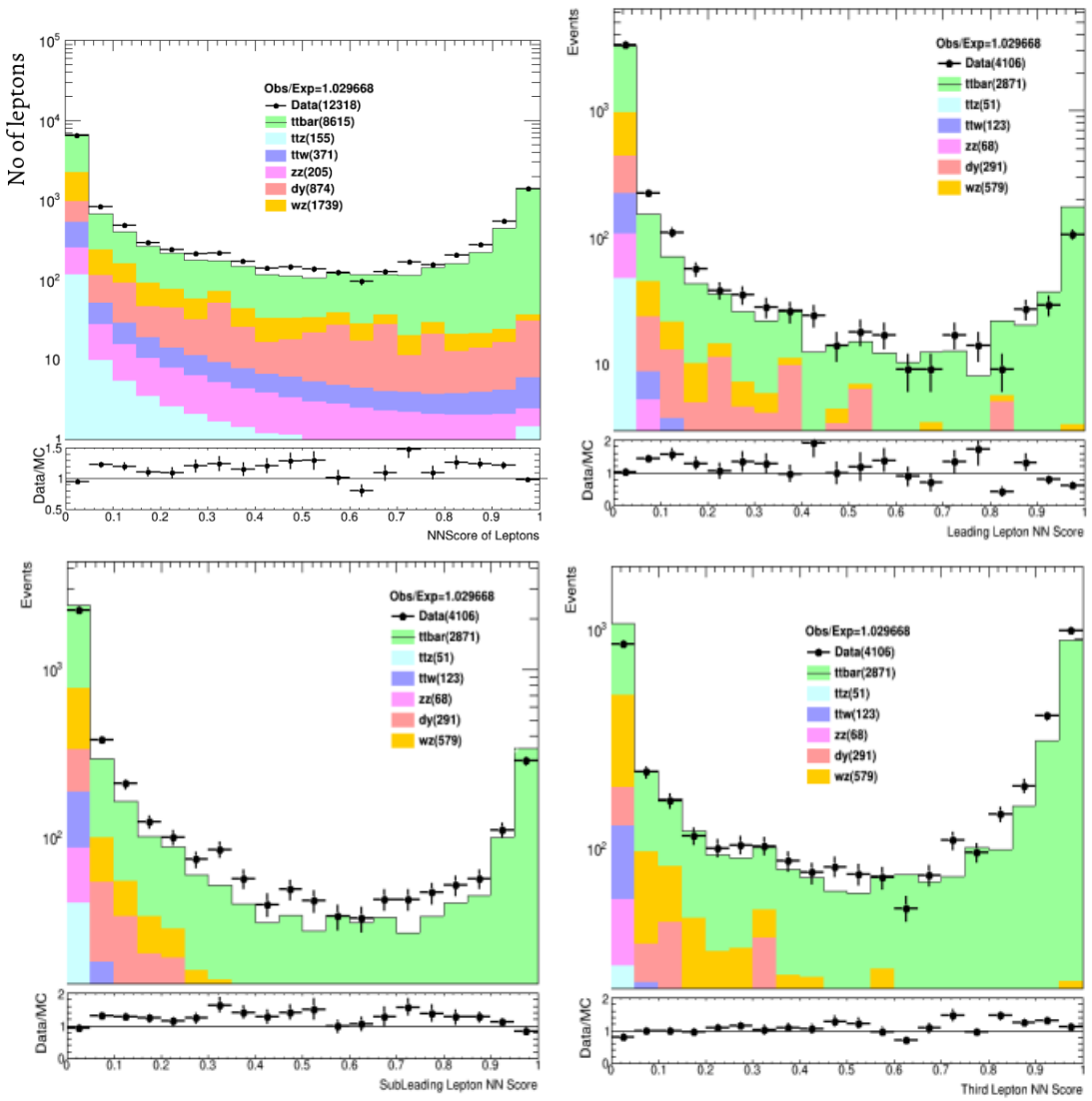
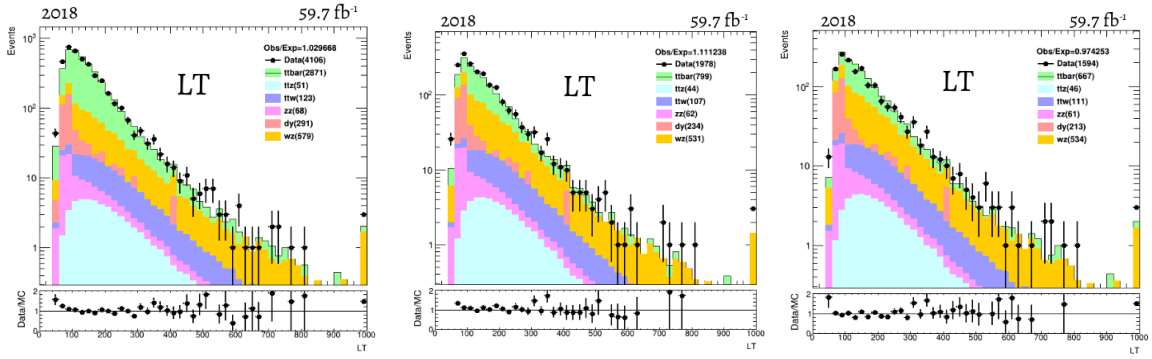


FIGURE 4.13: Evaluated score for all leptons (top left), score of leading (top right), subleading (bottom left), and third lepton score (bottom right) in 3L $t\bar{t}$ CR in data (left to right)



(A) Base selection (B) Box cut (C) Classifier score cut

FIGURE 4.14: LT in $t\bar{t}$ 3L CR for base selection, box cut, and classifier score cut (left to right) on all the leptons.

- **Base lepton selection:** Pass the analysis level criteria without SIP3D and Lepton DeepCSV score.
- **Box cut:** In addition, all the lepton must satisfy SIP3D<9.0 and DeepCSV<0.3 cuts.
- **Classifier score cut :** Base lepton selection + classifier score <0.7 for each lepton

LT spectrum shows good data/mc agreement after applying the classifier score selection criteria on each lepton of the event. Additionally, the fake (non-prompt) contribution is also suppressed compared to the usual box cut (SIP3D and LeptonDeepCSV) even for a conservative cut on the classifier score. Non-prompt backgrounds can be suppressed drastically by applying a hard cut on the soft (third) lepton (such as score <0.1).

Concretely, $t\bar{t}$ and WZ yield (where leptons should come from prompt sources like W or Z boson) are compared in Table 4.4 for the three cases itemized above. Lepton classifier (score<0.7) can remove more $t\bar{t}$ fakes than the usual SIP3D and DeepCSV selection criteria, while the WZ yield remains mostly the same. It indicates that the neural network performs better at reducing fakes without affecting the prompt lepton efficiency. The basic setup is developed and tested on data in the context of discriminating prompt leptons against leptons from b-hadron decays. This can be fine-tuned further depending on the specific application in the future (in Run-3 multilepton analysis that deals with multiple leptons in the final state).

Process	Base selection	Box cut	Classifier score cut
Non-prompt process ($t\bar{t}$)	2871	799	667
Prompt process (**WZ**)	579	531	534

TABLE 4.4: $t\bar{t}$ and WZ yield in $t\bar{t}$ 3L CR for base selection, box cut, and classifier score cut cases.

Chapter 5

Search for VLL in μjj final state: Strategy and Backgrounds

The search was conducted as a straightforward counting experiment. As the low mass VLLs primarily decay to $W\nu_\ell$, the subsequent hadronic decay of W is used to reconstruct the W resonance. However, this property has only been used to construct a signal region, and the presence of the signal is not sought through resonance peaks (excess of events) in data.

The counting experiment is carried out in a blind fashion; all selections and systematics are fixed prior to checking data in the signal region. The method can be summarized briefly as follows:

- Decide the trigger based on the signal parameter space in target. Process the datasets to be used for the analysis. This analysis used a much lighter data tier called NanoAOD format with the latest detector alignment and reconstruction corrections applied to the data.
- Define preliminary event selections and overlap removal cuts among the physics objects.
- This will also decide the significant SM backgrounds in the search. At this point, choose a variable of interest to identify the signal in the cleanest way. Make some selection criteria based on the signal MC samples to determine the most likely data region where the signal could be hidden. Data events satisfying these selection criteria are vetoed and kept for later stages to test the presence of the signal.
- Define a set of control regions targeting the participating backgrounds. Control regions are constructed such that the targeted signal selected in this region is negligible. Control regions are used to estimate corrections and develop methods for the background prediction.
- Define a validation region(s) based on the significant background(s). This validation region is used to evaluate our background estimation methods and add any systematics

based on the modeling discrepancies in our variable of interest. Good modeling of relevant event- and object-level properties in data is also important.

- Further optimize the signal region using machine learning techniques to create a region optimized for the discovery of the signal. This step is not necessary if machine learning based optimization is not used.
- Estimating systematic uncertainties of the backgrounds and signal.
- Examine the data in the optimized signal region. Either claim a discovery or put constraints on the parameter space.

In this chapter, an overview of the dominant SM backgrounds and preliminary event selection is discussed. Background estimation and control regions are discussed next, followed by a machine learning based signal optimization strategy to further optimize the preliminary selected signal region in the next chapter. Finally, the systematics and results will be discussed.

5.1 Standard model backgrounds

The main processes that can mimic the signature of the signal are of three types:

1. Processes which lead to the production of real W bosons constitute the irreducible background for this search. One of the most common such processes is W+ jets, where a W boson is produced in association with hadronic jets through electroweak and QCD interactions. Another major contribution comes from the top quark pair production ($t\bar{t}$+ jets), where each top quark decays into a W boson and a b quark, and the subsequent decays of W boson to leptons and quarks produce same signature as VLL in the detector in case the b-jets are not identified due to algorithmic inefficiency or acceptance. Similarly, single top production, particularly in the t-channel and s-channel processes, also results in real W bosons when the top quark decays. Additionally, diboson processes such as WW and WZ production involve W bosons and contribute to the background when one or both bosons decay leptonically or hadronically. These processes collectively mimic the experimental signatures of the VLL signal events involving leptons, jets, and missing transverse energy.

2. Another type of background may contribute to the search where the source of the muon is not a gauge boson (any prompt sources), but other objects in an event are misidentified as a muon, and pass the analysis selection criteria. A major source of such background is the QCD-multijet production, where one or more jets are misidentified as leptons. Another source is instrumental backgrounds that arise primarily from

detector effects and reconstruction, contributing to the total $p_{\mathrm{T}}^{\mathrm{miss}}$ artificially. These backgrounds are particularly challenging because they do not originate from prompt sources, and are typically estimated using data-driven techniques to capture the complex nature of detector misidentification of such leptons. However, in this search, event selection and ML techniques can reduce the QCD-multijet background significantly. Thus, it is a negligible background and is estimated using dedicated MC samples.

3. Z boson production: The Drell-Yan process with additional jets ($Z/\gamma^{*}+$ jets) where the Z boson decays into a pair of muons can be a background. If one of the muons is not reconstructed or falls outside the detector acceptance, the event may appear as a single muon plus jets, mimicking the signal topology. Similarly, ZZ production, where one Z decays into muons and the other into neutrinos or jets, can yield final states with a muon and additional jets in the event, along with real missing transverse energy from neutrinos. These processes can therefore resemble signal events, and should be estimated.

5.1.1 W + Jets MC samples

$W+$ jets events are the most dominant backgrounds in this search, irrespective of discriminative selection criteria to reduce such events, mainly due to their huge cross-section. Three types of MC samples are stitched together to estimate the $W+$jets background to cover the full phase space probed in this analysis. They are the following,

- **WJets bulk**: WJetsToLNu_Inclusive samples (MADGRAPH5_AMC@NLO at LO). These samples are inclusive and the W boson is mostly produced on-shell, but have low statistical precision at high HT, muon M_T tails, or jet multiplicity in events.

- **WJets HT-binned**: WJetsToLNu samples (MADGRAPH5_AMC@NLO at LO with parton level HT thresholds 70-100, 100-200, 200-400, 400-600, 600-800, 800-1200,1200-2500, and 2500-Inf GeV). These samples are produced to improve the statistical precision and cover the full HT phase-space defined by the HT boundaries in the sample. In these samples, the W boson is produced mostly on-shell, and is unable to cover the high muon transverse mass (M_T) phase-space, which is crucial for this type of signal, where the muons may originate from the VLL particle, and have higher M_T.

- **W-mass binned**: WToMuNu PYTHIA samples (with mass value as M100, M200, M500, M1000, M2000, M3000, M4000, M5000, and M6000 GeV) are specific simulation samples with the contribution from off-shell standard model W bosons as denoted by the mass value. These samples are generated to model high transverse mass

tails of W boson production, where the W boson is far off-shell. These samples are not "boosted" in the usual sense where a low-mass particle (like an 80 GeV W boson) has high p_{T}; rather, the W itself is off-shell, resulting in high M_T rather than Lorentz boosts of an on-shell particle.

Figure 5.1 demonstrates the relative contribution of these three kinds of samples in the muon M_T phase-space in 2018 at preselection. The yield in this plot is normalized to 2018 luminosity.

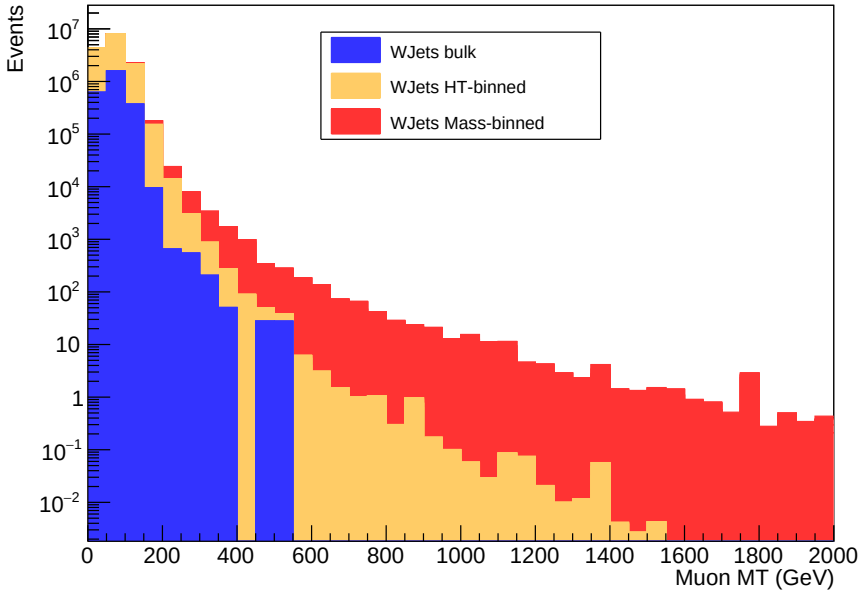


FIGURE 5.1: Muon M_T distribution for data-taking year 2018 at analysis event preselection level is shown. Three kinds of samples described in the text are stacked together to show their relative contribution in the M_T phase-space. W mass-binned samples are important to model the high muon M_T tails, where the high M_T is due to the large off-shell production of the W boson. This plot is made after overlapping the events from these three samples.

Overlap events removal

The three types of W +jets MC samples are made independently for different purposes, and are not necessarily mutually exclusive. These samples partially cover the same phase space, i.e., some simulated events might appear in both bulk and HT-binned samples, or both HT-binned and mass-binned samples. If such overlaps are not properly removed, they lead to double-counting, resulting in incorrect background normalizations and distorted kinematic distributions, which can bias physics analyses and incorrect estimated yields in sensitive signal regions. To remove overlap, LHE level variables such as LHEHTincoming (scalar

sum of final state parton transverse momentum) and mass of the $W(M_W)$ at generator level are used.

First, events with HT >70 GeV are taken from H_T-binned samples, and events with HT< 70 GeV are taken from the bulk sample. This creates non-overlapping HT regions between these two samples. Then, events with gen level W boson mass, $M_W < 100$ GeV, are taken from HT-binned and bulk samples. Events with $M_W > 100$ GeV are used from the W mass binned samples. Thus, the overlap in W mass regions is avoided in these three samples.

Figure 5.2 schematically shows the stitching strategy of three different samples to remove overlap events using generator-level quantities.

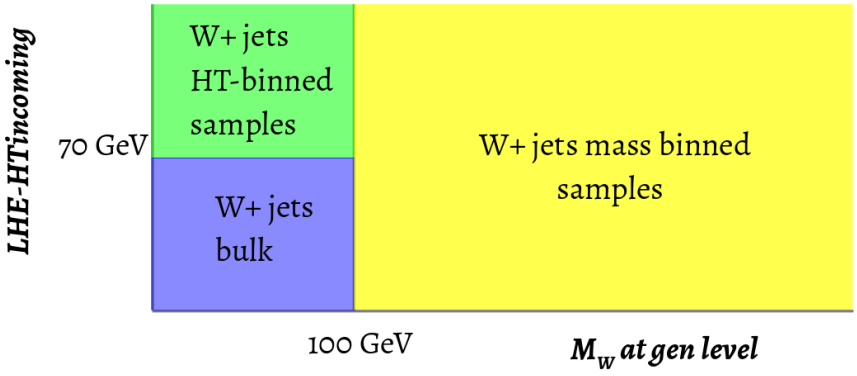


FIGURE 5.2: Stitching strategy for the three type of W+jets samples to remove overlap events. Overlap is removed using generator-level quantities such as mass of the W boson at gen level, and LHE (parton) level quantity, scalar sum of final state parton p_{T}, LHEHTincoming.

It might happen during stitching that the boundaries between the samples of a particular category are not smooth, resulting in an artificial "kink" in the reconstructed and gen-level distribution. Figure 5.1 illustrates that there is no such kink in the reconstructed level M_T distribution, and Figure 5.3 demonstrates a smooth transition of different samples in some specific category or across categories in a few key gen level quantities.

Before going into the event selections, let's look at the definition of some variables used in this analysis of one muon and at least two jets.

5.2 Variable definitions

The following variables are defined to facilitate the event selection in this analysis, and a subset of them is used for the neural network training discussed later.

- M_W: The SM W boson mass, 80.3 GeV.
- $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$: Particle flow-based missing transverse momentum vector in an event.

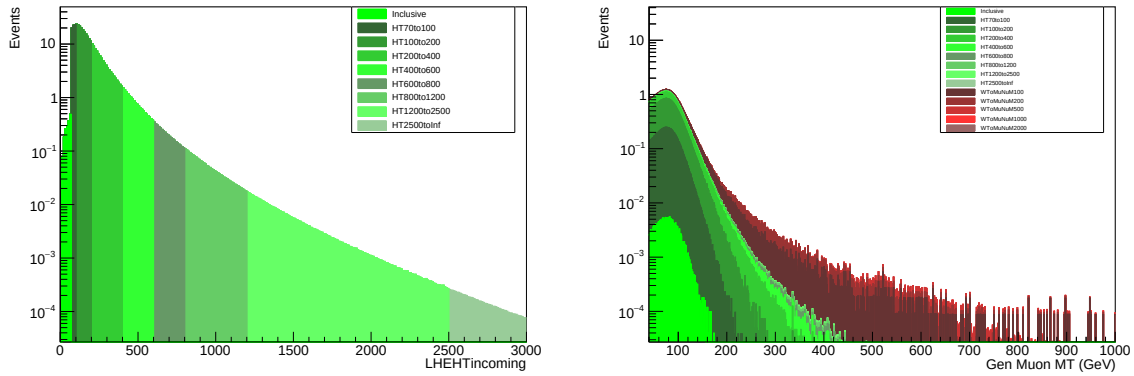


FIGURE 5.3: Smooth transition of different samples in some specific category or across categories in a few key gen level quantities to remove pathological cases of artificial kinks in combined samples.

- $p_{\mathrm{T}}^{\mathrm{miss}}$: Magnitude of the missing transverse momentum vector in an event.
- N_j: Number of jets in an event satisfying the jet selection criteria.
- N_b: Number of medium b-tagged jets in an event.
- Scalar sums: We define H_T as the scalar p_{T} sum of all jets satisfying the selection requirements. Additionally, the scalar sum of lepton p_{T}, H_T, and $p_{\mathrm{T}}^{\mathrm{miss}}$ is defined as S_T.
- Invariant and transverse masses: We define M_{j0j1} as the invariant mass of the two selected jets ordered in p_{T}. The transverse mass for a single object i is defined as $M_T^i = (2p_{\mathrm{T}}^{\mathrm{miss}} p_{\mathrm{T}}^i [1 - cos(\vec{p_{\mathrm{T}}^{\mathrm{miss}}}, \vec{p_{\mathrm{T}}^i})])^{1/2}$, where p_{T}^i is the p_{T} of the object i. M_T^{μ},M_T^{j0}, and M_T^{j1} define the transverse mass of the muon, leading, and subleading jets, respectively. Similarly, M_T^{ij} is the transverse mass calculated with the $p_{\mathrm{T}}^{\mathrm{miss}}$ and the resultant 4-momentum vector of objects i and j.
- $\Delta R(i,j)$ and $\Delta\phi(i,j)$: We define $\Delta R(i,j)$ and $\Delta\phi(i,j)$ as the distance between object i and j in the $\eta - \phi$ plane and difference in azimuthal angle ϕ of object i and j, where i, j can be the lepton, jets, and $p_{\mathrm{T}}^{\mathrm{miss}}$ in an event.
- DeepJetQG score: Value of the DeepJet Quark Gluon discriminator of each jet.

5.3 Preliminary event selection

Events are selected with exactly one muon and at least two jets. These objects must satisfy the object identification criteria described in Section 4.5. The muon is required to have

$p_{\mathrm{T}} > 26(29)$ for 2016, 2018 (2017) following the trigger criteria considered in this analysis. We veto events if more light lepton (electron or muon) is found in the event that satisfies the loose lepton criteria. Leading and subleading jets are chosen from the jet pairs (ordered in p_{T}) that give the best reconstructed W mass ($M_W = 80.3$ GeV). Events with exactly two jets, leading and subleading jet (in p_{T}), are considered to characterize the event. Events with three or more jets, leading and subleading jets refer to the pair of jets ordered in p_{T}, which gives the best reconstructed W mass. Given the signal characteristics, identifying the two jets coming from W may distinguish the signal over SM backgrounds. In any case, the leading and subleading jets must have $p_{\mathrm{T}} > 30$ GeV and satisfy other criteria described in the jet identification criteria. Selected muon must be separated by $\Delta R = 0.4$ from the leading or subleading jet. Also, we required the leading and subleading jets to be separated by $\Delta R = 0.4$. The invariant mass of the two jets,$M_{j_0 j_1}$, is required to be larger than 40 GeV to suppress multijet background arising from the soft QCD interactions or instrumental fake backgrounds. The signal mass range that the analysis probed is primarily dominated by the $W\nu_\ell$ decay mode; it is less likely to get a jet originating from a b-quark in the event. To veto events with b-jets, we used the medium working point of the DEEPJET tagger to identify b-jets and required $N_b = 0$, where N_b denotes the number of b-tagged jets in the event. This criterion removes the top background drastically. These selections define our analysis preselection criteria described in Table 5.1.

Criteria	Cuts
No of muons	N_μ = 1
No of jets	$N_{jets} \geq 2$
Dijet Invariant mass	$M_{j_0 j_1} > 40$ GeV
No of b-tagged jets	$N_b = 0$
Veto extra loose lepton	Yes
Overlap removal	$\Delta R(j_0, \mu) > 0.4$,$\Delta R(j_1, \mu) > 0.4$,.$\Delta R(j_0, j_1) > 0.4$

TABLE 5.1: Preselection criteria for this analysis

After preselection, the following variables are optimized to construct control and a preliminary signal region for this search.

- $\Delta R(j_0, j_1)$: separation between leading and subleading jet in $\eta - \phi$ plane and,
- $M_{j_0 j_1}$: invariant mass of the two selected jets.

- S_T: Scalar sum of the Muon p_{T}, $p_{\mathrm{T}}^{\mathrm{miss}}$ and HT.

The primary signal region (SR) is identified by exploiting the signal characteristics. Hadronic decay of W produces two energetic jets (resolved). Invariant mass of this dijet system ($M_{j_0 j_1}$) is restricted between 50 and 110 GeV to capture the on-shell W resonance. Finally, the $\Delta R(j_0, j_1) < 2.6$ criteria is imposed. These two criteria reduce the dominant W+Jets background, as the jets produced via QCD radiation in association with the W boson have no correlation (wrt to W boson), and are mostly produced back-to-back to balance the event in the transverse plane. $S_T > 250$ GeV is imposed on the event as it improves the signal significance. VLL Mass 150 GeV was chosen as a benchmark point to optimize the SR. The SR preselection criteria are summarized below:

- $50 < M_{j_0 j_1} < 110$ GeV
- $\Delta R(j_0, j_1) < 2.6$
- $S_T > 250$ GeV

A dedicated ML strategy is used downstream to separate signal events from the background events. To keep enough events for the training of the networks is deemed necessary. The above-mentioned signal region preselection criteria are chosen to be looser to moderately reject participating backgrounds. The ML algorithms may be trained optimally and achieve higher signal significance $\left(S/\sqrt{B}\right)$. This ML-based region is further divided into validation and final signal regions based on the signal sensitivity.

On the other hand, dedicated control regions (CR) are constructed to estimate W+jets and QCD multijet background, the two most dominant background contributions of this analysis. We inverted the muon isolation criteria for the QCD multijet CR, keeping the other preselection cuts unchanged. To estimate W+jets background, events are selected with $50\ \mathrm{GeV} < M_T^{\mu} < 130\ \mathrm{GeV}$ and $\Delta R(j_0, j_1) > 2.6$ to construct a region with high purity of W+jets events. This region is named WJets CR. Events are selected with $\Delta R(j_0, j_1) < 2.6$ (vetoing the preliminary SR events) to construct the WJets Validation Region (WJets VR) to validate the W+jets estimation strategy developed in WJets CR. Figs. 5.4 shows schematically the SR preselection, Wjets CR, and Wjets VR on the M_{j0j1}-$\Delta R(j_0, j_1)$ plane where other dimensions such as M_T^{μ} and S_T are not shown. Table 5.2 consolidates the SR preselection and various CRs used in this analysis. Additional details are discussed in the background estimation in Chapter 6. Different kinematics and event-level distributions are demonstrated in the preselected SR in Figure 5.5 for the full Run-2 dataset.

Fig 5.6 demonstrates the relative background composition in W+jets CR, VR, and SR preselection regions in the full Run-2 dataset.

Region	baseline	$\Delta R(j_0, j_1)$	$M_{j_0 j_1}$(GeV)	M_T^{μ}(GeV)	S_T (GeV)	μ_{iso}
SR presel.	Yes	< 2.6	$50 < M_{j_0 j_1} < 110$	–	> 250	< 0.15
QCD CR	Yes	–	–	< 50	–	> 0.15
WJets CR	Yes	> 2.6	–	$50 < M_T^{\mu} < 130$	–	< 0.15
WJets VR	Yes	< 2.6	$110 < M_{j_0 j_1} < 50$	–	–	< 0.15

TABLE 5.2: Selection criteria for the preselected signal region for ML training, control regions, and validation regions for estimating backgrounds in this analysis. "–" denotes no cut applied on the variable.

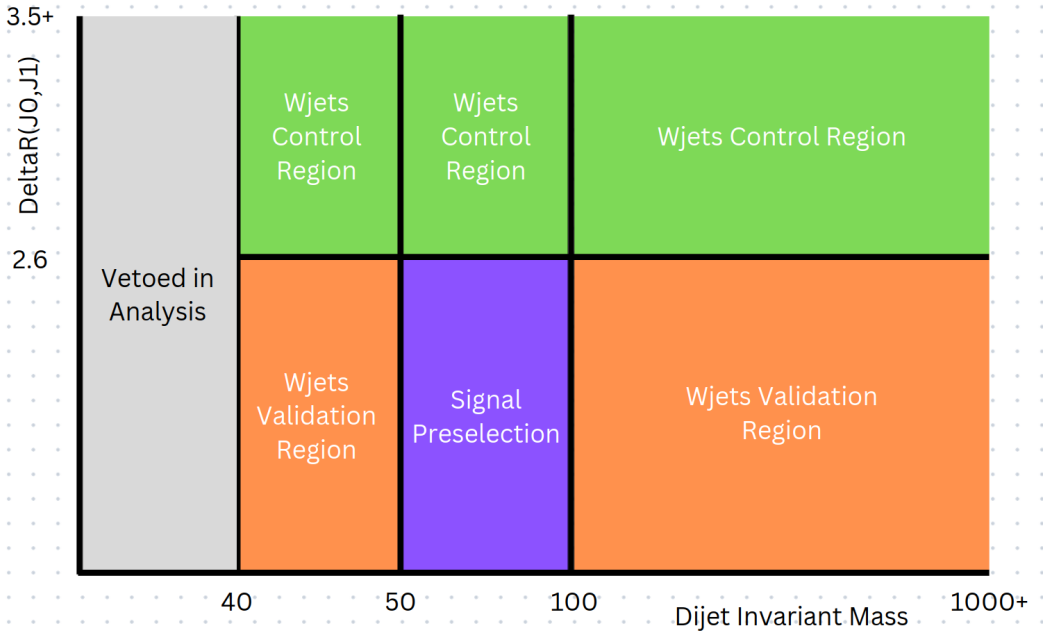


FIGURE 5.4: Schematic diagram of different event selection criteria required in this analysis for SR preselection, WJets control region, and validation region. This diagram is drawn on the 2D plane of M_{j0j1}-$\Delta R(j_0, j_1)$, where other dimensions M_T^{μ} or S_T are not shown for brevity.

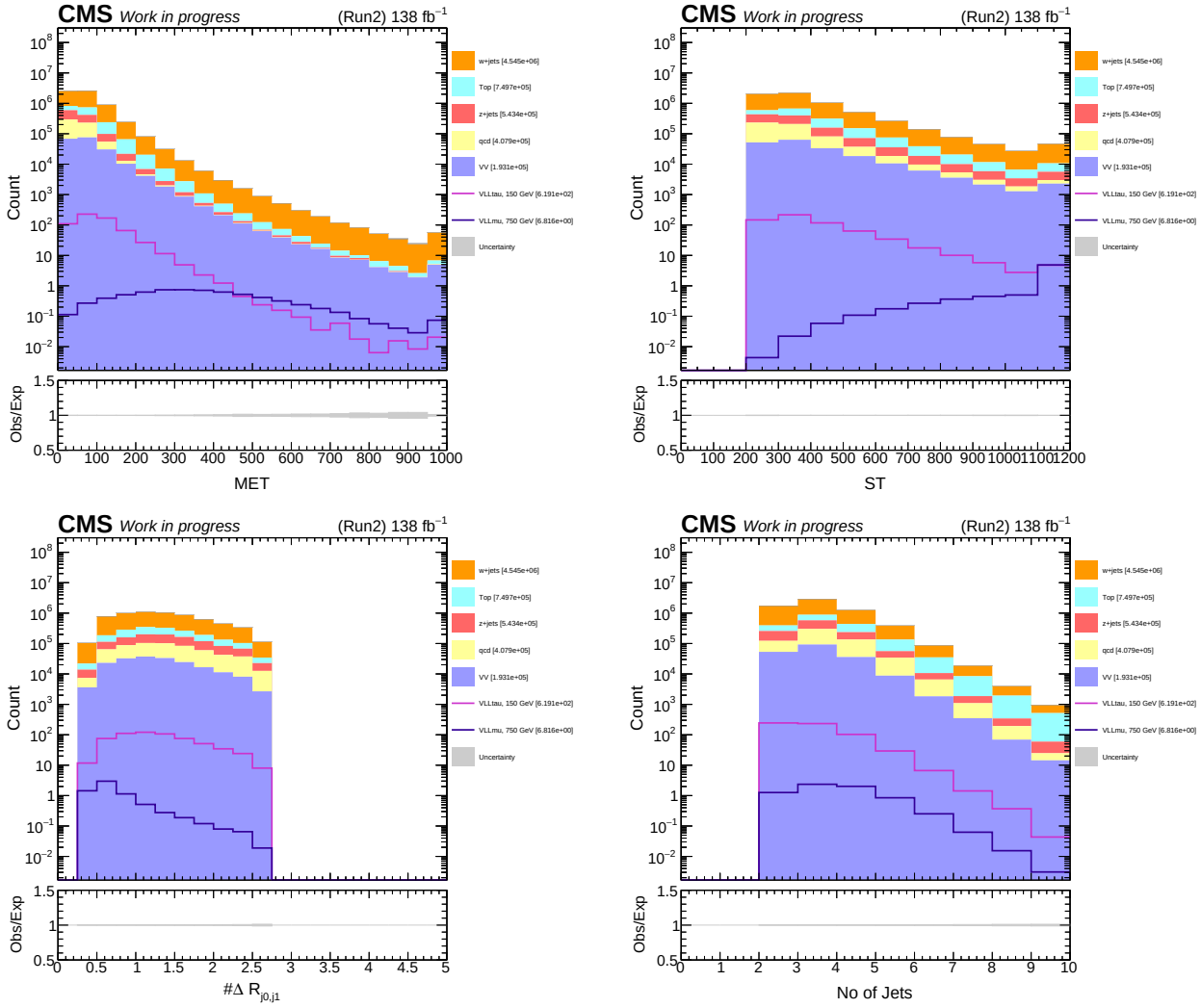


FIGURE 5.5: Distribution of backgrounds and a few representative signal mass hypotheses with events satisfying SR preselection criteria in the full Run-2 dataset. Backgrounds are stacked, and the signal is overlaid. Statistical uncertainties only. The first and last bins include underflow and overflow.

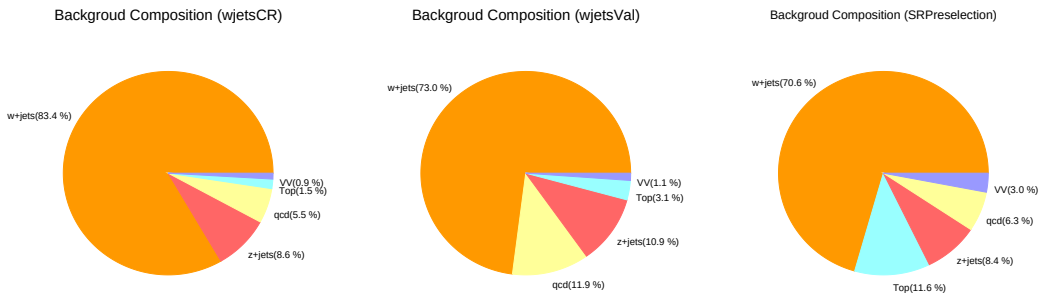


FIGURE 5.6: Background composition in W+jets CR, VR, and SR preselection region in full Run-2 dataset.

5.4 Background estimation

Devising and validating background estimation techniques are crucial for any experimental search. This is usually carried out in some data regions with minimal or negligible contamination of the targeted signal. In other words, in a data region dominated by the background. These regions are naturally called control regions. Multiple control regions can be devised to test different background contributions or specific background estimation methods.

In this section, the background estimation techniques are described. For the dominant background W+jets process, the background estimation methods are also validated in an orthogonal set of events to the W+jets control region.

5.4.1 QCD multijet background

Multijet background due to QCD interactions is a reducible background in the μjj final state, where a jet can fake a muon passing the analysis muon selection criteria. However, the QCD-multijet background can be reduced significantly by event selection and ML techniques. Thus, it turned out to be a negligible background.

The rest of the QCD contribution that satisfies the event selection criteria is estimated using MC samples. Dedicated Pt-binned QCD Mu-enriched samples are used. These samples are used to study the QCD multijet processes that contain muons, typically arising from the semileptonic decays of heavy-flavor hadrons like b and c quarks. These samples are binned in the transverse momentum of the hat variable ($\hat{p_{\mathrm{T}}}$), which refers to the scale of the hard scattering process. It is defined as the scalar sum of the transverse momenta of all partons (quarks and gluons) in the hard scattering process at the generator level. The binned samples provide adequate statistics in different energy ranges. These samples are usually generated using PYTHIA 8, which simulates the matrix element, the parton showering, and hadronization. Since PYTHIA is a leading-order (LO) generator, its cross-section must be corrected to consider the higher-order corrections. QCD contribution is normalized to the data to correct its cross-section.

Usually, muons that arise from jet fragmentation in a QCD process (from light quark or gluon jets or b-jets) are not isolated, i.e., the amount of energy around the lepton candidate is significantly different than the muons decayed from prompt sources (such as W, Z, H, or τ and VLL). This property can be exploited to create a data sample dominated by the QCD multijet events. Events that satisfy the preliminary event selection criteria are considered, except that the muon isolation criterion is inverted. Events where muons satisfy $\mu_{iso} > 0.15$ ($\& < 1.0$) are used to get a QCD control region. Signal contamination is negligible as the muons from the signal either come from gauge boson decays or VLL. This cut also made this region orthogonal to the preselected signal region events. An additional cut on the

transverse mass of the muon (M_T^{μ}) is imposed to get a high purity of QCD multijet events. The QCD cross-section is corrected to data in this QCD control region, where other SM contributions are negligible. A normalization factor is derived for each data-taking era as shown in Table 5.3. In this CR, we observed good data/mc shape agreement as shown in Fig. 5.7 for 2016preVFP, 2016postVFP, 2017, and 2018, respectively. The stability of this normalization factor is established by varying the M_T boundaries, using the $p_{\mathrm{T}}^{\mathrm{miss}}$ variable, and isolation boundaries. These events are triggered using HLT level isolated muons, where a looser isolation variable (than the defined analysis isolation) is typically calculated. A different non-isolated trigger, with varying p_{T} threshold, is also used to check the stability, and in all cases, $< 1 - 2\%$ variation is observed. Since the QCD contribution is negligible in the end, these estimates are reasonable. The derived normalization factors are used on the QCD yield from the MC samples in the signal region to estimate the expected QCD contribution. In other words, shapes of different variables are taken from QCD MC samples, but they are normalized to data to take into account the corrected cross-section.

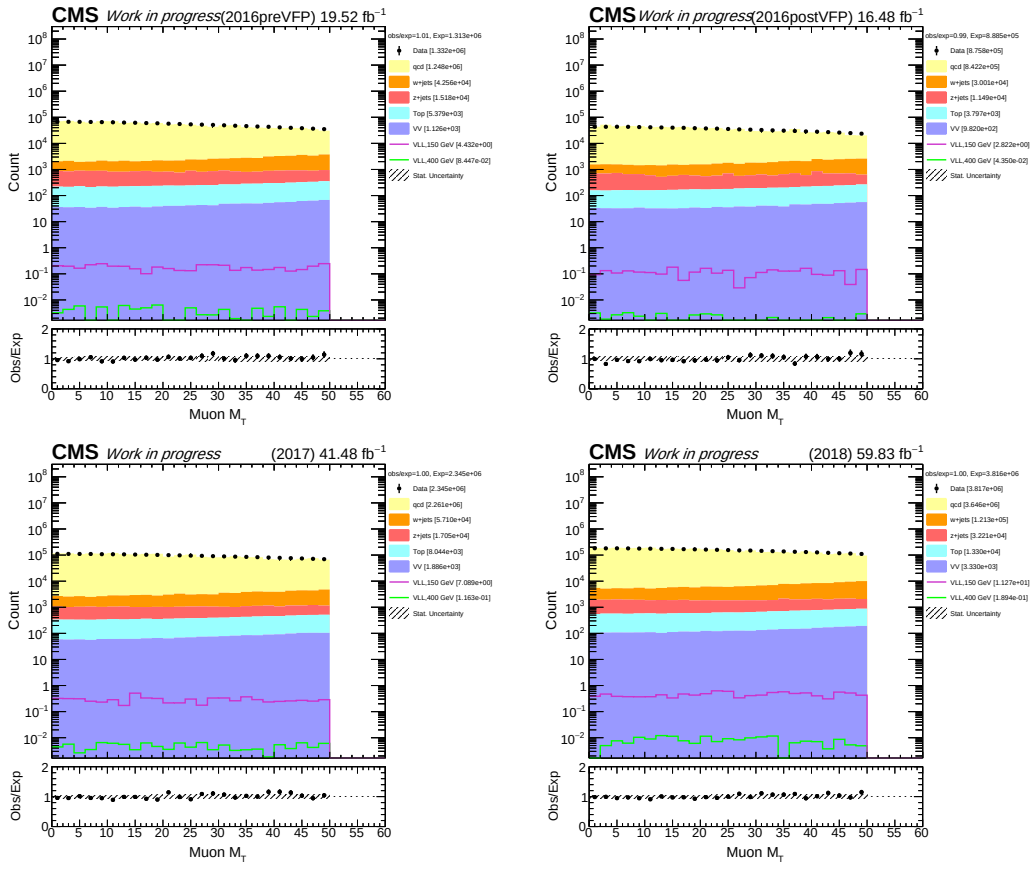


FIGURE 5.7: Muon M_T in QCD control region for 2016preVFP, 2016postVFP, 2017 and 2018 dataset after normalization. The lower panel shows the ratio of observed events to the total expected background prediction. The gray hatched band on the ratio represents the statistical uncertainties in the SM background prediction. Representative signals are also overlaid to show the negligible signal contamination in this control sample.

Era	Normalization Factor
2016preVFP	0.887 ± 0.016
2016postVFP	0.889 ± 0.019
2017	1.162 ± 0.018
2018	0.996 ± 0.015

TABLE 5.3: QCD multijet normalization factor for 2016(pre and post), 2017, and 2018. Pt-binned QCD mu-enriched simulation samples are used.

5.4.2 W + jets background

As discussed, the production of W+jets, where W decays leptonically to a muon along with jets from initial state radiation (ISR) or final state radiation (FSR), is the most dominant background for this search.

Events with large $\Delta\mathrm{R}$ separation between leading and subleading jets satisfying analysis preselection criteria are used to create the W+jets control region. An additional requirement of muon M_T^{μ} between 50 GeV and 130 GeV is imposed to increase the purity of W+jets events as described in Table 5.2. This additional requirement reduces the QCD background drastically as these events are expected to populate the low M_T^{μ} region. Contributions from other prompt processes (top, diboson, or Z+jets processes) are taken from their respective MC samples.

The combined MC sample (LO) of three types, W+jets inclusive, HT-binned, and mass-binned, is used to estimate W+jets background as described in Section 5.1.1. The combined sample will be called the W+jets background in text for brevity. First, the W+jets background is normalized to data in the control region as described earlier. Table 5.4 shows each data-taking era's normalization factor derived in the W+jets CR. The large normalization factor is attributed to the well-known NLO/LO (k-factor) discrepancy, as higher-order corrections are not captured by LO samples that are used to predict the background. It is worth remembering that data (or nature) has all the highest possible corrections for these processes. NNLO or NLO are our attempts to model those corrections in a mathematical framework and simulation. The shape of kinematic and event-level variables that will be used for ML training is investigated. While the LO sample can model many variables in data, i.e., the data/mc agreement is good, mismodeling is observed in a few key variables such as HT, muon p_{T}, or W p_{T}.

LO samples are known to underestimate hard emissions as they only include the lowest-order Feynman diagrams without loop corrections or additional real emissions, thereby restricting the number of partons in the final state. Radiation is added through parton showering, which doesn't always fully capture the phase space. As a result, important effects like initial and final state radiation (ISR/FSR), which can significantly boost the W boson or generate high-HT events, are inadequately modeled. This also impacts the jet multiplicity

Era	Normalization Factor
2016preVFP	1.331 ± 0.005
2016postVFP	1.423 ± 0.006
2017	1.467 ± 0.005
2018	1.393 ± 0.005

TABLE 5.4: W+jets normalization factor for 2016preVFP, 2016postVFP, 2017, and 2018. HT binned LO W+jets samples are used to normalize WJets CR to data.

of the event. Several differential cross-section measurements of the W+jets process shown in Figure 5.8 indicate that NLO may model the low jet multiplicity region well. Still, LO samples can better describe the higher jet multiplicity region.

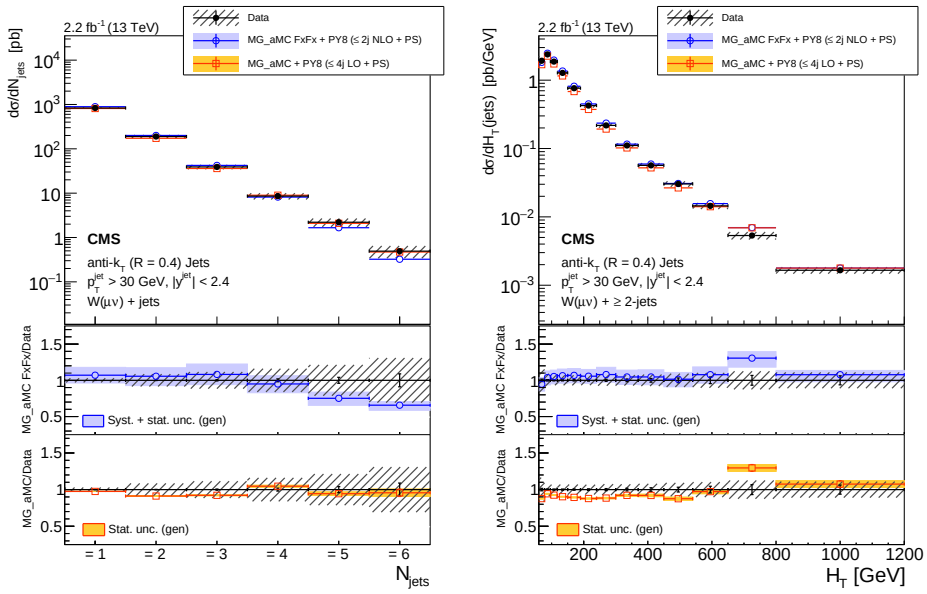


FIGURE 5.8: Differential cross-section measurement for the inclusive jet multiplicity (left), for the jets HT, shown for at least two jets (right), compared to the predictions of MG_aMC FxFx and MG_aMC. The black circular markers with the gray hatched band represent the unfolded data measurement and the total experimental uncertainty. The MG_aMC prediction is given only with its statistical uncertainty. The band around the MG_aMC FxFx prediction represents its theoretical uncertainty, including statistical and systematic components. The lower panel shows the ratios of the prediction to the unfolded data. Figure is taken from [78]

Figure 5.9 (left) demonstrates the limitation of NLO samples at low dijet invariant mass phase-space, where LO generated samples are modeling the data well [79]. The right panel shows the double differential cross-section measurement from [80] where the ratio of MADGRAPH at LO and data is plotted as a function of Z boson p_{T} in different rapidity regions. The plot demonstrates the issue in vector boson p_{T} (here Z boson p_{T}) modeling using LO generators. At LO, processes like $q\bar{q} \to W$ or $q\bar{q} \to Z$ generate the boson without any recoil, i.e., the boson has zero transverse momentum in the parton-parton center-of-mass

frame. The boson may get transverse momentum from initial state radiation (ISR) modeled by the parton shower, but it can simulate soft and collinear gluon emissions reasonably well; however, it doesn't accurately model hard, wide-angle QCD radiation. So, the choice of NLO and LO to describe data can be driven depending on the variable of interest and the analysis phase space. NLO samples based study is described in Appendix 11.2, where we found mismodeling of angular separation of leading and subleading jets, and jet multiplicity using NLO samples. Moreover, the p_{T}-binned NLO samples (where W bosons are mostly produced on-shell) could not cover the high muon-$\mathrm{M_T}$ phase space which was crucial for this analysis.

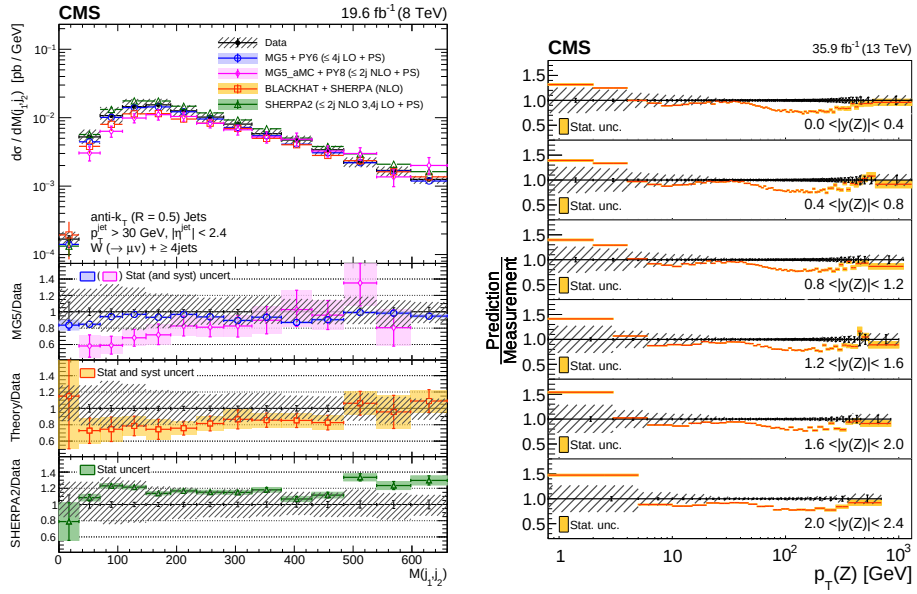


FIGURE 5.9: Differential cross-section measurement in dijet invariant mass (calculated from the two leading jets) for inclusive jet multiplicities 4 for a set of generators (left). The lower panel shows the ratios of the prediction to the unfolded data. The right plot shows the double differential cross sections as a function of Z boson p_{T} and rapidity for events with at least one jet, compared to the predictions of LO mg5_amc. Left figure is taken from [79] and right figure is taken from [80].

To model the NLO/LO effect in the W+jets events (predicted by LO samples), we applied a set of corrections in the control region. We derived corrections as a function of HT, muon p_{T}, and muon M_T in the CR. Finally, we can apply these correction factors to events predicted by the combined W+jets MC in SR following Equation 5.1.

$$N^{SR}_{W+jets} = N^{SR}_{W+jets}(MC) \times f_{corr}(HT, p^{\mu}_{\mathrm{T}}, M^{\mu}_{T})|_{CR} \tag{5.1}$$

where f_{corr} are the derived correction factors in the different ranges of HT, muon p_{T} or M_T.

The correction factors are derived sequentially as shown in Eqn 5.2. Other backgrounds are subtracted from data after normalization, etc, taken into account.

$$f^{i}_{corr} = \frac{N^{i}_{Data} - N^{i}_{OtherBkg.}}{N^{i}_{W+jets\ MC}};\ i = bin \tag{5.2}$$

The following points describe the corrections:

- **HT based**: HT-based correction factors are derived in events with 2 or 3 jets and more than 3 jets separately. HT ranges are optimized as 20 GeV binning between 80 to 100, 50 GeV binning between 100–500 GeV, and 100 GeV binning for 500 GeV and above for events with N_j = 2, 3. One bin for < 80 GeV and > 1000 GeV to cover the low and high HT events. For events with $N_j \geq 4$, 100 GeV HT bins are used between 100-1000 GeV (the last bin includes overflow events).

- **Muon p_{T} based**: The W boson p_{T} is corrected using muon p_{T} as a proxy object. In μjj final state, the W decays to the muon and a neutrino. Since the neutrino is undetected, four vectors of W can not be constructed. We make use of the muon as a proxy object to correct the W. However, we checked W p_{T}-based correction by constructing W boson p_{T} using muon and $p_{\mathrm{T}}^{\mathrm{miss}}$, or applying muon p_{T} based correction to check W p_{T}. The W p_{T} is well modeled after applying the muon p_{T} correction, and this strategy was chosen. Muon p_{T} bins of 10 GeV between 20–100 GeV, 25 GeV between 100–200 GeV, and 50 GeV for 200–500 GeV are used. > 500 GeV bin includes all high p_{T} events.

HT-based and muon p_{T}-based corrections, applied on the W+jets MC samples for different data-taking eras, are shown in Fig. 5.10. The corrections are found to be consistent between all eras. The correction factor in the first HT bin for $N_j \geq 4$ events is larger with uncertainty due to the binning, as a few events populate the region.

Muon M_T based correction

After normalizing W+jets to data, HT, and muon p_{T} based correction in WJets CR, the muon M_T criteria are removed to construct an extended WJets CR to carry out the muon M_T based corrections. This rolled out muon M_T distribution can be used to reweight the events in broad bins of M_T. Muon M_T bins of <100 GeV, 100-200 GeV, 200-450 GeV, 450-750 GeV, 750-1000 GeV, and >1000 GeV are used to get the reweighing correction factors. These weights are propagated throughout the analysis and validated in the W+jets validation region, resulting in improved data/mc agreement. Table 5.5 encapsulates the relevant regions for calculating different corrections and validation of the applied corrections. Fig. 5.11

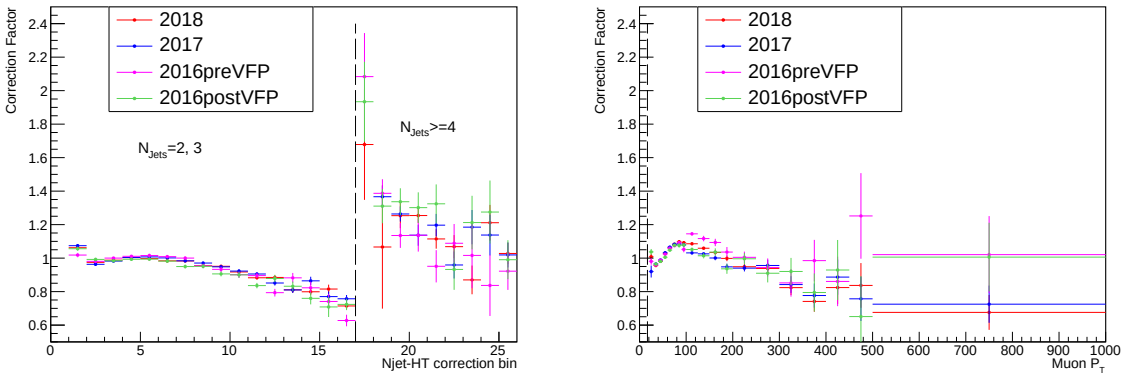


FIGURE 5.10: HT-based correction (left) and muon p_{T}-based correction (right) derived in W+jets CR for different eras. Statistical uncertainties only.

shows the correction factors in different muon M_T bins for all data-taking eras, and they were found to be consistent across eras.

Region	Selection criteria	Corrections
WJets CR	Preselection + 50 GeV $< M_T^{\mu} <$ 130 GeV + $\Delta R(j_0, j_1) > 2.6$	W+jets normalization, HT, and muon p_{T}-based correction.
WJets CR-extended	Preselection + $\Delta R(j_0, j_1) > 2.6$	Muon M_T-based correction.
WJets VR	Preselection + $\Delta R(j_0, j_1) < 2.6$ + $M_{j_0 j_1} < 50 \| \| M_{j_0 j_1} > 110$ GeV (veto SR)	Validating W+jets prediction.

TABLE 5.5: Selection criteria and applied corrections for WJets CR, extended WJets CR, and WJets validation region.

Key variables used in ML training are in good agreement with the data after applying the corrections. Fig. 5.13 shows good data/mc agreement in WJets CR for the key variables used in training the ML algorithm for the full Run-2 dataset. Signals are also overlaid for a few representative mass points to show negligible signal contamination in the CR.

5.5 W+jets background validation

To validate the method of estimating W+jets background developed in the CR, an orthogonal set of events is used, satisfying the criteria $\Delta R(j_0, j_1) < 2.6$. This validation region partially overlaps with our preselected signal region. To avoid biases, we vetoed the events with dijet invariant mass (M_{j0j1}) between 50 GeV and 110 GeV. Also, no selection is applied on the muon M_T. W+jets background yield in this validation region is predicted from the stitched MC samples using all the corrections derived in W+jets CR.

Key distributions shown in Figure 5.14- 5.16 for the full Run-2 dataset demonstrate good data/mc agreement in the validation region with good shape agreement. This indicates that

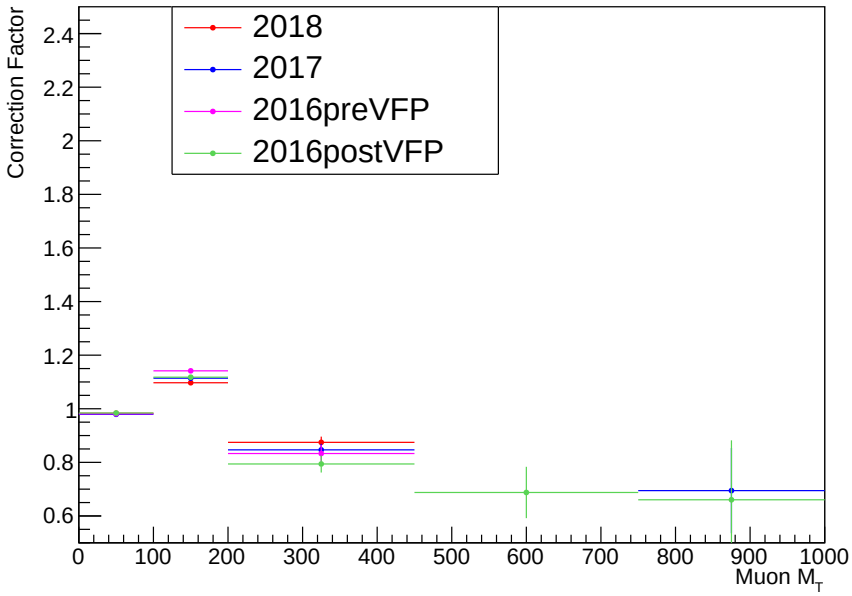


FIGURE 5.11: Muon M_T based correction derived in extended WJets CR as shown in Table 5.5 for different data taking eras. Statistical uncertainties only.

W+jets MC samples can be used to predict the normalization and shape of this dominant background in the signal region.

Other background contributions, such as $t\bar{t}$ semileptonic, single top, Z+jets, along with diboson processes (WW, WZ, and ZZ) that are subdominant, are estimated in SR directly using the MC samples, i.e., the normalization (cross-section is taken from the theory) and shape are taken directly from the respective MC samples.

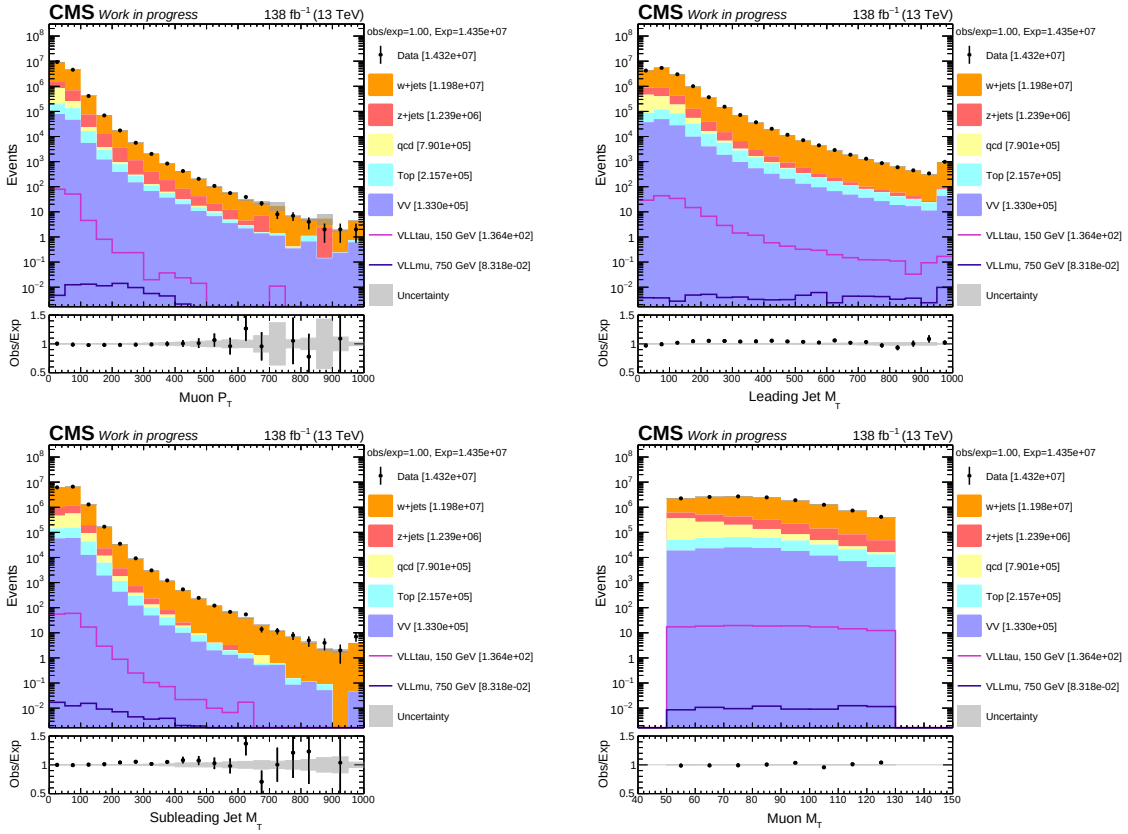


FIGURE 5.12: The distributions of muon p_{T}, muon M_T, leading jet M_T, subleading jet M_T in W+Jets CR events for the combined 2016-2018 dataset. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents statistical uncertainties only.

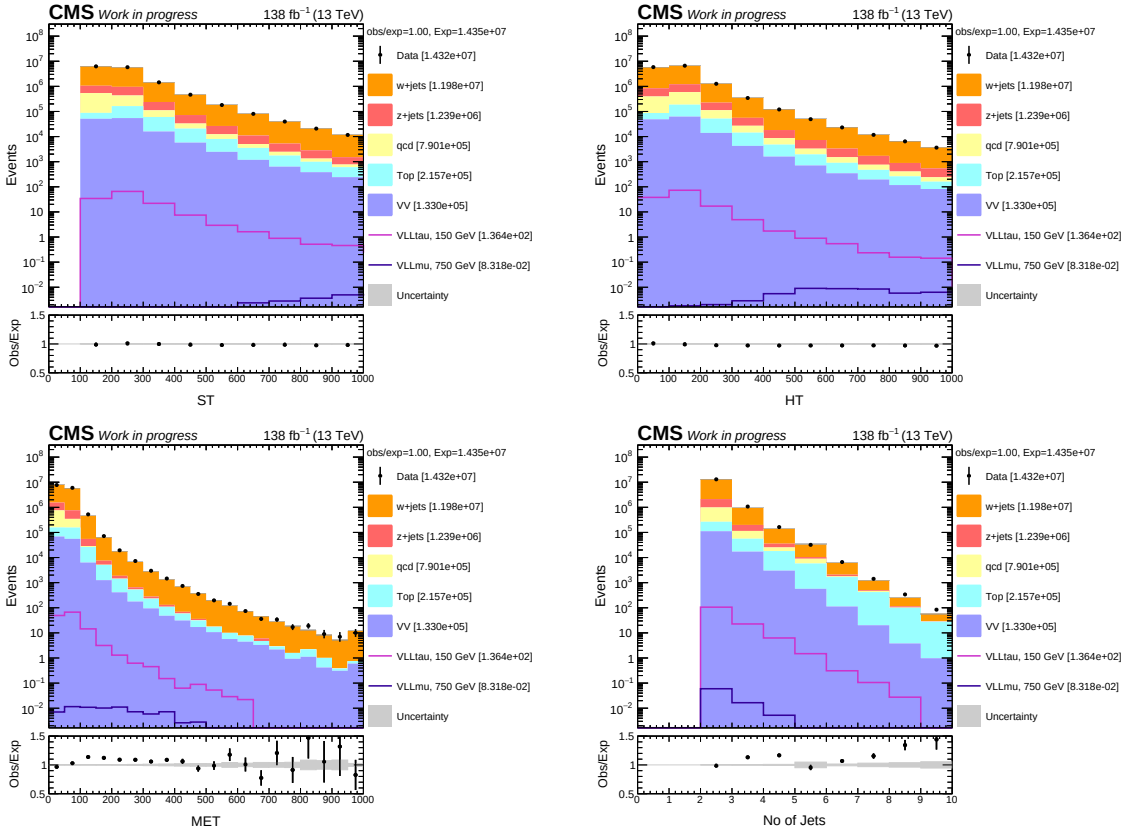


FIGURE 5.13: The distributions of S_T, H_T, $p_{\mathrm{T}}^{\mathrm{miss}}$, jet multiplicity in W+Jets CR events for the combined 2016-2018 dataset. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents statistical uncertainties only.

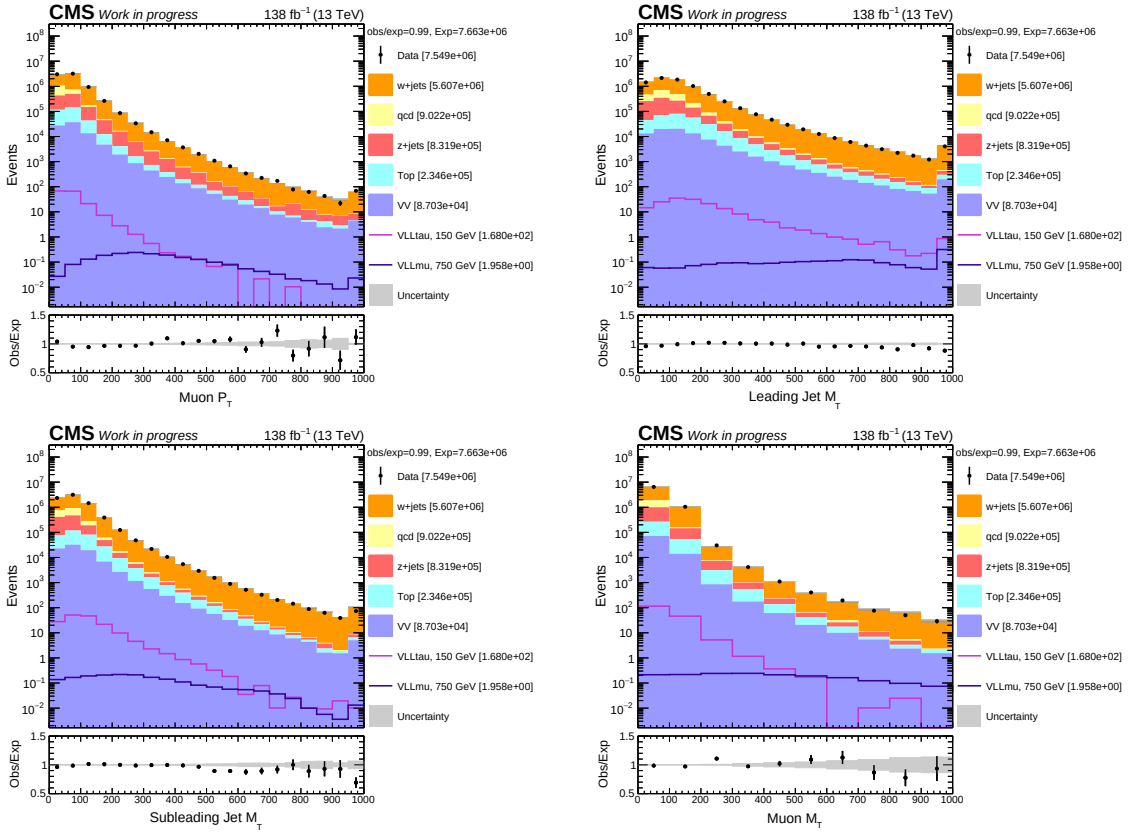


FIGURE 5.14: The distributions of muon p_{T}, muon M_T, leading jet M_T, subleading jet M_T in W+Jets VR events for the combined 2016-2018 dataset. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents statistical uncertainties only.

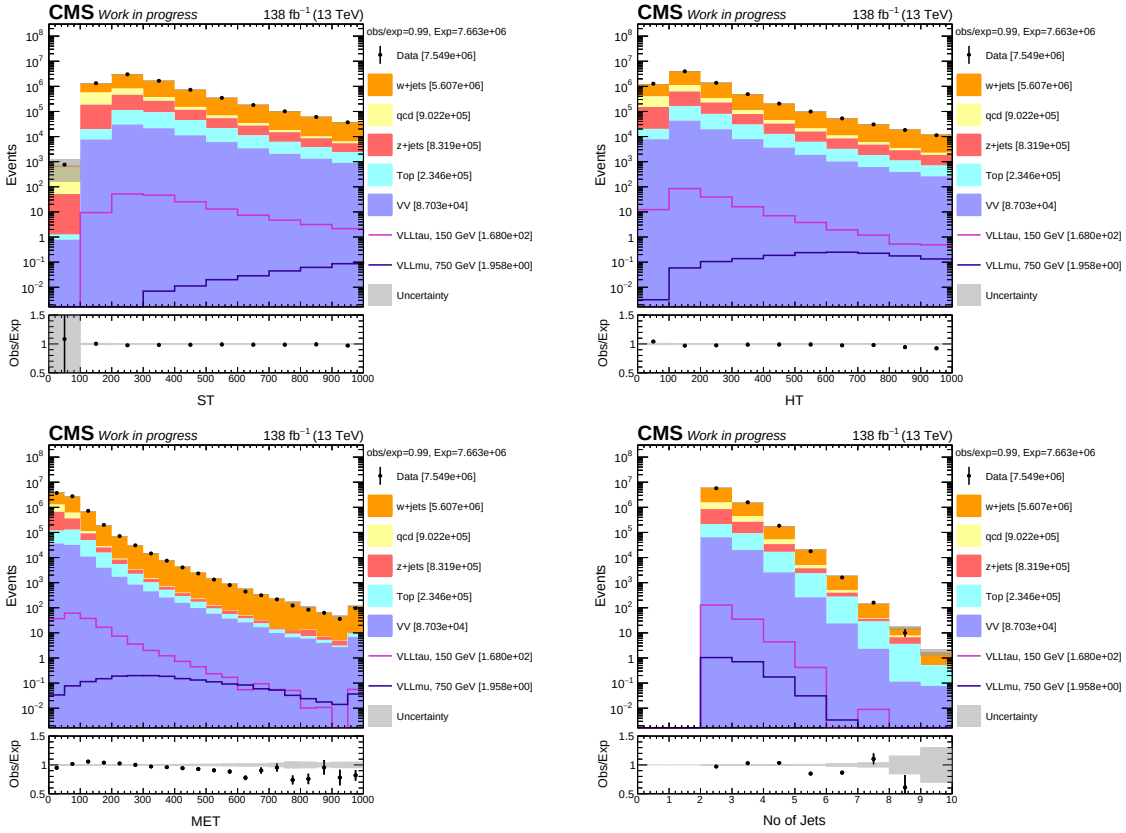


FIGURE 5.15: The distributions of S_T, H_T, $p_{\mathrm{T}}^{\mathrm{miss}}$, jet multiplicity in W+Jets VR events for the combined 2016-2018 dataset. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents statistical uncertainties only.

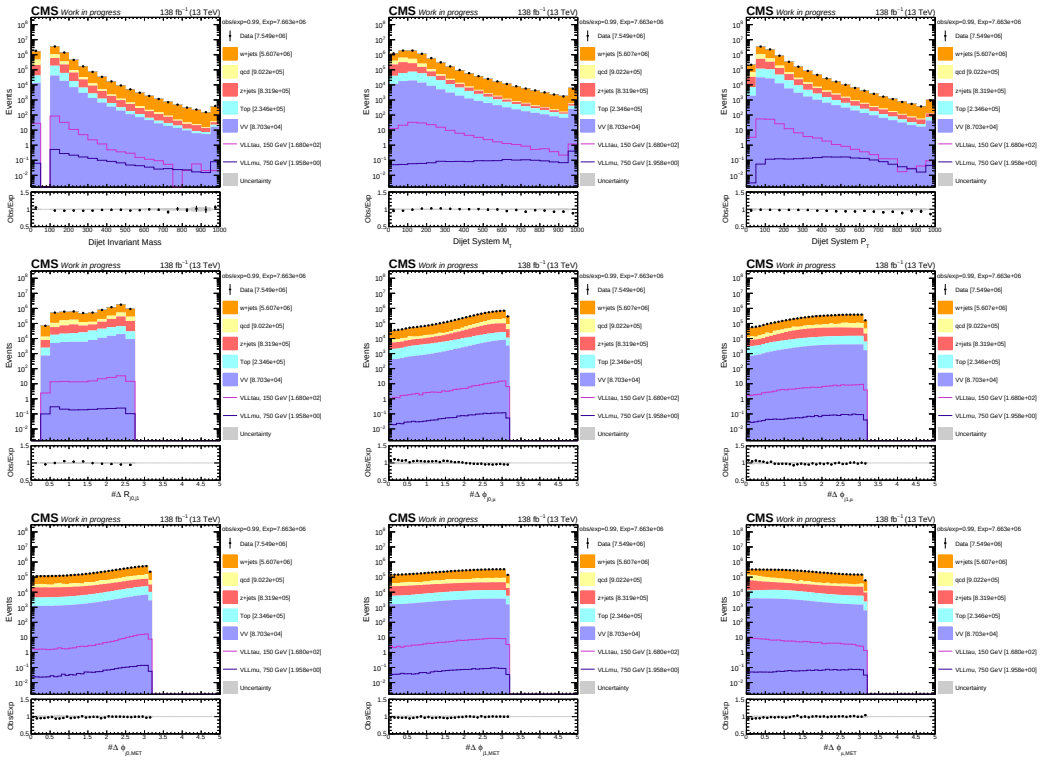


FIGURE 5.16: Key distributions of the dijet system and angular variables between objects in W+Jets VR events for the combined 2016-2018 dataset. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents statistical uncertainties only.

5.6 Systematic uncertainties

Precise estimation of SM backgrounds is necessary to establish the credibility of any potential discovery or exclusion claim. The estimation is limited by uncertainties that broadly fall into statistical and systematic categories. Statistical uncertainties arise from the finite size of data and simulated event samples, leading to fluctuations that can mimic signals. Systematic uncertainties, on the other hand, originate from imperfect knowledge of detector response, modeling of background processes, theoretical inputs such as parton distribution functions (PDFs), and assumptions in the signal modeling.

Backgrounds are estimated using different MC samples as described earlier. Statistical uncertainties due to the finite size of the available events in the MC samples are completely uncorrelated across the four eras of data-taking periods. However, for the bulk of distributions, systematic uncertainty plays a significant role, as does the statistical uncertainty in the measurements. Several systematic uncertainty sources are considered to account for differences between MC and data events. These uncertainties apply to signal processes and background contributions due to irreducible processes estimated using simulated samples.

The following per-object and per-event systematic uncertainties account for the differences in the modeling of pile-up, single muon trigger efficiency, jet energy scale, and jet energy resolution between data and MC events. Corrections factor derived in the W+jets CR in HT, $p_{\mathrm{T}}^{\mathrm{miss}}$, and M_T^{μ} also have associated uncertainties. The statistical uncertainties due to the finite size of the sample used to derive these correction factors were propagated through the analysis and taken into account to estimate the events in SR.

- **Trigger Efficiency:**
  The individual lepton trigger efficiencies in data and MC as illustrated in Appendix 11.1.1are assigned systematic uncertainties of 1-4% per lepton leg, and the final uncertainty on the trigger efficiency weight per event is conservatively estimated by varying these MC efficiencies up an down by 4%.

- **Jet energy scale and Jet energy resolution**
  The jet energy scale correction (JEC) and jet energy resolution (JER) corrections are applied at the per jet level, by shifting the corrected jet 4-momentum (up/down) within the recommended uncertainty ranges as derived in the orthogonal data samples in CMS centrally and recalculating all dependent observables, and propagated to the analysis.

- **Muon reconstruction,identification and isolation efficiency**
  The uncertainties on the reconstruction, identification, and isolation efficiency scale factors are applied as obtained by the dedicated tag-and-probe studies of the physics

object groups. They are cumulatively in the range of $< 2\%$ for muons. Custom identification (SIP3D and DEEPJET) criteria have associated scale factors close to unity, and their variations were also negligible. Hence, no associated uncertainty was propagated for this selection.

- **b-tagging efficiency**
  We also consider uncertainties on the b-tag efficiency scale factors applied to MC samples. They are divided into correlated and uncorrelated components across the data-taking eras. These uncertainty components are computed for light and heavy flavor(bc) jets to consider the uncertainty in the mis-tagging and tagging efficiencies between data and mc. This yields 10 uncertainties (nuisance parameters) for the full Run-2 dataset considered in this analysis.

- **Pileup**
  The pileup reweighting uncertainty is evaluated by varying the minimum bias cross-section used in the reweighting procedure up and down by 5%, and applied to all MC-based backgrounds.

- **W+jets HT , muon p_{T}, and muon M_T reweighting**
  Size of the HT based correction, muon p_{T} based corrections, and muon M_T based corrections due to finite sample size are propagated through the analysis by taking the up and down variation of those correction respectively. These uncertainties are treated as uncorrelated across the data-taking periods. For muon-pt based corrections, a similar variation is taken into account and treated uncorrelated across the data taking periods.

- **PDF and QCD scale uncertainties**
  The Parton Distribution Function (PDF) describes the probability of finding a parton (quark or gluon) carrying a fraction x of the proton's momentum at a particular energy scale Q^2. PDFs scale uncertainties stem from our incomplete knowledge of the internal structure of hadrons, like the proton, which are composed of quarks and gluons, collectively called partons. These functions are often tuned by the experimental data and have their associated uncertainties. In this analysis, the PDF uncertainties are evaluated using the PDF4LHC recommendations and computing the impact using different PDF sets and their variations on the analysis observable.

  On the other hand, the QCD factorization and renormalization scales, μ_F and μ_R, are non-physical parameters introduced in the perturbative QCD calculation of cross

sections. Although physical observables should be independent of these scales, fixed-order calculations have a residual dependence due to truncation at finite order. The envelope of estimated background (or signal) prediction is derived by varying the renormalization and factorization scale (μ_R, μ_F) in these pairs of variations: $(0.5\mu_0, 0.5\mu_0)$, $(2\mu_0, 2\mu_0)$, $(2\mu_0, 0.5\mu_0)$, and $(0.5\mu_0, 2\mu_0)$.

Depending on the process, the PDF and QCD scale uncertainties have a $\sim$5-10% effect on the background and signal yield. However, the effect is only in the normalization; no shape-dependent variations were observed.

The impact of statistical and systematic uncertainties is quantified in the final discriminant (observable) used to conduct the counting experiments. In the next chapter, neural network-based final discriminant is discussed, and the impact of such systematic uncertainties is measured in the final SR bins, and described later in Section 7.2.

Chapter 6

Search for VLL in μjj final state: Event categorization

Events selected using SR preselection criteria as described in Section 5.3 can further be reduced to a subset of events where the probability of finding the signal can be maximized. The signal topology of μjj final states indicates that multiple variables constructed using the selected muon, jets, or event-level properties could be sensitive to the presence of signal. For example, enhancement in the tails of S_T, scalar sum of muon p_{T}, selected jets p_{T}, and $p_{\mathrm{T}}^{\mathrm{miss}}$, could indicate the presence of such new particles. It is worth to emphasize that the categorization based on S_T into different ranges to improve the $\frac{S}{\sqrt{B}}$ (each bin acts as an independent set of counting experiments) only sets a *linear* boundary in the S_T phase space, where each signal and background instances can be defined by one number, that is S_T (other variables used to define the SRs can be thought as a boolean variable that encode whether the event is signal-like, or background-like). Sometimes, events are divided based on multiple properties of the signal or background processes to capture the full analysis phase space. For example, the pool of events selected using SR preselection criteria can be divided into different ranges of $\Delta\mathrm{R}(j_0, j_1)$ or M_T^{μ} to capture the change in background composition (or signal characteristics). Then the S_T distribution in each of these subsets of events can be scrutinized to test the presence of a signal. Note that, in this strategy, we are expecting to exploit the correlation between these physics variables, and constructing *linear* boundaries in this multidimensional phase-space created by $\Delta\mathrm{R}(j_0, j_1)$–$M_T^{\mu} - S_T$. Such multi-bin splitting aims to optimally separate signal from background while capturing the composition of the background profile.

However, the multidimensional phase-space can be reasonably complex, where signal and background events may not be separable optimally by a linear decision boundary; instead, a non-linear decision boundary may help to distinguish between different processes. One could think of transforming a variable(s) using the non-linear function, for example, $\exp -(S_T/M_T)$ may have more discriminating power to identify the signal. An extension of such transformation can be achieved using machine learning techniques with the advantage

of exploiting a much larger dimensional phase-space, correlation among the variables, and constructing a complex non-linear function that could potentially model the non-linear separation boundary (whose functional form we don't know) between signal and background events. Although machine learning algorithms can operate in both linear and non-linear regimes, we will restrict ourselves to the non-linear regime, which is extensively used in many studies described in this thesis.

Machine learning is not a new concept. But the recent **open source** software ecosystem, computing advances have made ML an essential tool in our analysis workflow. High energy physics datasets are a breeding ground for applying many ML algorithms due to their multiple features associated with any physics objects or processes (multidimensional phase space), a large number of events, and complexity. Supervised classification problem for identifying rare BSM signal to unsupervised clustering algorithm to identify EM showers on muon chambers by long lived particles, regressing energy of particles to correct the energy scale and resolution, using cutting edge transformer models to identify the source of jets to diffusion models generating simulation of p-p collision events, the usage of ML algorithms has undeniably seen exponential growth, with the latest addition of End-to-end physics object reconstruction, triggering an event, and detector optimization techniques. A set of such studies using ML algorithms applied to high energy physics problems is maintained in the HEP-ML-Living Review [81].

This chapter will discuss the usage of supervised ML techniques in the context of distinguishing VLL against the background. In supervised learning, models are trained using labeled data and take direct feedback to check whether they predict the correct output. The goal of supervised learning is to train the model so that it can predict the output when it is given new data. Since our output decision is either signal (1) or background (0), it is called a classification problem. There are a few challenges to design the ML strategy for this search,

- **Type of ML algorithms**: Deep Neural Networks (DNN) or Boosted Decision Trees (BDT)
- **Choice of classification**: Binary classification, or multi-label classification, as there is more than one background process in the SRs.
- **Wide range of VLL mass**: Two VLL hypotheses that are targeted have a wide range of VLL mass from 100 GeV to 1 TeV. The signal kinematics of a 1 TeV VLL is significantly different from 100 GeV VLL, even a 400 GeV VLL would differ significantly. So, should we use a parametric neural network strategy where the mass of VLL can be fed as a conditional parameter, or train networks of individual mass points?
- **Model topology**: Vector-like muons, and vector-like taus (singlet) are significantly different once the other decay modes ($VLL \rightarrow Z\ell(H\ell)$) start contributing, as the

selected muon may arise from the VLL, hence its p_{T} could be large compared to the muons that may be produced from the gauge bosons.

- **Dataset variability**: The Full Run-2 dataset comprises four eras due to their difference in reconstruction, alignment, and detector conditions. These four eras are 2016 preVFP, 2016 postVFP, 2017, and 2018. They can be considered as four independent datasets. Do we need separate training for each year? Or training using three years, and application on the other year? Or a single training for the whole Run-2 dataset to take advantage of improved statistics of training events? Additionally, the imbalance of events from different processes should be considered for a robust classification performance.

We used Deep Neural Networks (DNN) for this classification task. A detailed study on the choice of multi-classification vs binary classification against each background is performed, and binary classification is chosen for its better performance in the context of this classification problem. The intuitive explanation is, in multi-classifier, the network training is "awarded" even if the network learns to classify different background processes, rather than maximizing the separation between signal and each background process considered for training. In technical terms, the loss objective function (often cross-entropy function) gets rewarded for classifying different backgrounds correctly, ignoring the signal, which is not our aim for such network training. Instead, we can devise a strategy to train a set of binary networks to discriminate between signal and individual background. We can create a final discriminant by transforming the individual scores into a single score using mathematical operations.

To discriminate maximally between the signal and dominant SM backgrounds, such as W+jets, Z+jets, $\mathrm{t\bar{t}}$+jets, and QCD multijet in the μjj channel, a set of binary classifiers targeting each background for each signal mass hypothesis is trained. This results in a total of 4 classifiers to be trained on for each mass point, namely,

- wjets classifier
- zjets classifier
- ttjets classifier
- qcd classifier

At the end of the training, we had four neural network outputs (or scores) from these four classifiers, discriminating signal from their corresponding backgrounds. To construct the final discriminant score, we studied multiple strategies to combine these four classifier scores based on signal significance as a figure of merit (FOM), which is discussed later in detail. The strategy is explained in the schematic diagram in Figure 6.1.

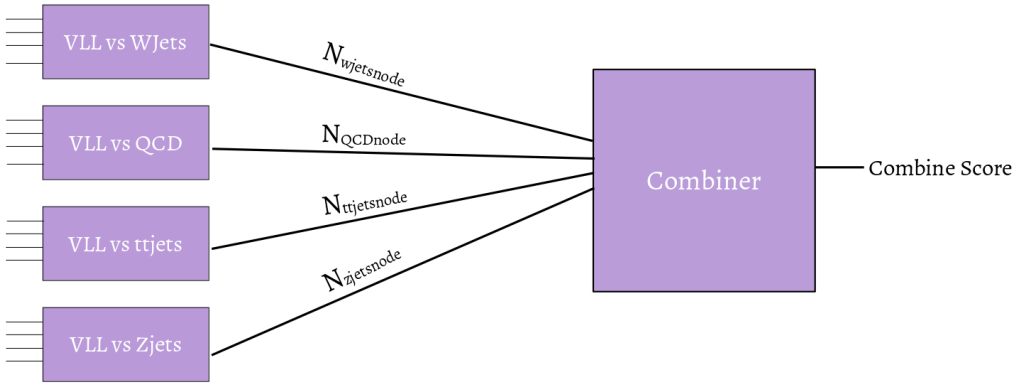


FIGURE 6.1: ML strategy devised for this analysis to train a set of binary classifiers to discriminate the signal maximally against the dominant SM backgrounds.

6.1 Discriminant training strategy

The training is performed in the Keras/Tensorflow software package, following the training strategy itemized below.

- This analysis probed vector-like tau model from 100 GeV to 400 GeV and vector-like muon model from 100 GeV to 1000 GeV. Given such a wide signal mass range, the properties of the signal vary considerably at high mass compared to the low mass, and have model dependence. Hence, we decided to perform the training for each signal mass hypothesis for both vector-like tau and muon models separately. At this point, we can count the total number of trainable classifier networks given a single full dataset training: four individual classifiers $\times$ eight mass points for vector-like taus and four individual classifiers $\times$ 12 mass points for vector-like muons. So, a total of 80 classifier networks were trained.

- We investigated the training strategy for this analysis in two directions. Firstly, we can train the networks for each data-taking period separately. Secondly, we can have a single training for the full Run-2 dataset. In the first approach, for training of the networks in a given year of the Run-2 period, signal and background MC samples of the other two years are used, i.e., for training a NN classifier to be used in the 2018 dataset, the training is done using samples generated for 2016 (preVFP and postVFP combined) and 2017 datasets. This ensures the statistical independence of the training and application samples and minimizes the possibility of overtraining. We check for overtraining by comparing training and application performance, and the ROC curves give close to similar AUC. Hence, we did not find any signs of overtraining. Additionally, we have also checked that this choice of using samples from independent years in the training does not compromise the performance while evaluating.

In the second approach, we performed a combined training using the full Run-2 dataset and compared it against the individual year training to assess the performance and simplify the training strategy. To demonstrate this, we apply full Run-2 training and individual training on the 2018 dataset and compare their performance, which is shown in Figure 6.7. We evaluate this for the vector-like tau model, for all four classifiers at low mass (150 GeV) and high mass (400 GeV). The ROC curve shows a slightly improved performance at low mass for full Run-2 training, which is expected because of the boost in statistics by combining all data-taking periods. The performance is comparatively the same at high mass for all training strategies. Thus, we chose to do a full Run-2 training for this analysis. Also, it is worth noting that individual year training performs almost similarly, which is a good reason to combine individual training into full Run-2 training to take advantage of more statistics.

- All control region selections are vetoed for selecting events in the training. These control regions thus serve as a cross-check while evaluating the performance of the trained networks.

A detailed study has been conducted to optimize the network to capitalize on the available signal and background training events. DNN model architecture and the choice of hyperparameters used for the training are described in Table 6.1. Different network architectures are also studied for various backgrounds, and comparable performance is observed. Thus, the same architecture is used for all classifiers targeting each background for simplicity. Classifier networks are trained for 60 epochs with a batch size of 1024. We did not observe any significant performance gain by training for longer epochs; actually, it led to overtraining. We also optimized the batch size and did not observe any significant change in performance for a set of batch sizes used in the optimization study. A low batch size value often resulted in the overtraining of the network and was thus avoided. Dropout layers with deactivating 30% of the neurons in different layers randomly (thus not changing their weights for that epoch) found to be useful to avoid overtraining.

6.1.1 Input features

There are a total of 25 training input variables for the μjj channel training. The input variables include object or event-level transverse momenta, invariant or transverse masses, and angular variables. For W+jets classification, extra jet-level features are used. DEEPJET Quark-Gluon discriminator score of the leading and subleading jets is used for this purpose. Hadronic decays of W produce a quark-antiquark pair in the signal topology. However, the source of jets in W+jets processes is mostly gluon radiation, but sometimes gluons can also

Name	Details
Hidden Layer 1	Dense neuron = 128, activation = relu, kernel initializer='he_normal'
Dropout	Dropout(0.3)
Hidden Layer 2	Dense neuron = 100, activation = relu, kernel initializer='he_normal'
Dropout	Dropout(0.3)
Hidden Layer 3	Dense neuron = 32, activation = relu, kernel initializer='he_normal'
Hidden Layer 4	Dense neuron = 8, activation = relu, kernel initializer='he_normal'
Output Layer	Dense neuron = 1, activation = sigmoid, kernel initializer='he_normal'
Epoch	60
Batchsize	1024

TABLE 6.1: Optimized DNN architecture and hyperparameters used in this analysis.

be fragmented into quark-antiquark pairs. Overall, these jets are found to be discriminative against the W+jets process. Before feeding to the network, the continuous spectra of Quark-Gluon discriminator score are divided into discrete score bins of $<$0.1, 0.1–0.4, 0.4–0.7,$>$ 0.7. We call the modified score the DEEPJET QG category. This variable may not be well modeled in data. Thus, the discretization may help reduce the mismodeling effect and related systematics. Training variables are listed in Table 6.2. Since the muon p_{T}, $p_{\mathrm{T}}^{\mathrm{miss}}$, and angular distribution between muon and $p_{\mathrm{T}}^{\mathrm{miss}}$ is already provided, M_T^{μ} is not explicitly provided, as the network has the power to learn the correlation.

6.1.2 Training performance

W+jets classifiers performance

Fig 6.3 shows the neural network output score and ROC curve for W+jets training for two VLL mass points of 100 (left) and 400 (right) GeV in the vector-like tau model. Training and validation performance significantly improved for high VLL mass as the kinematic properties differ from the W+jets process. At low mass, signals look very similar to the W+jets background, resulting in degraded performance.

Fig 6.2 illustrates the accuracy and loss between the training and validation datasets during training. The confusion matrix for W+jets training is also shown. All of them are shown for two VLL mass points of 100 (left) and 400 (right) GeV training in the vector-like tau scenario.

Variable type	Variables
Object variables	μ p_{T}, j_0 p_{T}, j_1 p_{T} μ eta, j_0 eta, j_1 eta DEEPJET QG score category of leading jet DEEPJET QG score category of subleading jet
Event variables	$p_{\mathrm{T}}^{\mathrm{miss}}$, HT, N_j average DEEPJET QG score category of the event Dijet system p_{T}
Angular variables	$\Delta R(j_0, j_1)$, $\Delta\phi(p_{\mathrm{T}}^{\mathrm{miss}}, j_0)$, $\Delta\phi(p_{\mathrm{T}}^{\mathrm{miss}}, j_1)$ $\Delta\phi(p_{\mathrm{T}}^{\mathrm{miss}}, \mu)$, $\Delta\phi(j_0, \mu)$, $\Delta\phi(j_1, \mu)$ $\Delta\phi$(Dijet system, μ), $\Delta\phi$(Dijet system, $p_{\mathrm{T}}^{\mathrm{miss}}$),
Mass variables	$M_T^{j_0}$, $M_T^{j_1}$ $M_T^{\text{Dijet system}}$

TABLE 6.2: Input variables used for the NNs trained for the VLL singlet model.

Classifier performance overview for all backgrounds

Fig 6.4– 6.5 shows the performance of all background-specific classifiers for 2017 training for 100 GeV and 400 GeV VLL-tau mass training. Performance is significantly improved for all background classifiers in the 400 GeV mass training compared to the 100 GeV mass training.

This is expected as the properties of signal events vary significantly as the VLL mass differs from the W boson mass. For the QCD multijet background, the high $p_{\mathrm{T}}^{\mathrm{miss}}$ due to significant momenta carried by the neutrino from VLL, and hard muon p_{T} properties are discriminative features. However, discriminating against QCD multijet backgrounds is relatively easier for all the mass points than for other backgrounds. Performance in the tt+jets classifier is degraded at low mass training due to a similar event topology as the signal, where the semileptonic decay of $t\bar{t}$ produces a lepton and jets coming from the leptonic and hadronic decay of the W boson, respectively. However, as the VLL gets heavier than the top mass, the momentum share between the decay products changes, and the training performance is improved. The training procedure focused on reducing the W+jets background, as this is the most dominating background in this search.

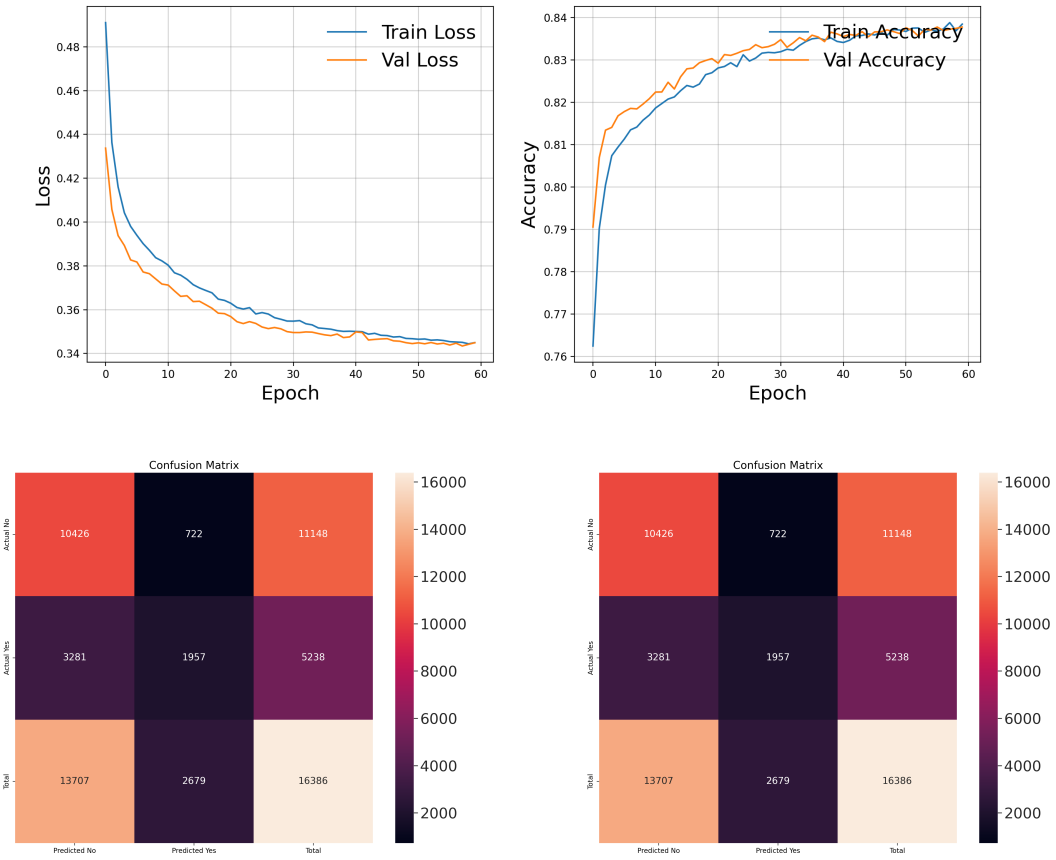


FIGURE 6.2: Training and testing (validation) performance for the W+jets classifier in 2018. The upper panel shows the loss and accuracy metrics during training, and the lower panel shows the confusion matrix for 100 GeV (left) and 400 GeV (right) VLL-tau mass training.

To get an overview of the testing performance for all SM background processes in different years, ROC curves of different SM processes are overlaid, and AUC values are compared. Fig. 6.6 shows the ROC curve at testing using 2016 and 2017 training. QCD discrimination is maximum, and tt+jets events are found to be challenging to discriminate due to identical event topology at low mass VLL training. At high VLL mass, discrimination power increases against all the SM background processes.

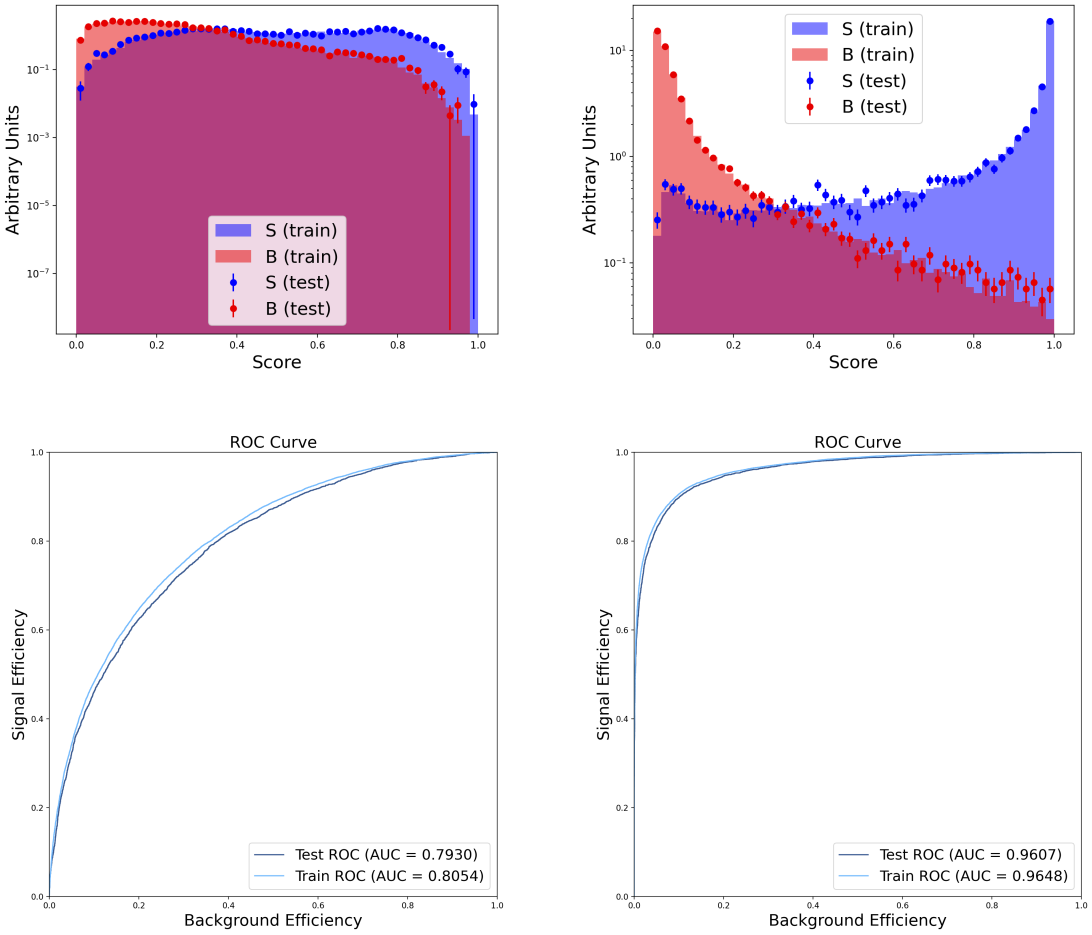


FIGURE 6.3: Training and testing (validation) performance for the W+jets classifiers in full Run-2 dataset training. The upper panel shows neural network output, and the lower panel shows the ROC curve for 100 GeV (left) and 400 GeV (right) VLL-tau mass training.

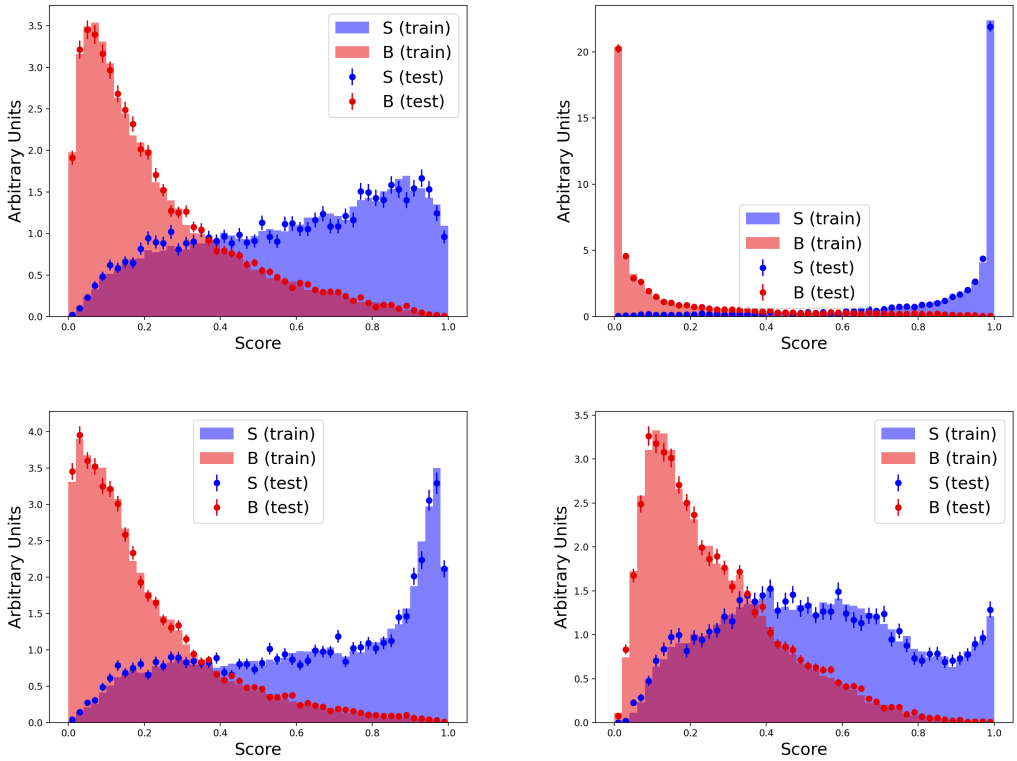


FIGURE 6.4: Training and testing network score for the four background-specific classifiers in 2017. Neural network scores for W+jets classifier (upper left), QCD multijet classifier (upper right), Z+jets classifier (lower left), and tt+jets classifier (lower right) are shown for 150 GeV VLL-tau mass training.

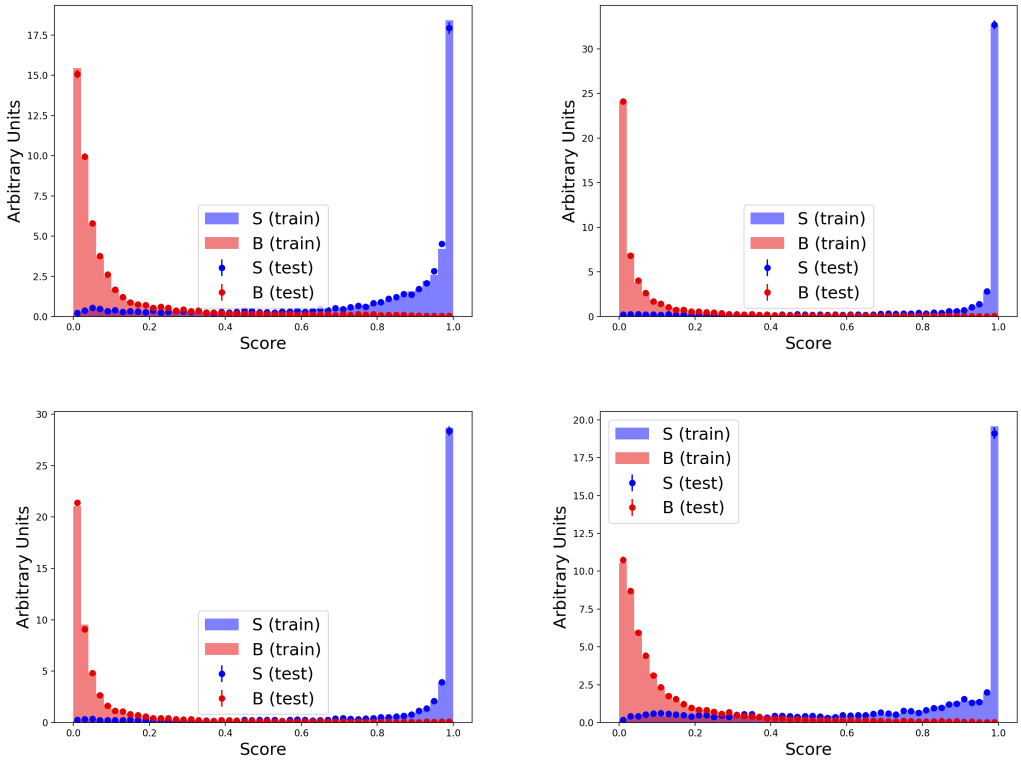


FIGURE 6.5: Training and testing network score for the four background-specific classifiers in 2017. Neural network scores for Wjets classifier (upper left), QCD multijet classifier (upper right), Zjets classifier (lower left), and ttjets classifier (lower right) are shown for 400 GeV VLL-tau mass training.

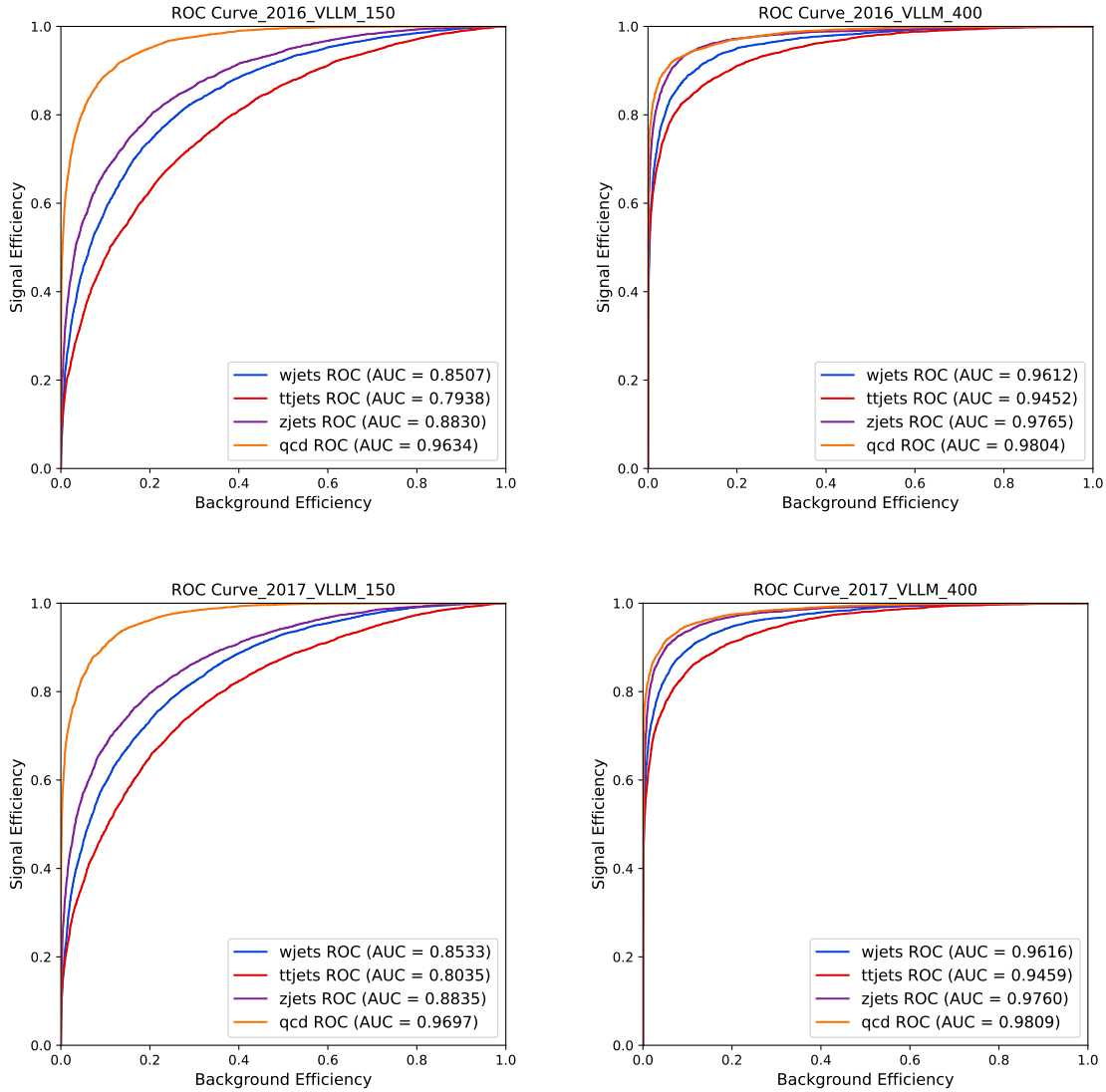


FIGURE 6.6: Testing performance for the four background-specific classifiers in 2016 and 2017 in terms of the ROC curve. The upper panel shows the ROC curve for all processes in 2016, and the lower panel shows the ROC curve for all processes in 2017 for 150 GeV (left) and 400 GeV VLL-tau mass training.

Full Run-2 training vs year specific training

Finally, the testing performance for full Run-2 training is compared with year-specific training for vector-like tau mass of 100 GeV and 400 GeV for each background classifier. The plots illustrate negligible differences in performance between year-specific and full Run-2 training (indicated by different colors) for 100 GeV (solid) and 400 GeV (dashed) cases.

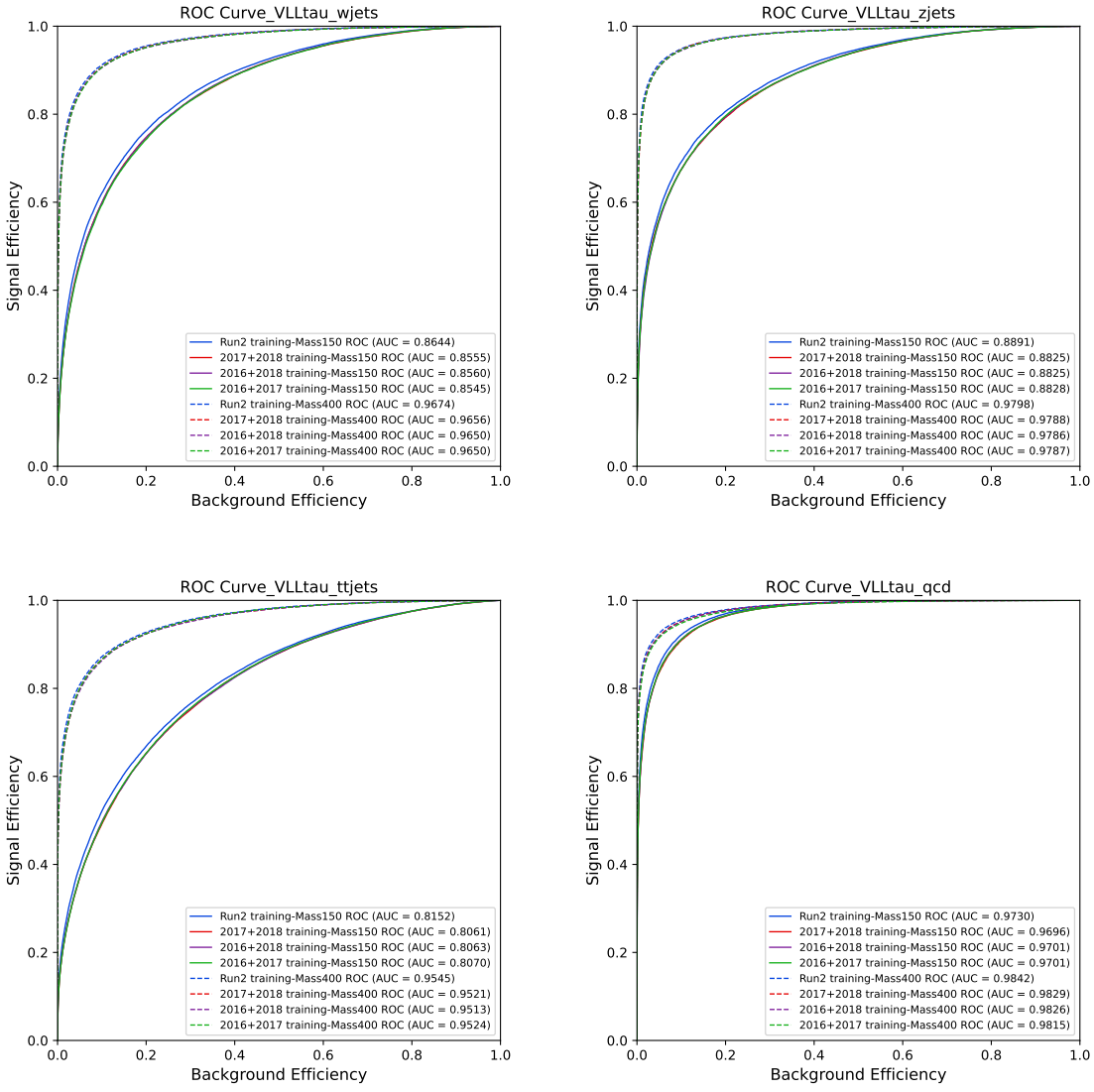


FIGURE 6.7: Comparing full Run-2 training strategy against training performed for an individual data-taking period after evaluating the trained network on the preselected SR events from the 2018 dataset. ROC curve is shown for a low mass (150 GeV) in solid line and high mass (400 GeV) in dashed line for W+jets, Z+jets, tt+jets, and QCD classifier in the vector-like tau scenario.

Several distributions of various training input variables are illustrated in Figure 5.13–5.14 for different control and validation regions. Backgrounds for all considered input variables are well modeled and found to be in good agreement with the data. As the training is complete, neural network scores shown in the following few sections implicitly mean scores at evaluation of the network on events satisfying the SR preselection criteria.

6.1.3 Feature importance

Figure 6.9 demonstrates the importance of features in the training procedure for both third and second generation VLL models, for a few representative mass points. The feature permutation method of the scikit-learn package is used to rank input variables for NN training.

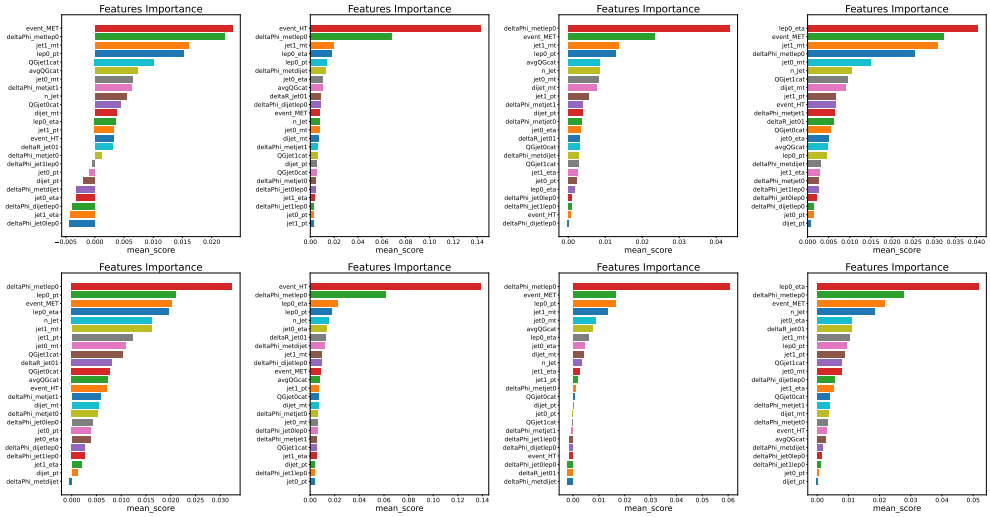


FIGURE 6.8: Feature importance for 150 GeV mass training in third generation VLLs model (upper panel) and second generation VLLs model (lower panel) for wjets, qcd, ttjets, and zjets classifier (from left to right).

6.2 Constructing the final discriminant

Different strategies were investigated for the optimization study to construct the final discriminant score combining the neural network output of 4 background-specific classifiers. We studied five different ways of combining the output of the individual classifier as follows,

- Option 1: CombineNNScore_1 = $N_{wjets} * N_{ttjets} * N_{zjets} * N_{qcd}$
- Option 2: CombineNNScore_2 = $1 - \sqrt{\frac{1}{4} \times [\sum_i (1 - N_i)^2]}$, i = wjets, ttjets, zjets, and qcd
- Option 3: CombineNNScore_3 = $\frac{N_{wjets} * N_{ttjets} * N_{zjets}}{N_{wjets} * N_{ttjets} + N_{ttjets} * N_{zjets} + N_{zjets} * N_{wjets}} + N_{qcd}$

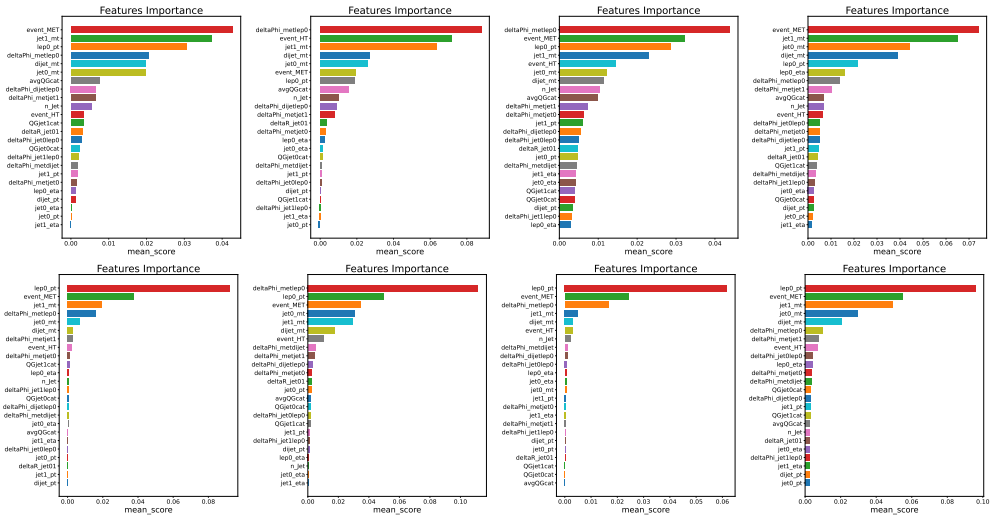


FIGURE 6.9: Feature importance for 400 GeV mass training in third generation VLLs model (upper panel) and 750 GeV mass training for second generation VLLs model (lower panel) for wjets, qcd, ttjets, and zjets classifier (from left to right).

- Option 4: CombineNNScore_RMS = $\sqrt{\frac{N_{wjets}^2+N_{ttjets}^2+N_{zjets}^2+N_{qcd}^2}{N_{wjets}+N_{ttjets}+N_{zjets}+N_{qcd}}}$
- Option 5: CombineNNScore_WjZj = $\frac{1}{2}(\frac{N_{wjets}*N_{zjets}}{N_{wjets}+N_{zjets}})$

Fig. 6.10- 6.11 shows the neural network score for options 1 and 2, with the events passing the SR preselection criteria in the full Run-2 dataset. Luminosity-weighted scores are shown on the left for 100, 150, and 400 GeV VLL mass training in a vector-like tau scenario. Normalized histograms (right) of the signal and total background stack are shown to compare the shape between the signal and total background in the computed combined scores. This demonstrates the variation of the discriminating power of the same combining strategy as a function of VLL mass.

6.2.1 Choosing the best combining strategy: Asimov significance

We plotted Asimov significance as Figure of Merit (FOM), described in Eq. 6.1 for all combining strategies where N_{sig} denotes the number of signals and N_{bkg} denotes the background yield above a threshold cut on the spectra. Fig. 6.12- 6.13 shows the Asimov significance as a function of threshold cut on the combined score in the full Run-2 dataset for 100-400 GeV VLL mass training in the vector-like tau scenario. The plots demonstrate that option 2 of the combining strategy at low mass can achieve the highest significance. Combining strategy is important below 200 GeV mass as the classifier performance is moderate, but as we approach higher mass training, similar significance can be achieved by three combining strategies (1, 2, and 5) as the training performance improved significantly against all backgrounds.

$$\text{Asimov Significance} = \sqrt{2 * ((N_{sig} + N_{bkg}) * log((1 + \frac{N_{sig}}{N_{bkg}}) - N_{sig}))} \tag{6.1}$$

Option 2 is chosen as the final discriminant based on the optimization study we performed based on Asimov significance. Option 2 also produced the best expected limit compared to the other combining options.

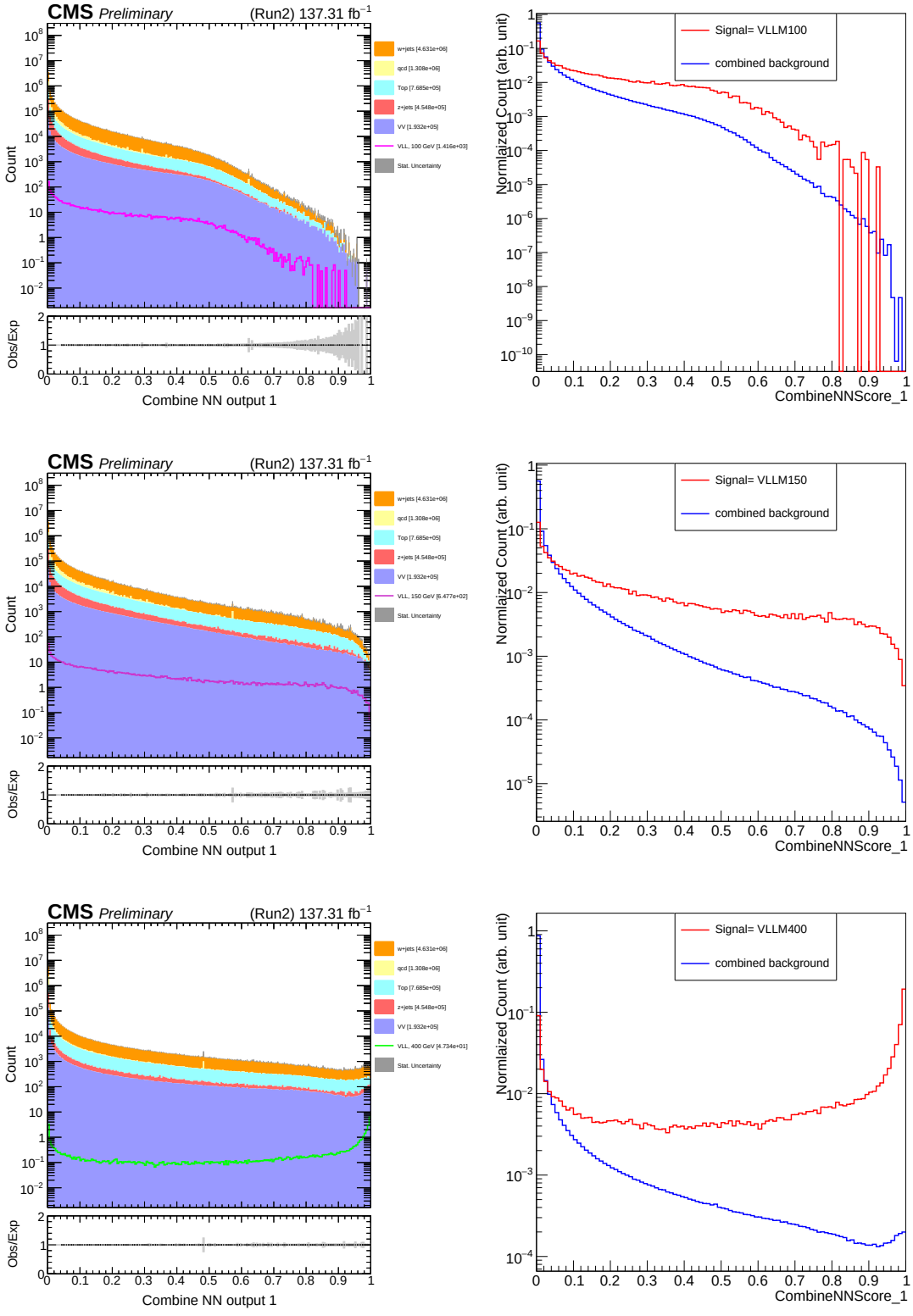


FIGURE 6.10: Combine NN Score (option 1) distributions (left) and normalized histograms of signal and total background (right) distributions in the full Run-2 dataset. The plots are for 100 GeV (first row), 150 GeV (second row), and 400 GeV (third row) VLL mass training in a vector-like tau scenario. Statistical uncertainties only.

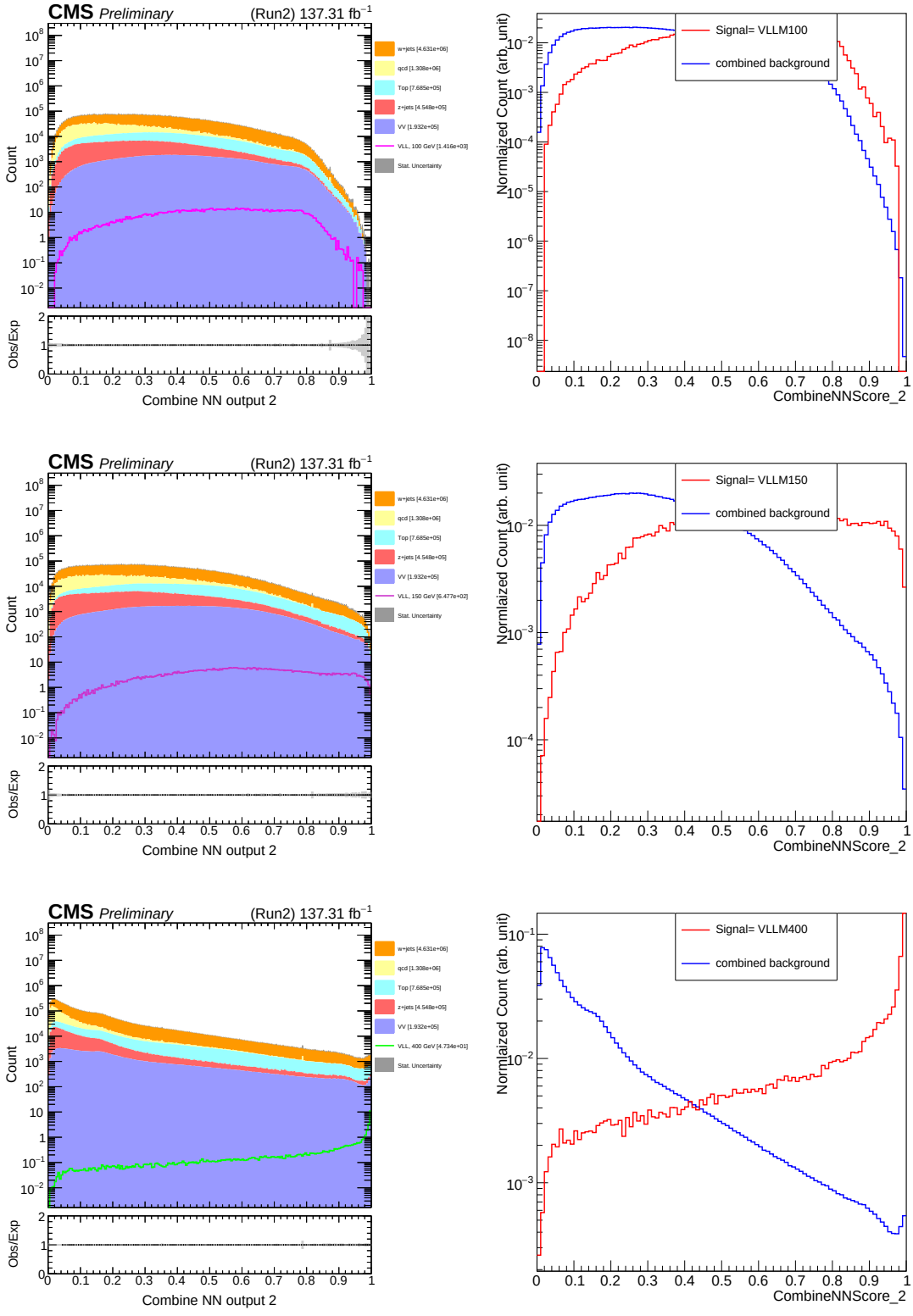


FIGURE 6.11: Combine NN Score (option 2) distributions (left) and normalized histograms of signal and total background (right) distributions in the full Run-2 dataset. The plots are for 100 GeV (first row), 150 GeV (second row), and 400 GeV (third row) VLL mass training in a vector-like tau scenario. Statistical uncertainties only.

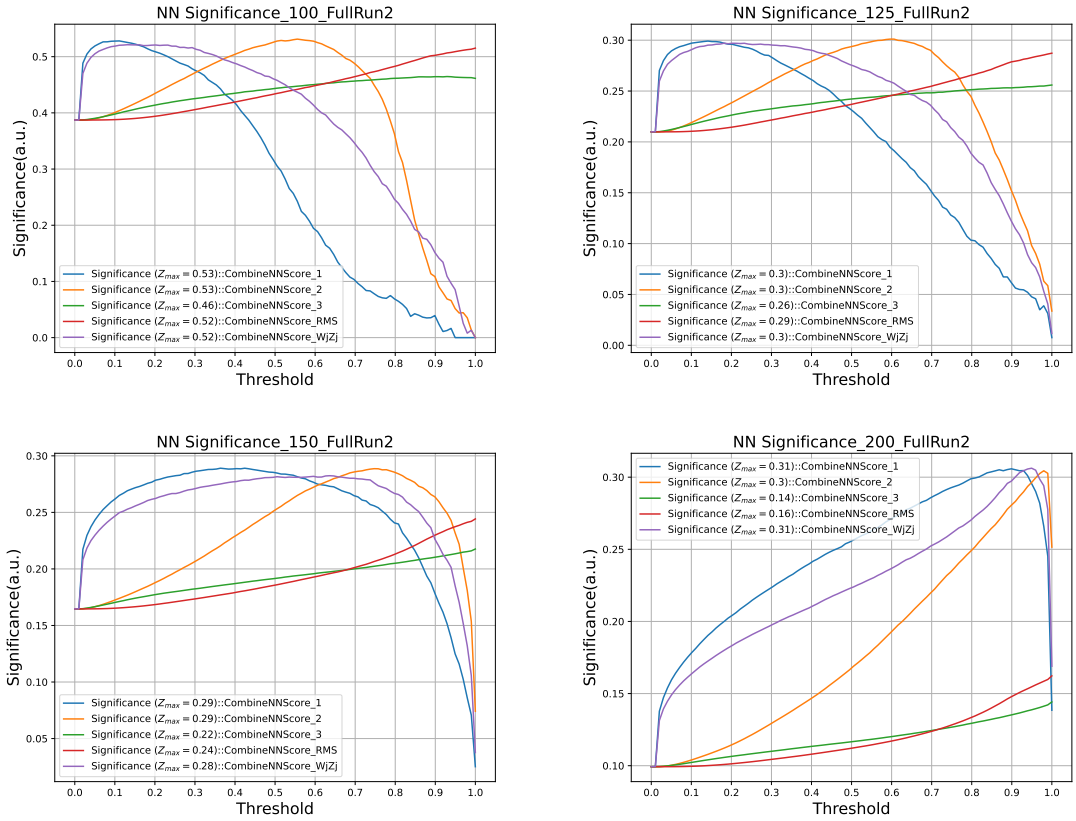


FIGURE 6.12: Asimov significance calculated as a function of threshold cuts on the combined NN score spectra for various combining strategies listed above. Full Run-2 signal and total background histograms are used to calculate the significance in the vector-like tau scenario. The upper panel shows the significance curves of the various combined scores from 5 different combining strategies for 100 GeV (left) and 125 GeV (right) VLL mass training. The lower panel shows the significance curves for 150 GeV (left) and 200 GeV (right) VLL mass training.

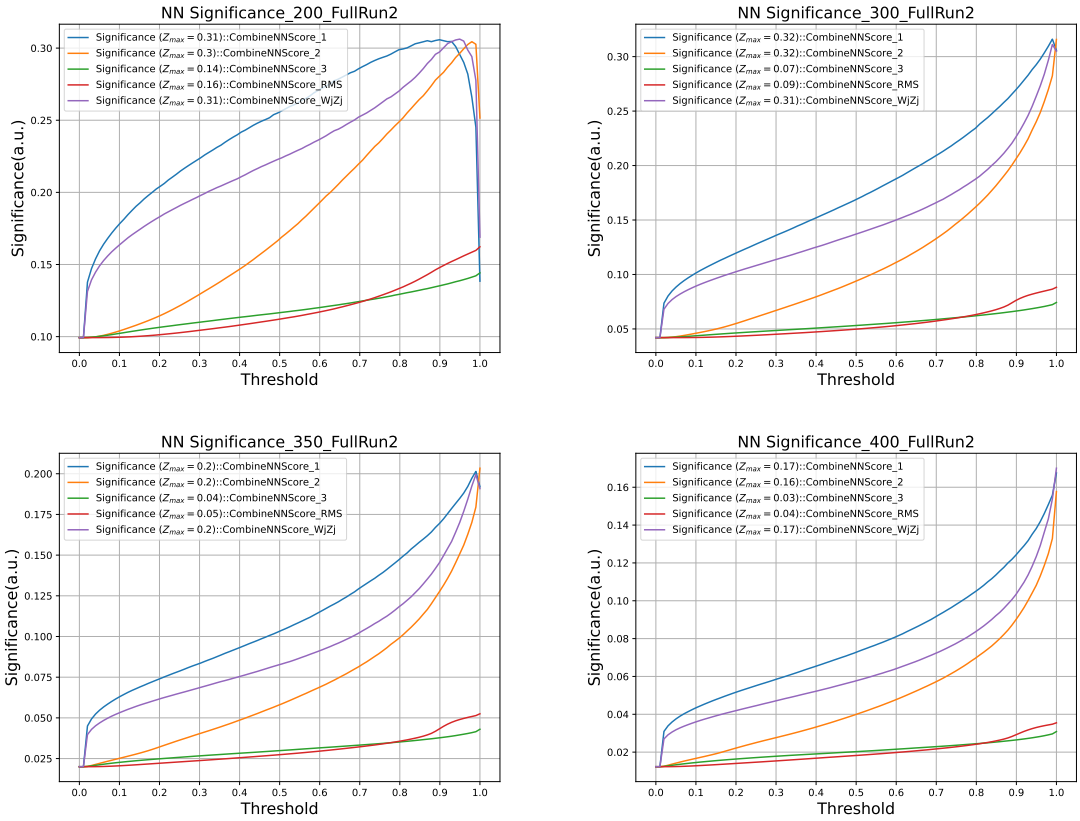


FIGURE 6.13: Asimov significance calculated as a function of threshold cuts on the combined NN score spectra for various combining strategies listed above. Full Run-2 signal and total background histograms are used to calculate the significance in the vector-like tau scenario. The upper panel shows the significance curves of the various combined scores from 5 different combining strategies for 250 GeV (left) and 300 GeV (right) VLL mass training. The lower panel shows the significance curves for 350 GeV (left) and 400 GeV (right) VLL mass training.

6.3 Validation using data

An important aspect of the neural network strategy is to validate its performance against the data. As expected, the signal is pushed right in the combined score, while backgrounds are populated mostly towards the left side of the combined neural network score spectrum. To validate the neural network strategy, events are selected with 0.5<Combine NN score < 0.7, and data is not looked at beyond > 0.7 for the blind strategy that has been followed so far. These give orthogonal regions to final signal regions to check data/mc agreement in neural network score and the underlying physics variable. This region is highly populated with background-like events and not too far from signal regions, i.e., the background events have similar characteristics to a signal. Hence, this region can be used to check data/mc agreement in the neural network score and key training variables used in this analysis.

6.4 Data MC agreement in preselected signal region

Distributions of key training variables in the preselected signal region are shown in Figure 6.16. Predicted SM background can describe the data well in all distributions. A subset of events satisfying a specific threshold of combine NN score for their respective model-mass networks are used to test the presence of VLL signal.

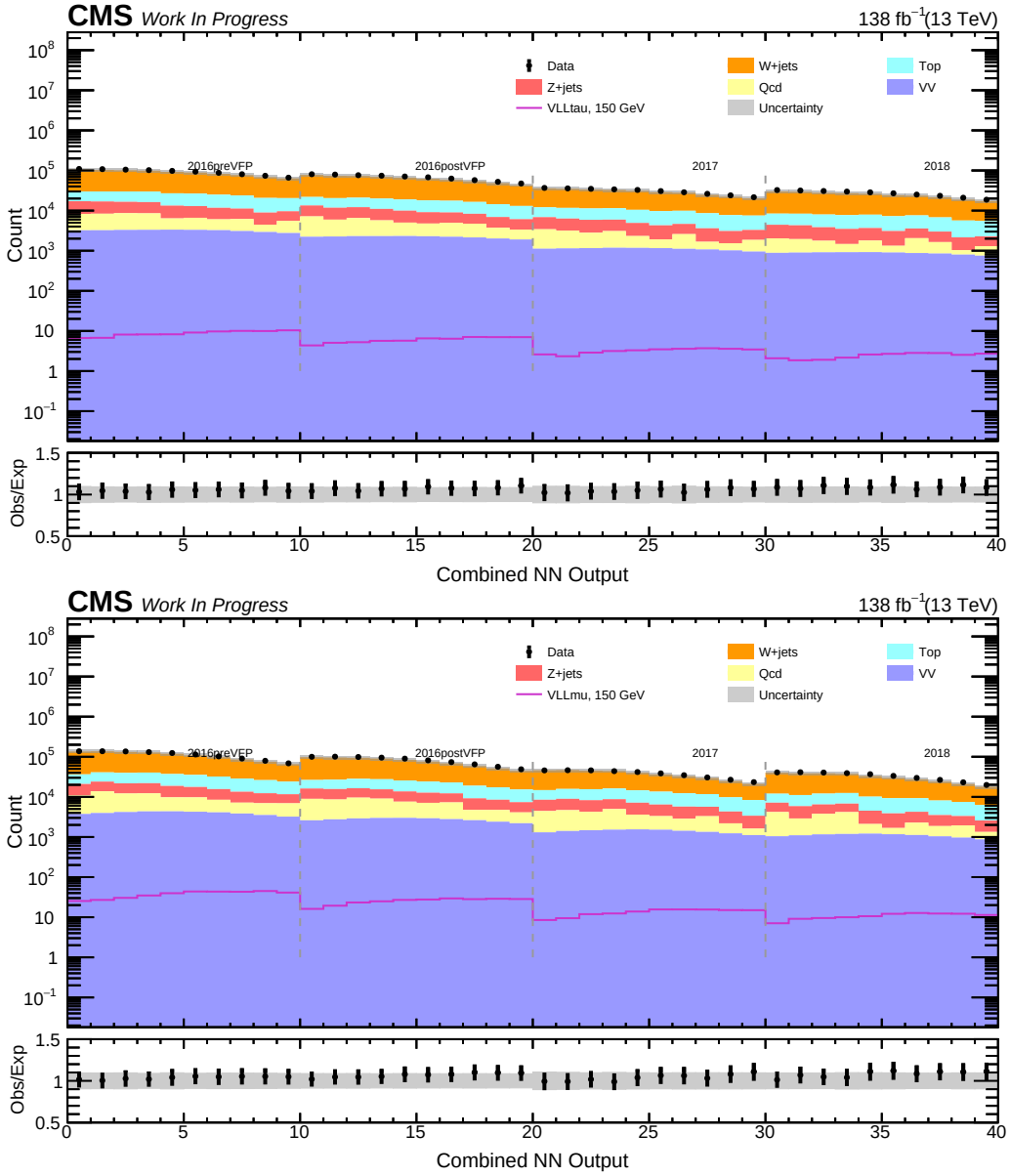


FIGURE 6.14: Distribution of the combine NN score with preselected SR events satisfying 0.5< Combine NN Score_2 < 0.7 cut for vector-like taus (upper) and vector-like muons (lower) of 150 GeV mass. These plots are demonstrated with bins with an equal width of 0.02 from 0.5 to 0.07, resulting in 10 bins per data-taking period. This region in data is highly populated with background-like events, hence providing a good validation region to check data/mc agreement in NN input variables. Data is shown in black marker, backgrounds are stacked, and the signal is overlaid to show the signal contamination in this region. The lower panel shows the ratio of data and predicted SM background with prefit uncertainties (including systematic and statistical uncertainties).

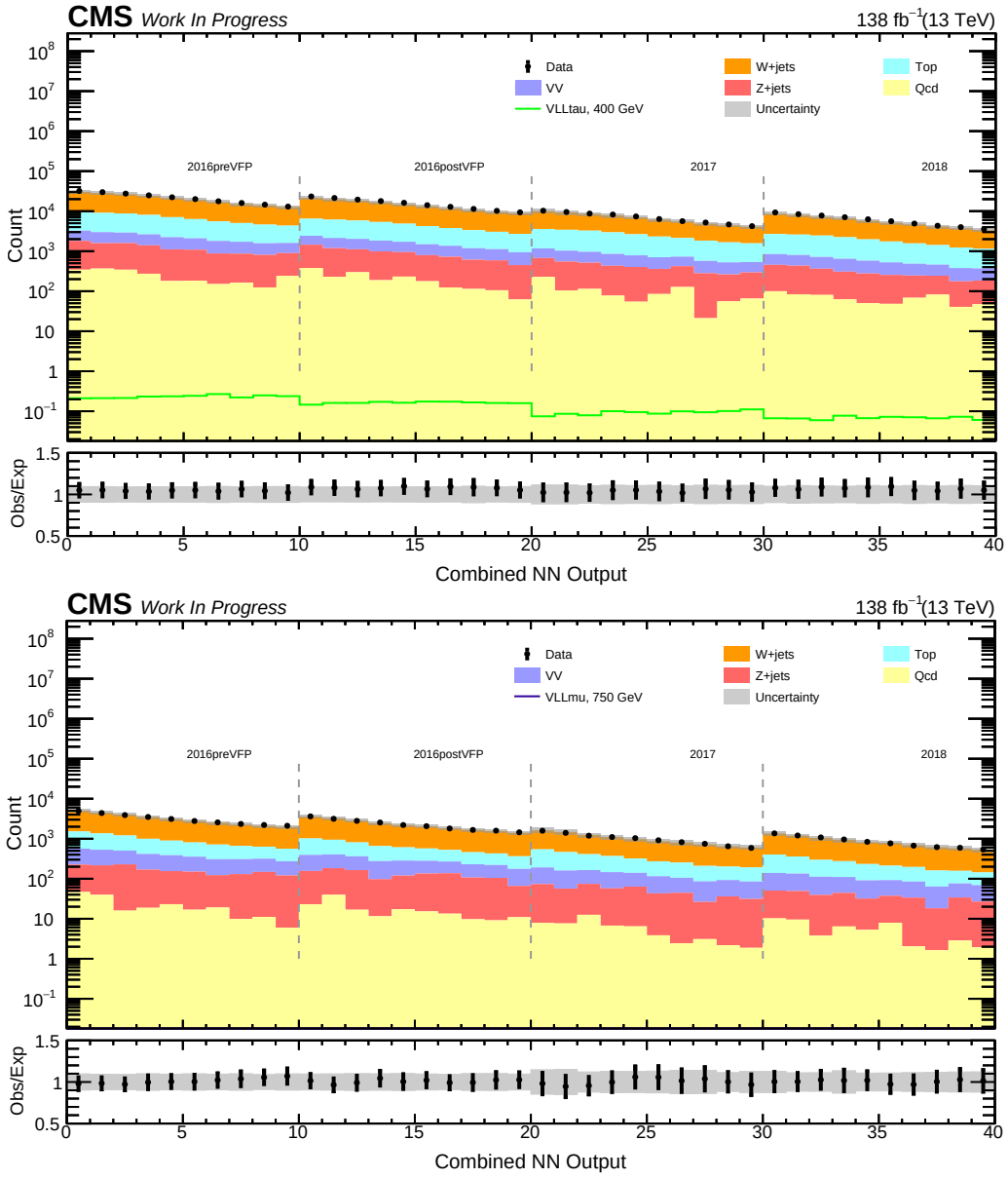


FIGURE 6.15: Distribution of the combine NN score with preselected SR events satisfying 0.5< Combine NN Score_2 < 0.7 cut for vector-like taus (upper) of 400 GeV and vector-like muons (lower) of 750 GeV mass. These plots are demonstrated with bins with an equal width of 0.02 from 0.5 to 0.07, resulting in 10 bins per data-taking period. This region in data is highly populated with background-like events, hence providing a good validation region to check data/mc agreement in NN input variables. Data is shown in black marker, backgrounds are stacked, and the signal is overlaid to show the signal contamination in this region. The lower panel shows the ratio of Data and predicted SM background with prefit uncertainties (including systematic and statistical uncertainties).

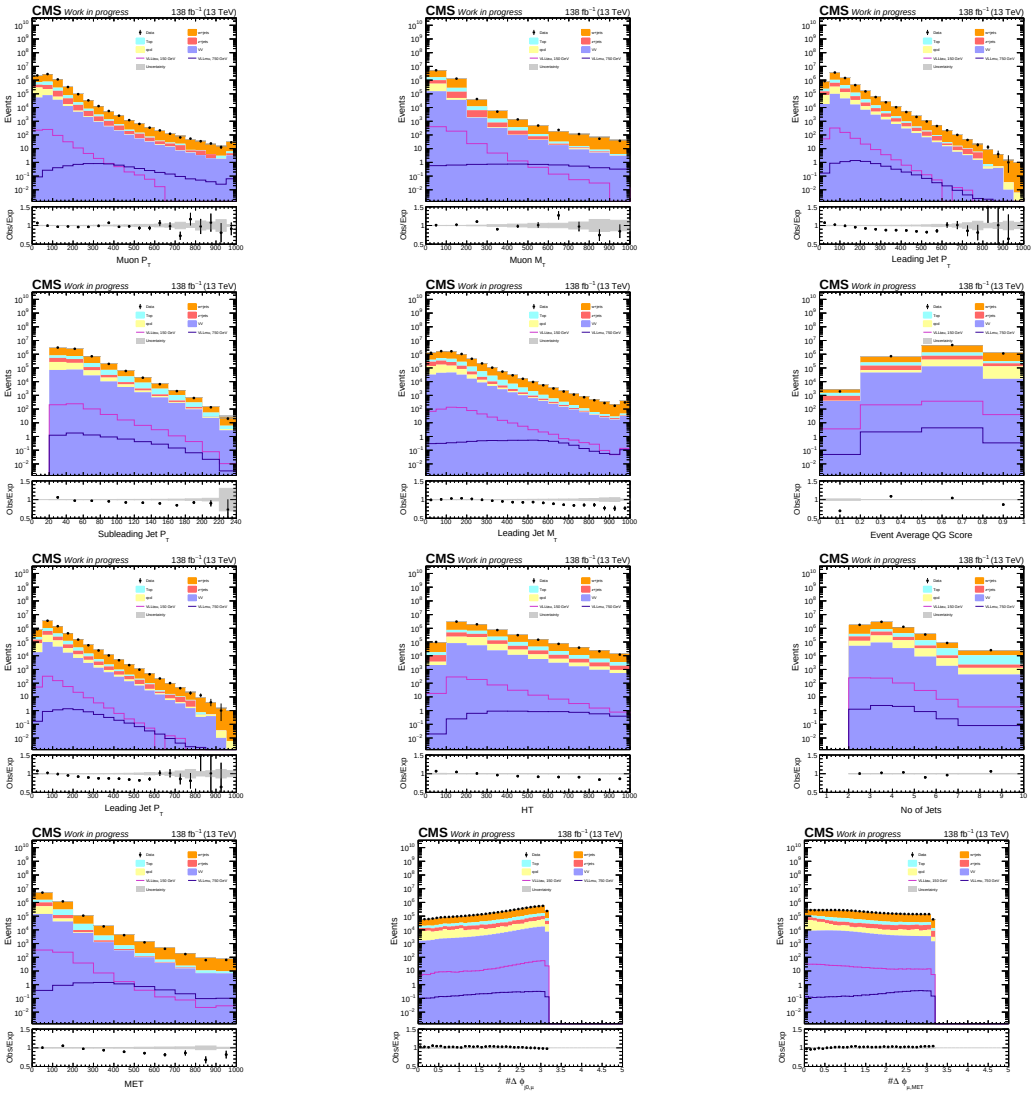


FIGURE 6.16: Distribution of the key training variables with preselected SR events for the combined 2016–2018 dataset. Data is shown in black markers, backgrounds are stacked, and two representative signal mass points are overlaid. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents statistical uncertainties only.

Chapter 7

Search for VLL in μjj final state: Results

7.1 Signal regions based on combine NN output

Each of the 20 distinct NN trainings provides a combined output score in the range of (0,1) for each event to be background- or signal-like. This combine neural network score is then treated as the primary discriminating variable to perform counting experiments for the vector-like lepton search. As illustrated in the score distribution of any of the networks, the region with the highest sensitivity to the signal is typically very close to the maximum combine score output. On the other hand, the region around the low end (here ≤ 0.7) of the combine 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 combine score spectrum are defined. More boundaries are defined on the high score side to achieve a good signal-to-noise ratio, maintaining a smooth and well-behaved expected background yield.

7.1.1 Binning strategy for final signal regions

Since this analysis probes a wide mass range for both vector-like models, the binning strategy is optimized for low and high-mass VLLs in vector-like tau or muon scenarios. Extremely fine binning is avoided to have a sufficient number of events at higher values of the NN spectrum, where most of the signal is populated. After optimization, we found that the following binning strategy yields a stable fitting strategy, avoiding statistical fluctuation when computing the limit.

- *Low mass binning*: 0.7, 0.75, 0.79, 0.82, 0.84, 0.85, 1.0
- *High mass binning*: 0.8, 0.818, 0.844, 0.868, 0.89, 0.91, 0.928, 0.944, 0.958, 0.97, 0.98, 0.988, 0.994, 0.998, 1.0

This is achieved by starting at the rightmost end of the combine score and then defining further bins towards the left with increasing step size equal to bin width 0.01 (0.002) such

that $(i-1)$ th bin is larger than i th bin by bin width$\times$ position of the bin from the rightmost end. The constraint was that each bin must have at least one background event. This variable binning strategy gives a smoothly increasing spectrum towards the left.

For low mass binning, the background yield at the high score was very low; hence, the bins are merged to have an appropriate background yield and then execute the binning strategy from 0.85 to 0.7 with a bin width of 0.01. A similar approach is taken for high mass binning, but since more backgrounds populate at high score, a bin width of 0.002 is chosen from 1.0 to 0.8. The network performance is comparatively better at high mass, so the combine network score threshold is selected as 0.8. The rest of the region can be used to validate the ML strategy.

Low mass binning is applied on Mass 100 and Mass 125 GeV combine output score spectrum in the vector-like taus scenario, and Mass 100 GeV combine output score in the vector-like muon model. The high mass binning strategy is applied in the vector-like taus case for Mass 150 GeV to 400 GeV combine NN score, and on Mass 150 GeV to 1000 GeV combine NN score in the case of the vector-like muon scenario.

7.2 Application of systematic uncertainties on the final discriminant

To assess the effect of systematic uncertainties, as covered in Section 5.6, on the final discriminant (combine NN output), the uncertainties were propagated from each of the sources on the final signal regions per year corresponding to every network (total 20 SRs as discussed before). The impact on the major backgrounds was analyzed, and the relative variation with respect to the nominal yield was taken as the final systematics band along with the statistical uncertainties, while computing the constraints on the VLL signal. Table 7.1 encapsulates the sources, magnitudes, impact (on the 20 SRs), and correlation model of systematic uncertainties considered in this analysis.

Figure 7.1 shows example variations of jet energy resolution, PDF and QCD scale uncertainties, and pile-up uncertainties on the W+jets combine NN score for the vector-like tau model with 125 GeV mass network for 2018. The nominal variation is set at unity, and the up variation (red) and down variation (blue) are shown with respect to the nominal. The gray band is the statistical uncertainty per bin.

Figure 7.2 shows example variations of jet energy resolution, PDF and QCD scale uncertainties, and btag heavy flavor (bc) correlated uncertainties on the combine NN score for VLL-tau 125 GeV signal with VLL-tau 125 GeV mass network for 2018. The nominal variation is set at unity, and the up variation (red) and down variation (blue) are shown with respect to the nominal. The gray band is the statistical uncertainty per bin.

Uncertainty source	**Magnitude**	**Type**	**Processes**	**Variation**	**Correlation**
Statistical	1-100%	per event	All MC samples	1-100%	No
Luminosity	0.2-1.3%	per event	All MC samples	1.2-2.5%	Mixed
Muon reco., ID and iso. efficiency	1-5%	per lepton	All MC samples	2-5%	No
Trigger efficiency	1-4%	per lepton	All MC samples	<3%	No
b tag efficiency	1-10%	per jet	All MC samples	2-5%	Mixed
Pileup	5%	per event	All MC samples	<3%	Yes
Jet energy scale	1-10%	per jet	All MC samples	<5%	No
Jet energy resolution	1-10%	per jet	All MC samples	<5%	No
Prefiring (2016 and 2017)	20%	per event	All MC samples	<2%	No
HT-based corrections	1-10%	per event	W+jets	<5%	No
Muon p_{T}-based corrections	1-20%	per event	W+jets	<5%	No
Muon M_T-based corrections	1-20%	per event	W+jets	5-10%	No
PDF, fact./renorm. scale	<20%	per event	All MC samples	<15%	Yes
W+jets normalization	<0.2%	per event	W+jets	<0.2%	No
QCD normalization	<2%	per event	QCD	<2%	No

TABLE 7.1: Sources, magnitudes, impact, and correlation model of systematic uncertainties in the signal region. Uncertainty sources marked as Yes under the correlation model have their nuisance parameters correlated across the data-taking era.

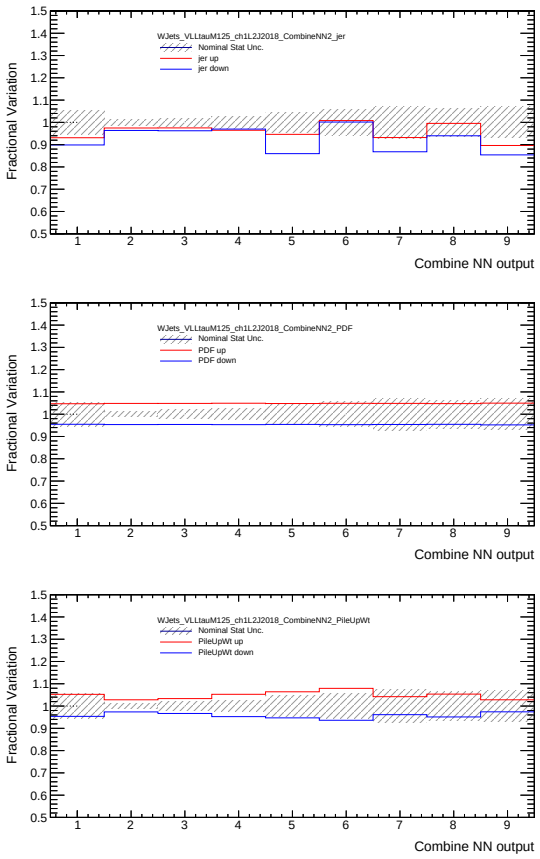


FIGURE 7.1: Example variations of jet energy resolution, PDF and QCD scale uncertainties, and pile-up uncertainties on the W+jets combine NN score for vector-like tau model with 125 GeV mass network for 2018. The nominal variation is set at unity, and the up variation (red) and down variation (blue) are shown with respect to the nominal. The gray band is the statistical uncertainty per bin.

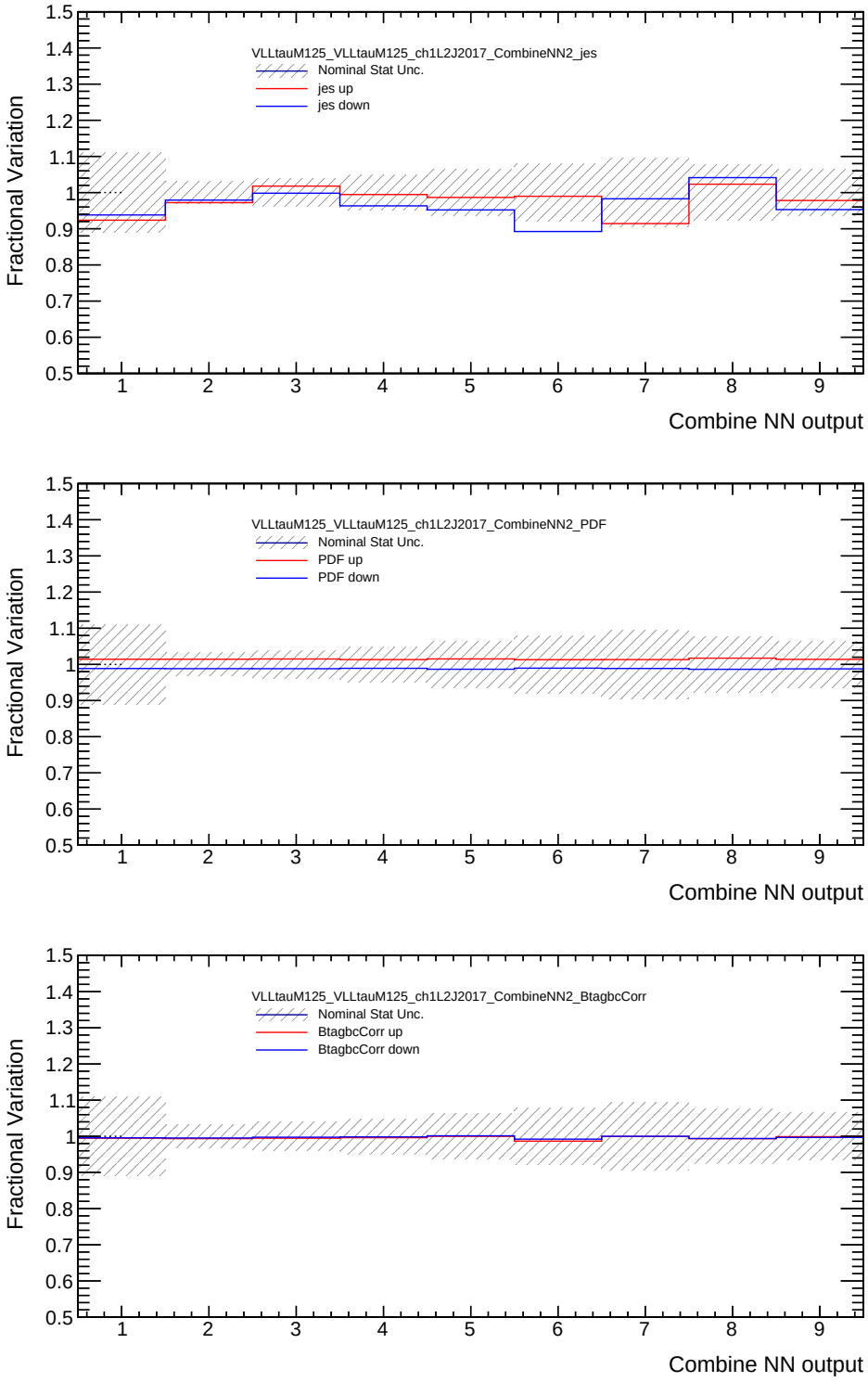


FIGURE 7.2: Example variations of jet energy resolution, PDF and QCD scale uncertainties, and btag heavy flavor (bc) correlated uncertainties on the combine NN score for VLL-tau 125 GeV signal with VLL-tau 125 GeV mass network for 2018. The nominal variation is set at unity, and the up variation (red) and down variation (blue) are shown with respect to the nominal. The gray band is the statistical uncertainty per bin.

Figure 7.3 shows example variations of HT (left), muon p_{T} (middle), and muon M_T (right) based correction uncertainties on the combine NN score for VLL-tau 125 GeV in 2016postVFP, VLL-tau 400 GeV in 2017, VLL-muon 150 GeV in 2017, VLL-muon 400 GeV in 2018, and VLL-muon 750 GeV in 2016preVFP, from top to bottom, respectively. The nominal variation is set at unity, and the up variation (red) and down variation (blue) are shown with respect to the nominal. The gray band is the statistical uncertainty per bin.

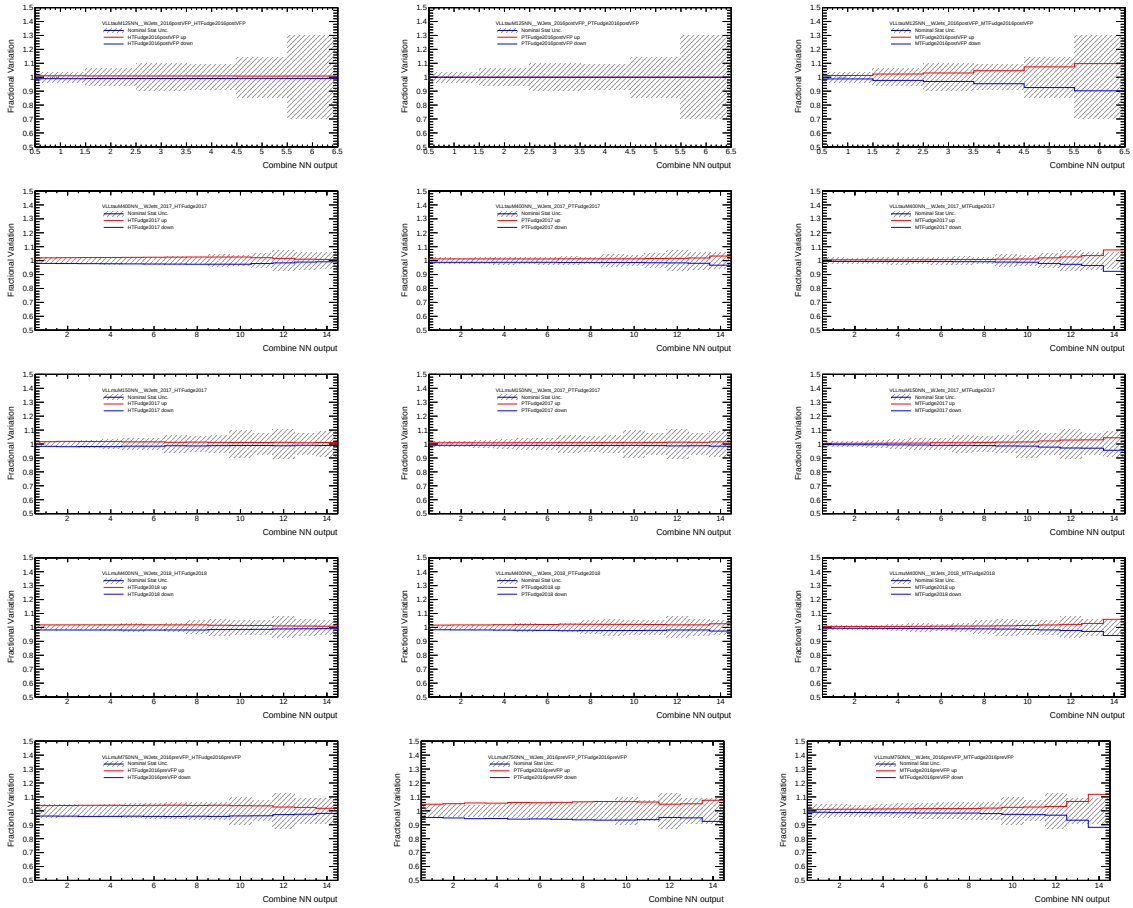


FIGURE 7.3: Example variations of HT (left), muon p_{T} (middle), and muon M_T (right) based correction uncertainties on the combine NN score for VLL-tau 125 GeV in 2016postVFP, VLL-tau 400 GeV in 2017, VLL-muon 150 GeV in 2017, VLL-muon 400 GeV in 2018, and VLL-muon 750 GeV in 2016preVFP, from top to down, respectively. The nominal variation is set at unity, and the up variation (red) and down variation (blue) are shown with respect to the nominal. The gray band is the statistical uncertainty per bin.

Figure 7.4 shows example variations of PDF (left), QCD scale (middle) and pile-up (right) uncertainties on the combine NN score on W+jets process in VLL-tau 125 GeV network in 2016postVFP, VLL-tau 400 GeV network in 2017, VLL-muon 150 GeV network in 2017, VLL-muon 400 GeV network in 2018, and VLL-muon 750 GeV network in 2016preVFP, from top to down, respectively. The nominal variation is set at unity, and the up

variation (red) and down variation (blue) are shown with respect to the nominal. The gray band is the statistical uncertainty per bin.

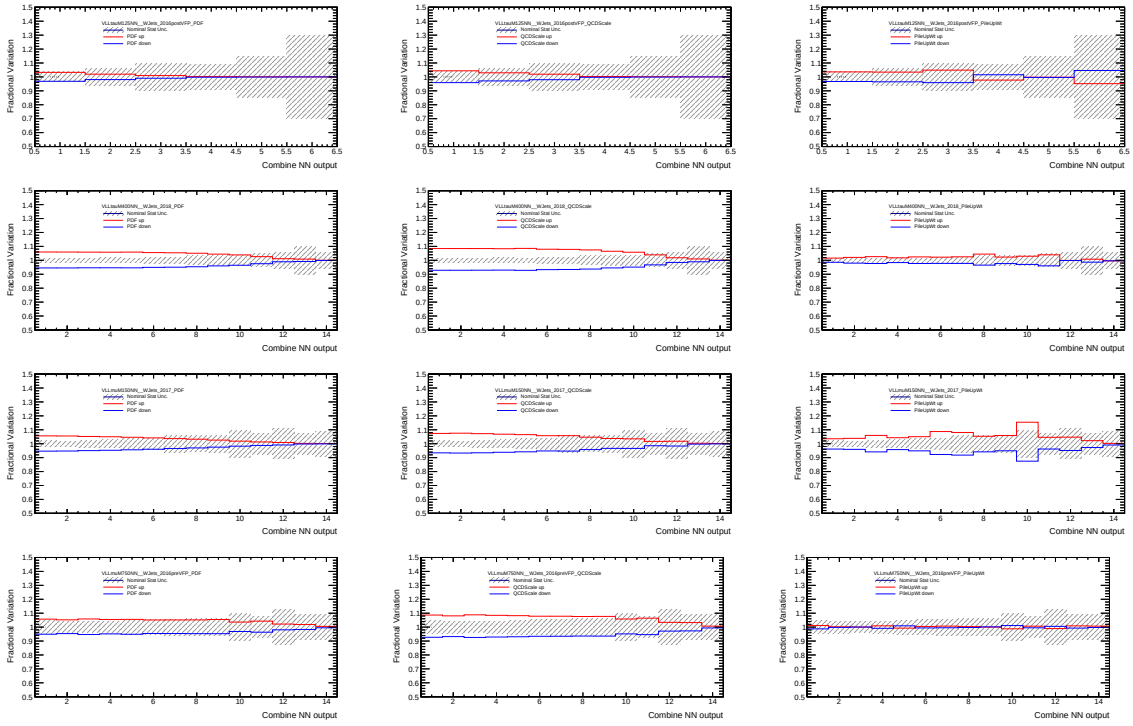


FIGURE 7.4: Example variations of PDF (left), QCD scale (middle), and pile-up (right) uncertainties on the combine NN score on W+jets process in VLL-tau 125 GeV network in 2016postVFP, VLL-tau 400 GeV network in 2017, VLL-muon 150 GeV network in 2017, VLL-muon 400 GeV network in 2018, and VLL-muon 750 GeV network in 2016preVFP, from top to down, respectively. The nominal variation is set at unity, and the up variation (red) and down variation (blue) are shown with respect to the nominal. The gray band is the statistical uncertainty per bin.

Figure 7.5 Example variations of PDF (left), QCD scale (middle), and pile-up (right) uncertainties on the combine NN score on the respective signal process in VLL-tau 125 GeV network in 2016postVFP, VLL-tau 400 GeV network in 2017, VLL-muon 150 GeV network in 2017, VLL-muon 400 GeV network in 2018, and VLL-muon 750 GeV network in 2016preVFP, from top to down, respectively. The nominal variation is set at unity, and the up variation (red) and down variation (blue) are shown with respect to the nominal. The gray band is the statistical uncertainty per bin.

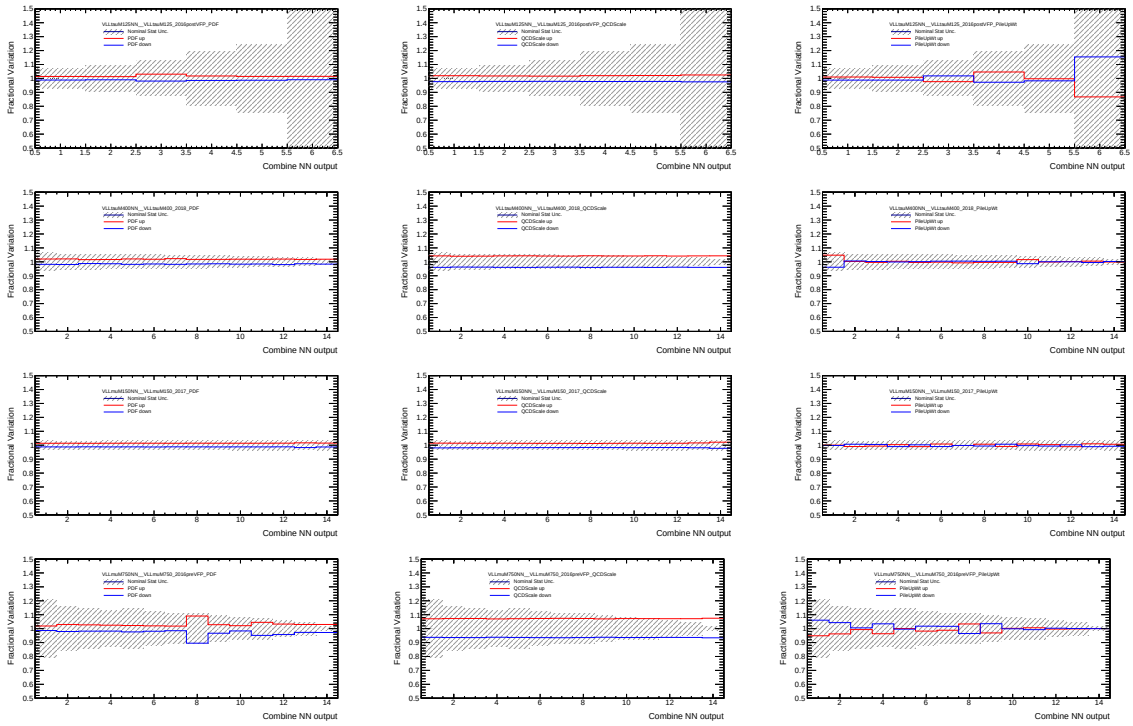


FIGURE 7.5: Example variations of PDF (left), QCD scale (middle) and pile-up (right) uncertainties on the combine NN score on respective signal process in VLLtau 125 GeV network in 2016postVFP, VLL-tau 400 GeV network in 2017, VLL-muon 150 GeV network in 2017, VLL-muon 400 GeV network in 2018, and VLL-muon 750 GeV network in 2016preVFP, from top to down, respectively. 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 Results

Figure 7.6 shows the combine NN score regions for 100, 125, and 150 GeV mass hypotheses in the vector-like taus model. Figure 7.7– 7.8 shows the combine NN score regions for 200–400 GeV mass hypotheses in the vector-like taus model.

Similarly, Figure 7.9– 7.12 shows the combine NN score regions for 100 GeV to 1000 GeV mass models in the vector-like muon scenario.

No significant deviations in data from the expected SM background prediction were observed in any bin of the combine score regions of each VLL mass hypothesis in the three eras of the data-taking. Consequently, the observed and expected upper limits at 95% CL were calculated on the production cross section of the doublet vector-like lepton model.

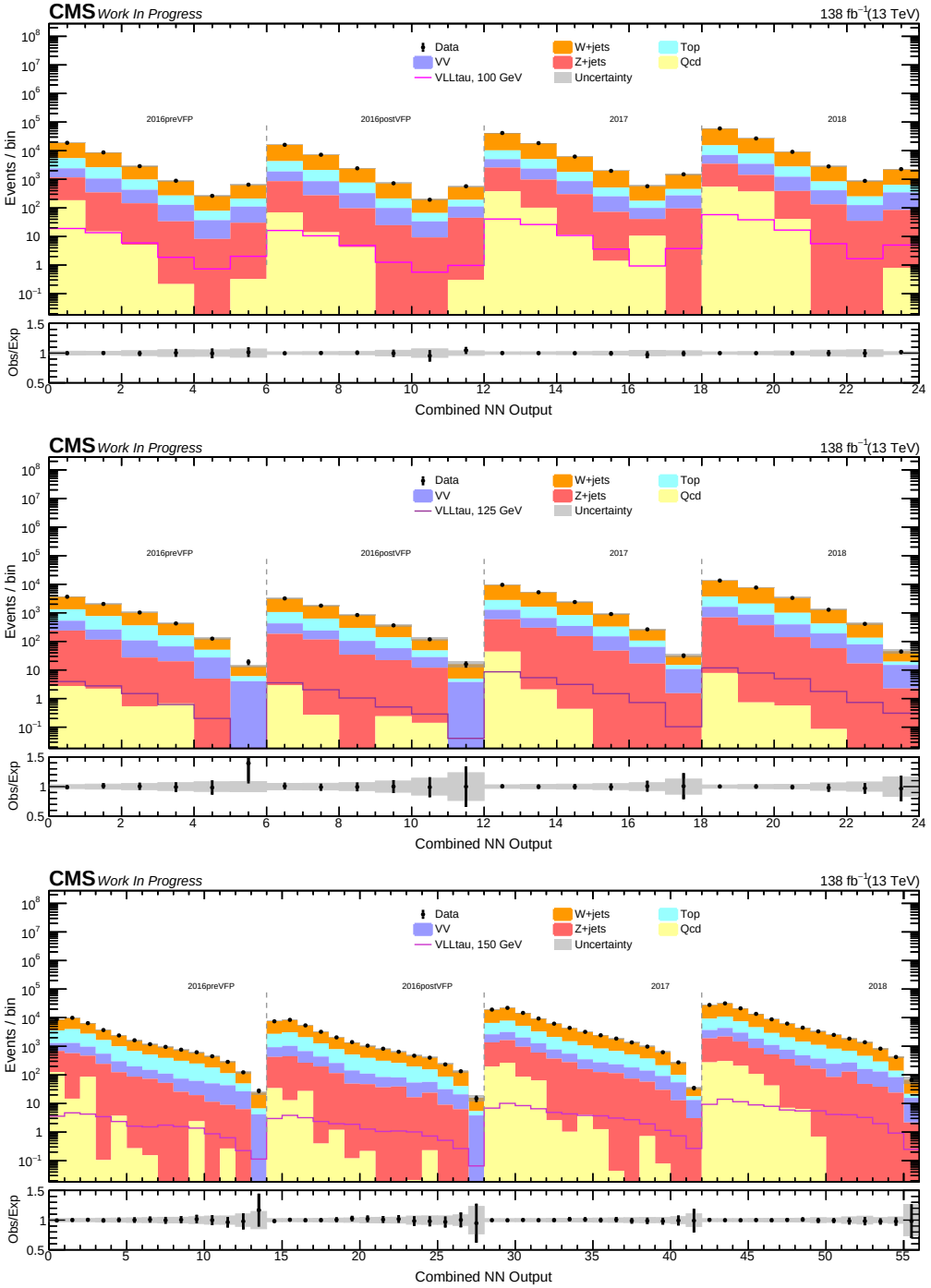


FIGURE 7.6: 100, 125 and 150 GeV signal region for the vector-like tau model in full Run-2 dataset. Individual eras are shown in each plot from left to right. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum 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, the signal with the respective mass hypothesis before the fit is also overlaid.

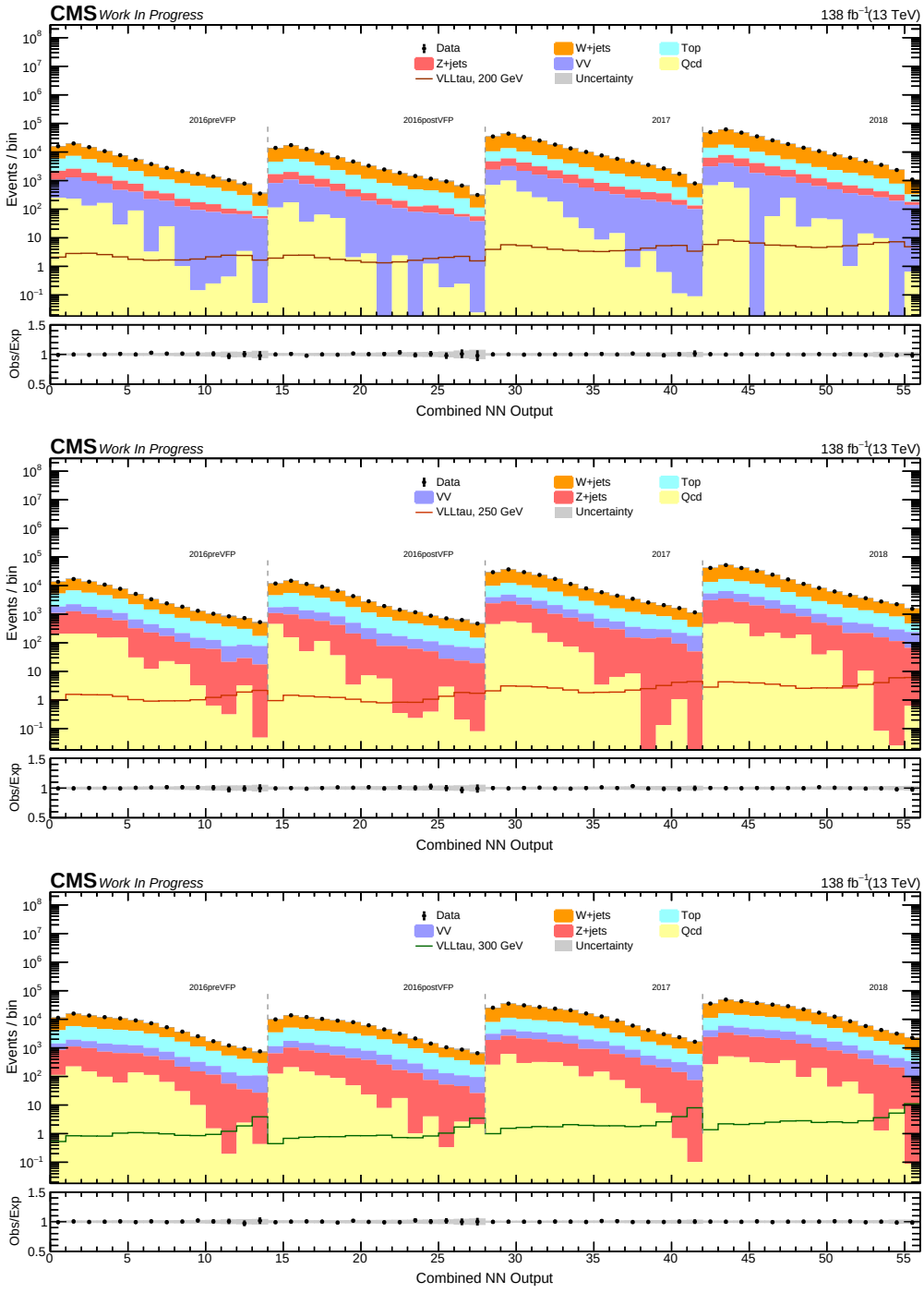


FIGURE 7.7: 200, 250 and 300 GeV signal region for the vector-like tau model in full Run-2 dataset. Individual eras are shown in each plot from left to right. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum 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, the signal with the respective mass hypothesis before the fit is also overlaid.

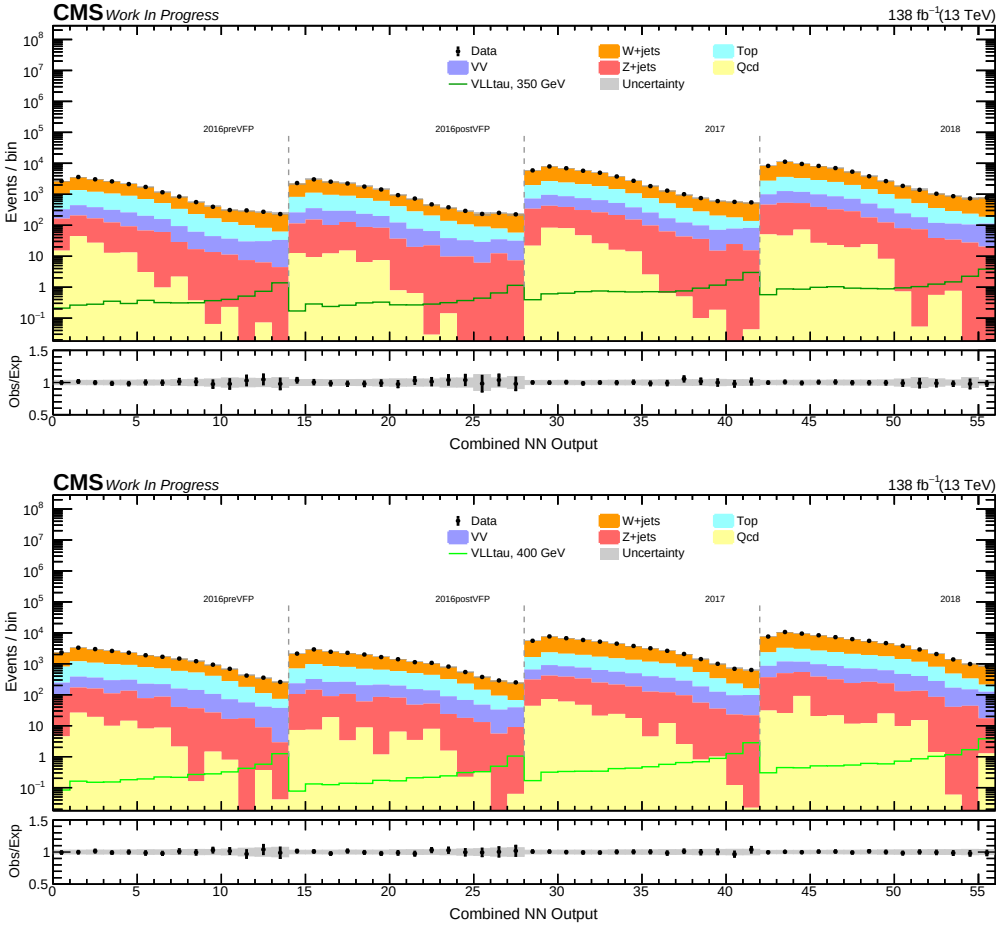


FIGURE 7.8: 350 and 400 GeV signal region for the vector-like tau model in full Run-2 dataset. Individual eras are shown in each plot from left to right. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical 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, the signal with the respective mass hypothesis before the fit is also overlaid.

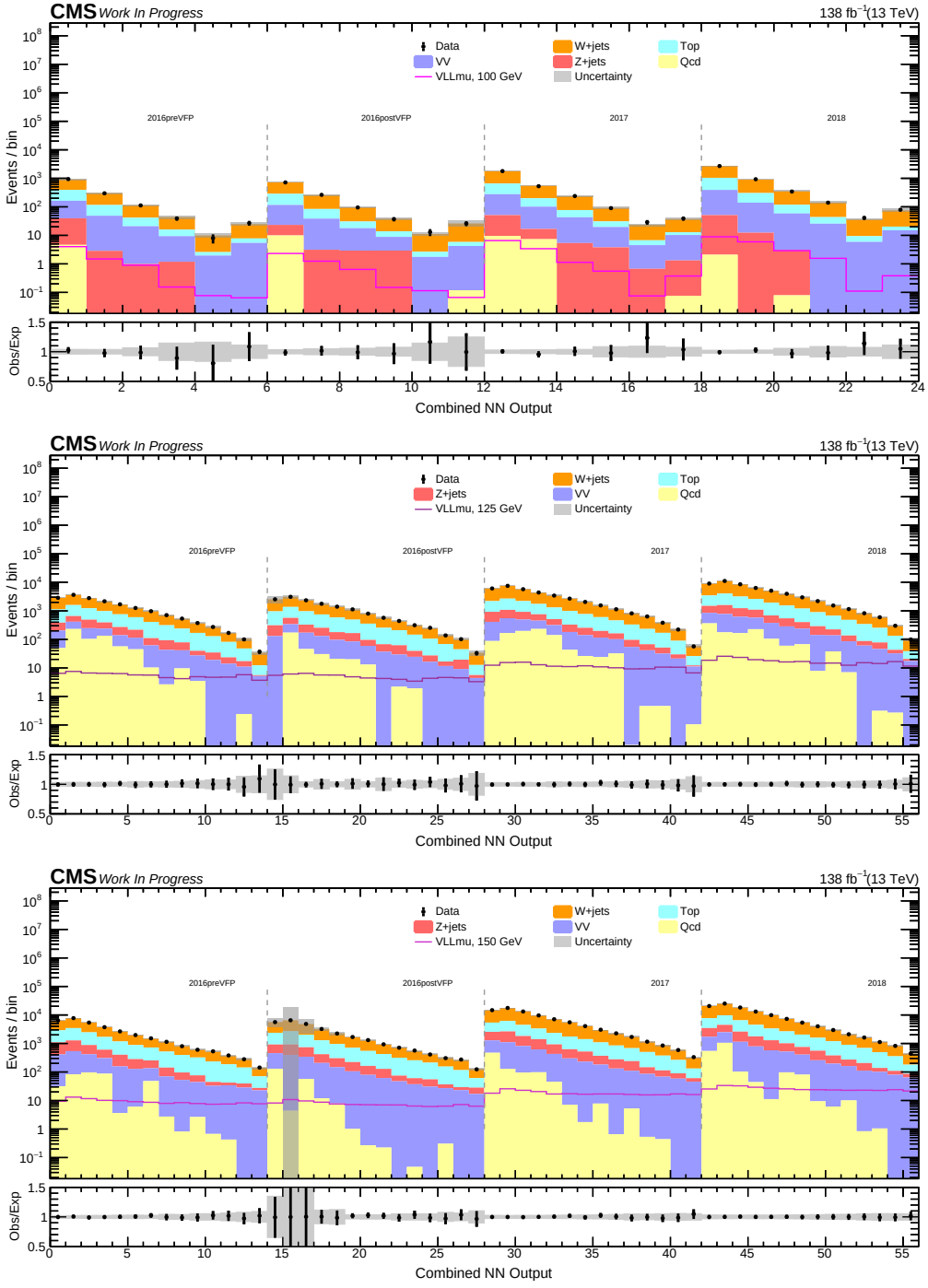


FIGURE 7.9: 100, 125 and 150 GeV signal region for the vector-like muon model in full Run-2 dataset. Individual eras are shown in each plot from left to right. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum 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, the signal with the respective mass hypothesis before the fit is also overlaid.

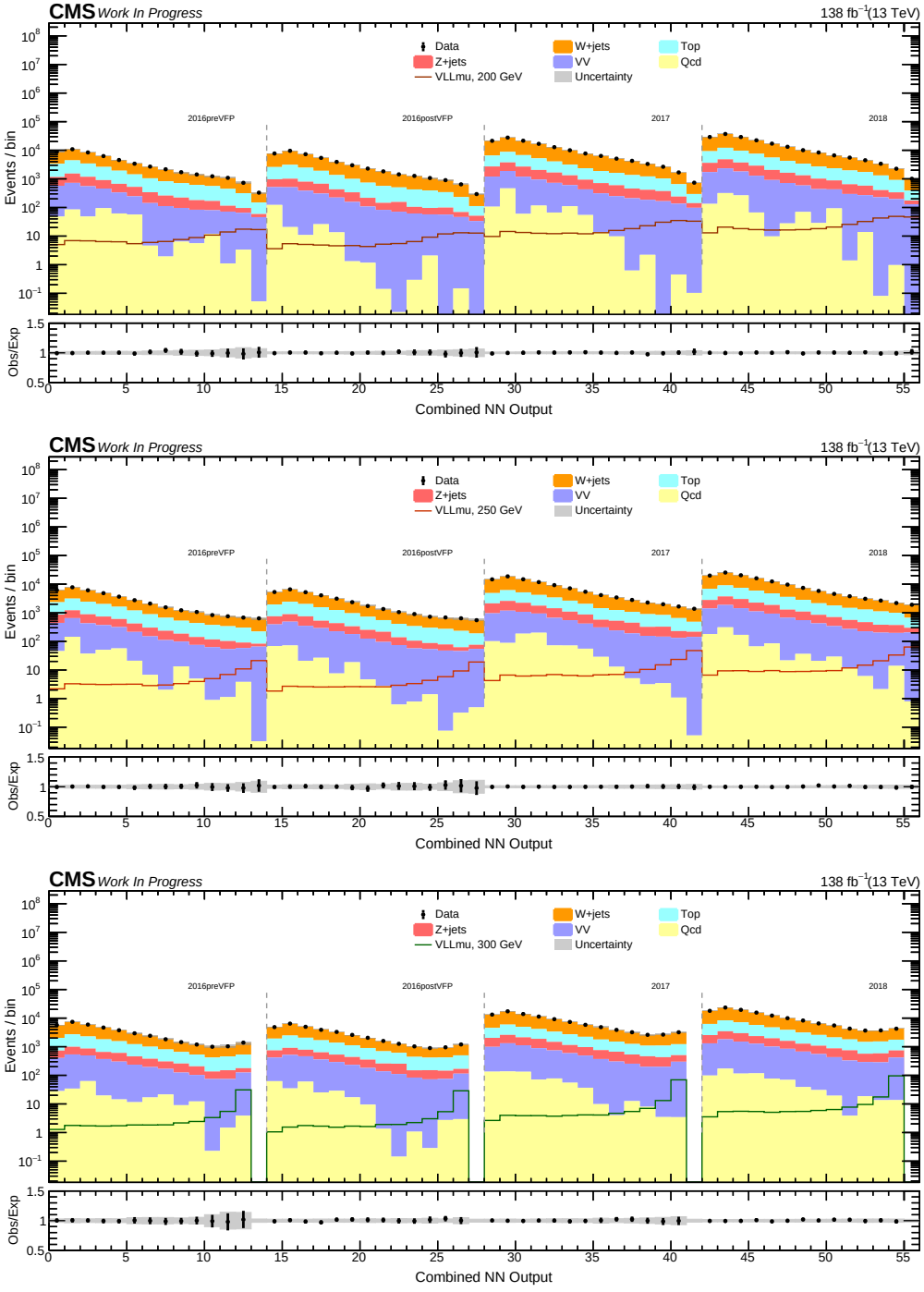


FIGURE 7.10: 200, 250 and 300 GeV signal region for the vector-like muon model in full Run-2 dataset. Individual eras are shown in each plot from left to right. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum 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, the signal with the respective mass hypothesis before the fit is also overlaid.

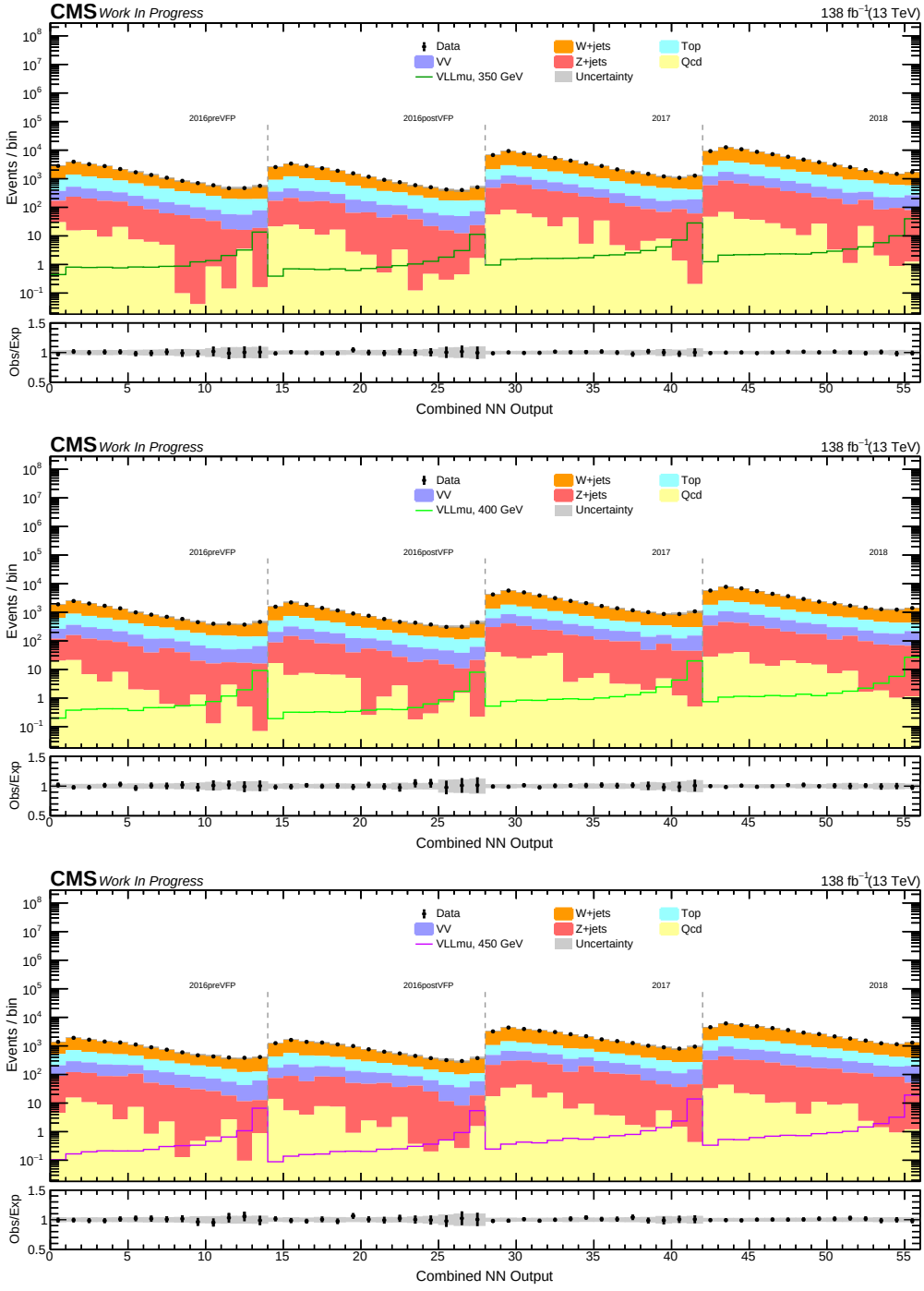


FIGURE 7.11: 350, 400 and 450 GeV signal region for the vector-like muon model in full Run-2 dataset. Individual eras are shown in each plot from left to right. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum 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, the signal with the respective mass hypothesis before the fit is also overlaid.

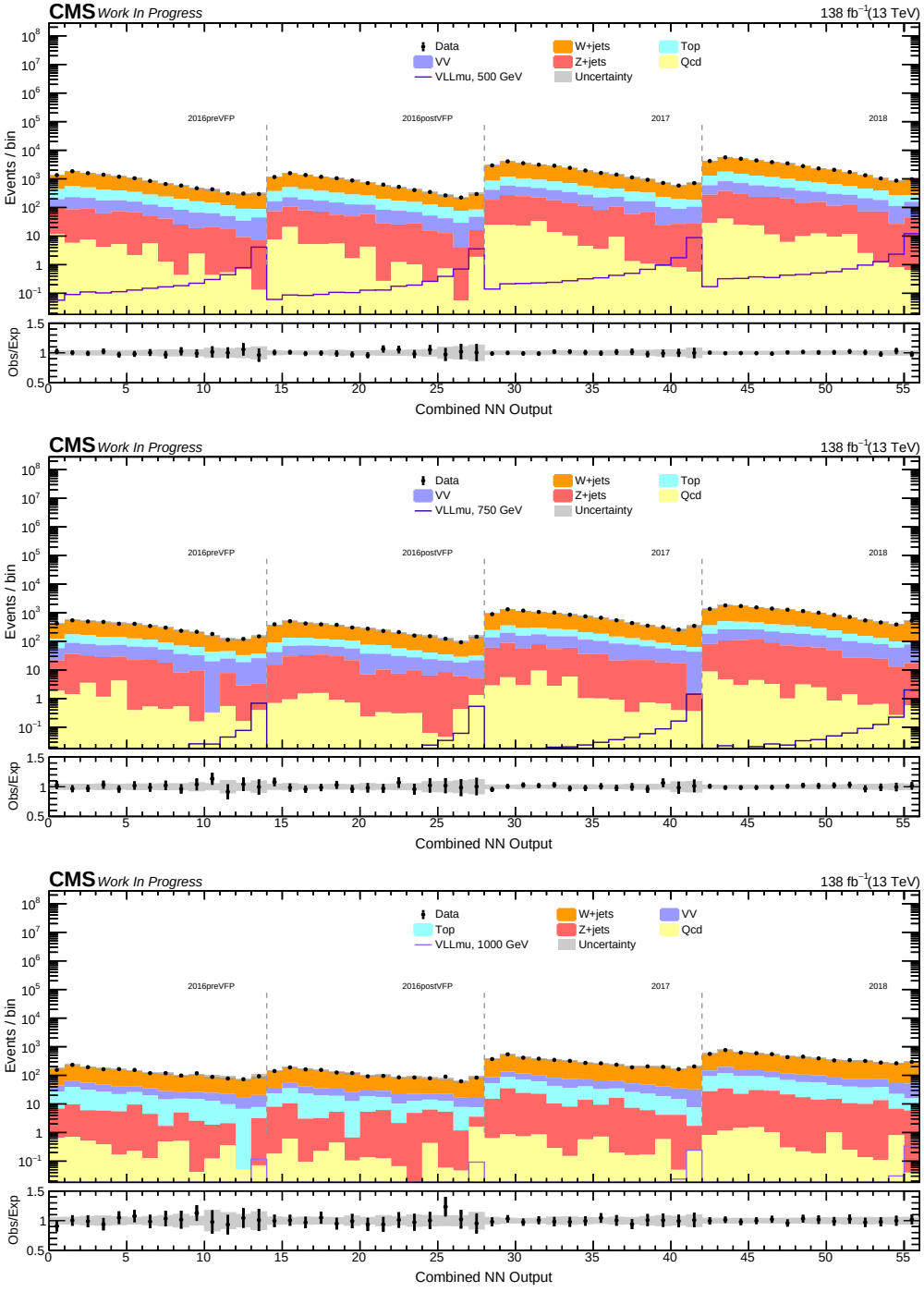


FIGURE 7.12: 500, 750 and 1000 GeV signal region for the vector-like muon model in full Run-2 dataset. Individual eras are shown in each plot from left to right. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum 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, the signal with the respective mass hypothesis before the fit is also overlaid.

Figure 7.13 illustrates the background composition in the final signal regions for all mass and model networks considered in this analysis.

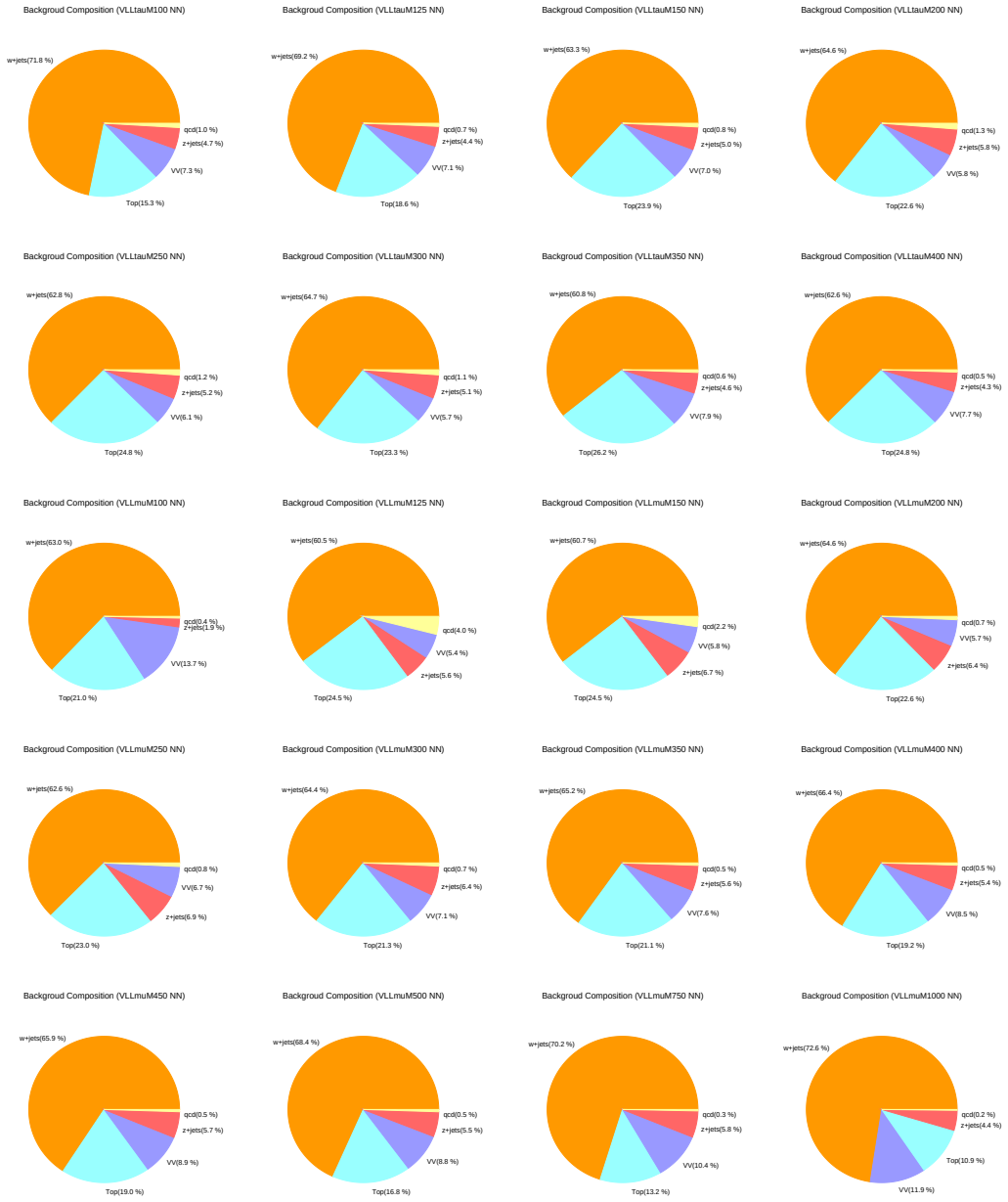


FIGURE 7.13: Background composition in the final neural network based signal regions for the combined 2016–2018 dataset in the singlet vector-like tau and muon models.

7.4 Calculating limits

The observation of events is consistent with the standard model predictions; this leads naturally to the question of what we can infer about the vector-like lepton models under the test using these results. However, we need to quantify the term "consistent" with the standard model and how to objectively constrain the parameters of the new physics model, such as VLLs.

The statistical inference performed in high energy physics can be the following [82, 83, 84]:

- Estimate and constrain the parameters of the model under the test
- Testing if the standard model hypothesis is favored or any alternative hypothesis (e.g, vector-like leptons) is favored given the observation
  - *Null hypothesis (H_0)*: This is the standard model for new physics searches.
  - *Alternative hypothesis (H_1)*: Presence of an exotic particle not predicted by the standard model but by an alternative theory (such as vector-like leptons).

  This null hypothesis (H_0) is also referred to as the background-only hypothesis, and the alternative hypothesis (H_1), in which additional new physics signal processes contribute, is referred to as the signal-plus-background hypothesis.

Firstly, we need to estimate the unknown parameters of the model. The likelihood function, which quantifies the probability of observing the data, is constructed given a model. The likelihood can be constructed for n measurements (events) that follow the Poisson distribution as;

$$\begin{aligned} L(n|\mu,\theta) &= Poisson(n|\mu s(\theta)+b(\theta)) \times \pi(\theta) \\ &= \frac{e^{-(\mu s(\theta)+b(\theta))}(\mu s(\theta)+b(\theta))^n}{n!} \times \pi(\theta) \end{aligned} \tag{7.1}$$

where μ is signal strength multiplier, s is the expected number of signal events, b is the expected background event, n is the number of observed events, and θ are the nuisance parameters that consider the systematic effects. Here μ is an unknown parameter. $\mu = 0$ defines the null hypothesis. $\pi(\theta)$ is a prior for nuisance parameters. For example, for a Gaussian prior, $\pi(0,\theta) = e^{-\frac{\theta^2}{2}}$. This formula can be extended for N sets of measurements (or independent bins). We use the maximum likelihood method (MLE) to estimate the unknown parameters of this model.

Profile likelihood: The likelihood function is profiled to remove the nuisance parameter dependency, with nuisance parameters treated as free fit parameters. Profiling the nuisance parameters refers to finding the values of these parameters that maximize the likelihood for each value of the μ. The profiled likelihood, $\mathcal{L}(\mu) = \mathcal{L}(\mu, \hat{\theta}_\mu) \equiv max_\theta(L)(\mu,\theta)$.

Test statistic: The next step is to define a test statistic from our data sample that discriminates between the two hypotheses H_0 and H_1. The profiled test statistic that is used for analysis following the Neyman-Pearson lemma,

$$t_\mu = -2\ln(\frac{\mathcal{L}(\mu, \hat{\theta}_\mu)}{\mathcal{L}(\hat{\mu}, \hat{\theta})}) \tag{7.2}$$

We define a *critical region* ($t_{\mu,c}$) in the data space, and if the measured value of the test statistic ($t_{\mu,obs}$) lies within the critical region, the hypothesis H_0 is rejected; if $t_{\mu,obs}$ lies in the acceptance region, we fail to reject H_0.The following quantities are useful to give quantitative information about a test:

- *significance level*,α: Probability to reject H_0 if H_0 is assumed to be true. This is called the size of the test. We choose a value of α before performing the experiment (looking at the data). This is also called a type I error. To claim a discovery, the α is taken to be a very small value, and for an exclusion it is taken as 0.05 (for 95% CL).

- *Misidentification probability*, β: Probability to reject H_1 if H_1 is assumed to be true. This is also called a type-II error. $1 - \beta$ is the power of the test.

- *p-value*: The higher the values of t_μ, the greater the incompatibility between the data and μ. To quantify the level of disagreement, we compute the p-value,

$$p_\mu = \int_{t_{\mu},obs}^{\infty} f(t_\mu|\mu)dt_\mu \tag{7.3}$$

  Assuming the null hypothesis is true, the p-value quantifies whether an experiment, if repeated many times, would obtain data as far away (or more) from the null hypothesis. The p-value is a measurement of the observed level of significance. It is a function of the data and therefore a random variable; it is not the same as the size of the test α, which is a predefined constant. The size does not depend on $t_{\mu,obs}$, the observed value of the test statistic, but on $t_{\mu,c}$, the cut defining the critical region.

In addition to a "single-value" estimate of μ, the result is usually expressed in a confidence interval (with a certain probability that may contain the true parameter value of μ), reflecting the statistical precision of the measurement. Frequentist confidence interval of μ can be obtained by performing a hypothesis test for each set of μ values. If data is observed in the critical region, reject the value μ. In terms of the p-value, the parameter value of μ is rejected if $p_\mu < \alpha$. The confidence interval at $\mathrm{CL} = 1 - \alpha$ consists of those values of μ that are not rejected. Thus, an upper limit on μ is the greatest value satisfying $p_\mu \geq \alpha$. Confidence level (CL) is given as 1 - α; thus, α = 0.05 corresponds to 95% CL. We use a modified

frequentist approach to exclude a hypothesized signal, where the p-value ratio of signal and background hypothesis to background-only hypothesis, $\mathrm{CL_s} = \mathrm{p_{s+b}}/(1 - \mathrm{p_b}) < \alpha$, is used.

The combined neural network score for each VLL mass training is used to compute the expected and observed limit for that respective mass hypothesis. To obtain upper limits on the signal cross section at 95% confidence level (CL), a modified frequentist approach is used with a test statistic based on the profile likelihood in the asymptotic approximation and the $\mathrm{CL_s}$ criterion. A linear interpolation of the expected event yields is used between the simulated signal samples in the limit calculations. Systematic uncertainties are incorporated into the likelihood as nuisance parameters with log-normal probability distributions. In contrast, statistical uncertainties are modeled with gamma functions (using auto MC stat in the Higgs combine tool) [85, 86, 87].

This analysis is comparatively more sensitive to the vector-like muons than the vector-like tau models. Figure 7.14 compares the shapes between the W+jets background and a few representative VLL-muon and VLL-tau signal mass points for a few key variables used in this search. At low mass (eg, 100 GeV), the VLLs branching fraction of $E \to W\nu_\ell$ is almost of 85%. The muon selected mostly comes from the gauge boson decay for both models in the low mass scenario, and their p_T spectrum looks similar. Other jet-related variables also look the same in low mass for both models. This is the reason for similar sensitivity at 100 GeV. But at high mass, other decay modes also start contributing. The possibility of the muon originating from the VLLs increases due to the $E \to Z\ell$ or $E \to H\ell$ decay modes for the vector-like muon model. However, for the vector-like tau model, the muon can either come from the gauge boson decay or the subsequent leptonic decays of the tau that come from the vector-like tau particle. It can be seen that the muon p_T spectrum for the 100 and 400 GeV VLL signal is quite different for vector-like muon models compared to the vector-like tau models for the reasons discussed above. Conversely, the leading jet p_T distributions are similar for both models, at any representative mass points, as they can mostly originate from the decays of gauge bosons from the VLLs. The leptonic activity was crucial to probe the parameter space of these two models.

Fig. 7.15 shows the expected and observed limit for the third-generation vector-like lepton model (left) and the second-generation vector-like lepton model (right) in a singlet scenario for the full Run-2 dataset. No third-generation vector-like lepton (vector-like taus) and second-generation vector-like lepton (vector-like muons) in the singlet model are excluded in this analysis. It is quite interesting that even if we exclude a large BSM parameter space at the TeV scale, such simple VLL models (which appear in SUSY, GUT, etc) can evade detection primarily due to overlapping signatures with the dominant SM processes (control regions in usual searches), or the analysis choices that are tuned to probe massive particles.

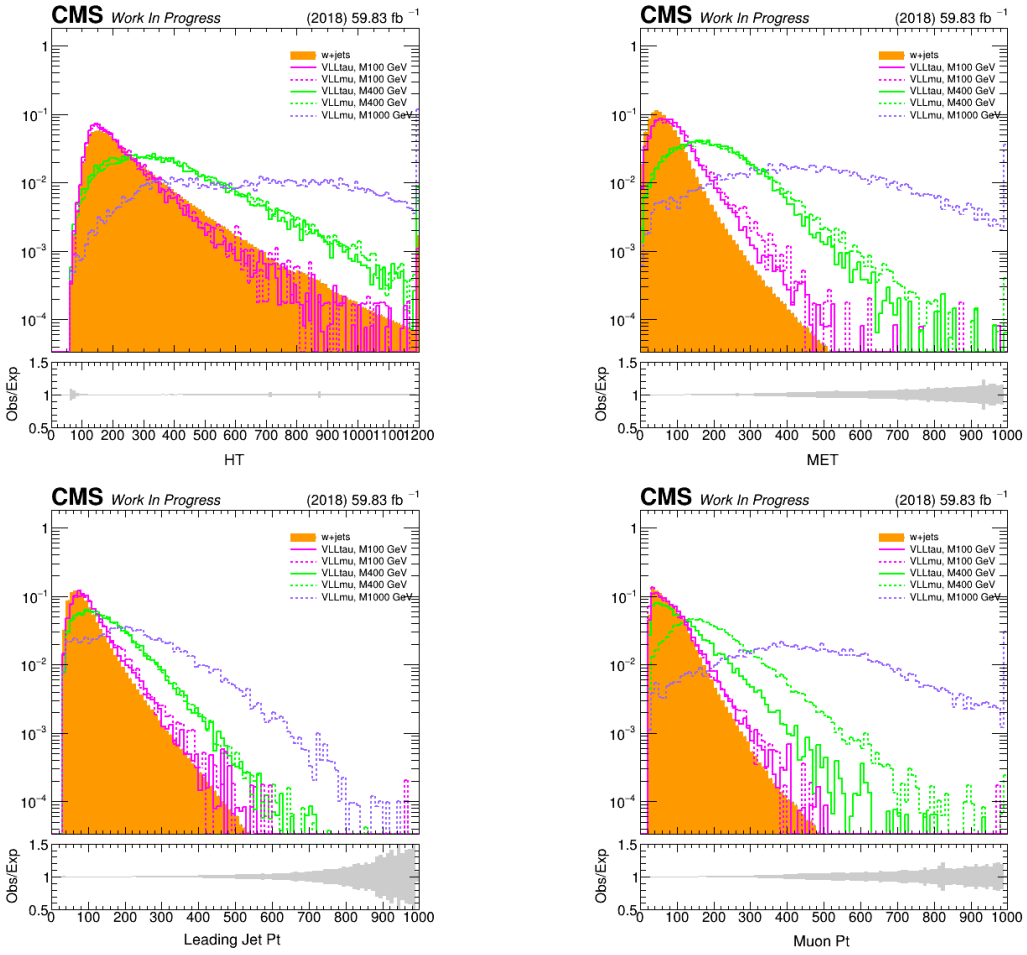


FIGURE 7.14: Shape difference between the most dominating backgrounds, W+jets and a few representative signal mass points for H_T, p_T^{miss}, Leading jet p_T, and muon p_T for the 2018 samples.

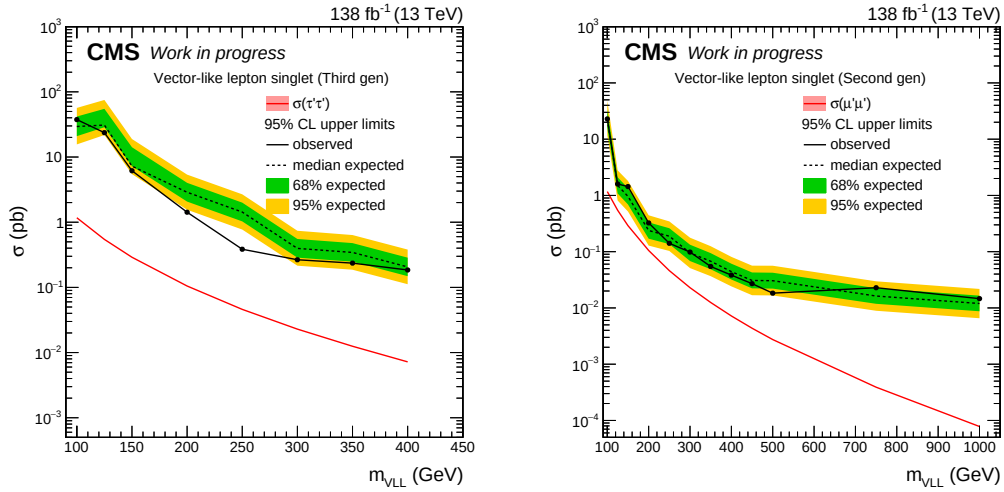


FIGURE 7.15: Observed and expected upper limits at 95% confidence level on the production cross section for the third-generation vector-like leptons (left) and second-generation vector-like leptons (right) in the singlet model with the full Run-2 dataset. The theoretical prediction for the production cross-section of a vector-like lepton singlet coupling to the third and second-generation SM leptons is also shown.

While this search did not have enough sensitivity to exclude phase space for the considered models, it exercised the power of good analysis to improve the signal-to-background significantly.

The dip in the observed cross-section upper limit at 250 GeV mass-point in the vector-like tau scenario is due to data fluctuating downwards in the most signal significant bins of the combine NN score spectrum corresponding to the VLL-tau 250 GeV network. Figure 7.16 shows the combine score distribution of VLL-tau 250 and 400 GeV networks for the full Run-2 dataset. Note that no such dip exists at the 400 GeV mass point. The trained neural network sculpted a different phase space of the background while separating the 250 GeV signal and 400 GeV signal from the same background distribution. We had a separate classifier for each model and mass point. The NN output shape of the background is different for 250 GeV and 400 GeV mass points. Also, underlying discriminating variables (from the rank of variables) are different for 250 GeV training vs 400 GeV training. This behavior is expected as the decay topology differs significantly, and a high mass of VLL gives rise to different lepton kinematics (and jet features). This was the reason for conducting separate training per mass to achieve optimal classifier performance. For example, the probability that jets originating from a quark or gluon is an important feature for 200/250 GeV training, while high muon M_T drives the 400 GeV mass point performance.

The expected limit for 125 GeV VLL-tau is worse than the adjacent 150 GeV mass point for two reasons. The preliminary SR selection criteria were based on a 150 GeV mass point. So, some acceptance is lost for lower mass points and due to the different binning transition

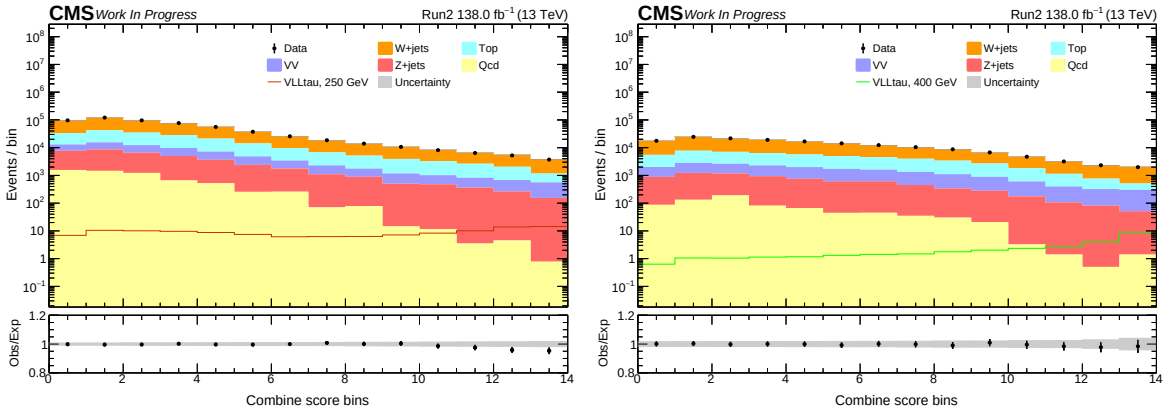


FIGURE 7.16: Combine score distribution of VLL-tau 250 and 400 GeV networks for the full Run-2 dataset. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, the signal with the respective mass hypothesis before the fit is also overlaid.

from low mass to high mass binning as described in Section 7.1.1. For 125 GeV mass points, the background yield at the high combine NN score was very low; hence, a few bins were merged to have an appropriate background yield.

Chapter 8

Prospects of vector-like leptons at HL-LHC

The CMS detector will undergo a significant upgrade, known as Phase-2, in preparation for the High-Luminosity LHC (HL-LHC) era beginning in 2029 [53, 88]. The determination of Higgs boson properties, and their connection to EW symmetry breaking (EWSB), is the primary target of the HL-LHC physics programme. This includes precision measurements of Higgs boson couplings, self-interactions, the stability of the Higgs potential, and rare decay modes to test the standard model and probe possible deviations. High statistics datasets and increased detector acceptance in HL-LHC should increase sensitivity to new phenomena such as supersymmetry, dark matter candidates, extra dimensions, and heavy resonances, and enable to explore higher mass scales and lower cross-sections inaccessible to the current LHC dataset to cover a wider BSM phase-space. Rare processes (e.g., multi-boson production, flavor anomalies) can be investigated with unprecedented statistical precision to reveal indirect signs of new physics. The SM processes like vector boson scattering (VBS), triple and quartic gauge couplings, and diboson production rely on forward jets and low-rate final states. High-granularity detectors improve jet resolution and forward object identification that would be crucial to isolate these processes in high-pileup. The upgraded inner tracking systems will allow better impact parameter resolution and secondary vertexing, enhancing the tagging performance of heavy-flavor jets, enabling stringent checks on the lepton flavor universality. A comprehensive discussion on the physics capabilities at HL-LHC is presented in this Ref. [89]- [90] and the references therein.

This upgrade aims to handle higher data rates and pileup conditions. The Level-1 (L1) hardware trigger will be enhanced to process events at an output rate of 750 kHz with a latency of 12.5 μs, while the High-Level Trigger (HLT) will further reduce the event rate to 7.5 kHz, approximately a factor of 100 reduction [91, 92]. The computing model of event processing and software management under this unprecedented data-intensive scenario is discussed in this Ref [93].

The tracking system, including both the pixel and strip detectors, will be completely

replaced to offer finer granularity, radiation tolerant, and reduced material budget, with an extended pseudorapidity coverage reaching $|\eta| < 4$ [52, 94]. The new strip modules with closely spaced sensors will enable a novel track-trigger capability. This expanded coverage and upgraded tracking will enhance jet and b-jet identification [95], especially in forward regions.

The endcap calorimeters will be replaced by a highly granular sampling calorimeter (HGCal) with improved spatial resolution in longitudinal and transverse directions [96]. New front-end electronics in the ECAL barrel enable precision timing for photons, and the HCAL barrel region read out by silicon photomultipliers (SiPM) [97] improves detector performance by enhancing signal-to-noise ratio, increasing photon detection efficiency, and enabling increased longitudinal segmentation, leading to better background rejection and energy resolution. The muon system will be improved through upgraded electronics for RPCs, CSCs, and DTs, along with the addition of new detectors using advanced RPC and GEM technologies [98]. These additions will provide geometrical coverage up to $|\eta| < 2.8$, and better performance and trigger efficiency in the forward region. This is particularly important for reconstructing physics objects at high η, improved muon shower reconstructions benefiting searches like vector-like leptons in some model-specific cases [95].

Moreover, a timing detector (MTD) for minimum ionizing particles (MIPs) will be installed in both the barrel and endcap regions [99], providing precise timing information with a resolution of around 30–40 picoseconds. This will enable 4D vertex reconstruction, which is essential for distinguishing tracks originating from different collision events and hard-interaction vertices among the nearly 200 pileup interactions per bunch crossing at the HL-LHC. This is important to maintain pile-up mitigation performance, especially as complex objects become harder to reconstruct under high pileup. Finally, heavy fermion searches will benefit from these detector enhancements, increased center-of-mass energy (14 TeV), and unprecedented integrated luminosity (3000 fb^{-1}) of the HL-LHC.

The physics capabilities of the Phase-2 upgrade of CMS for the HL-LHC have been studied in this chapter by projecting the sensitivity for vector-like electrons, muons, and taus in singlet and doublet scenarios, assuming a final integrated luminosity of 3000 fb^{-1} at $\sqrt{s}$ =14 TeV. Model-independent signal regions of a published CMS analysis [23] that performed a search for new BSM phenomena in multilepton final states may be utilized to extrapolate sensitivities for these three generations of vector-like lepton models (both singlet and doublet scenarios) at the HL-LHC scenario [100]. Before going into the detailed projection strategy, let us look at the summary of the multilepton search and model-independent signal region selections.

8.1 Brief overview of Run-2 multilepton results

The multilepton search analyzed events in seven orthogonal channels based on the number of light-charged leptons (electrons or muons) and hadronically decaying tau leptons (τ_{h}), defined as:

- at least four light leptons and any number of τ_{h} candidates (4ℓ)
- exactly three light leptons and at least one τ_{h} candidates ($3\ell1\tau_{\mathrm{h}}$)
- exactly three light leptons and no τ_{h} candidates (3ℓ)
- exactly two light leptons and at least two τ_{h} candidates ($2\ell2\tau_{\mathrm{h}}$)
- exactly two light leptons and exactly one τ_{h} candidates ($2\ell1\tau_{\mathrm{h}}$)
- exactly one light lepton and at least three τ_{h} candidates ($1\ell3\tau_{\mathrm{h}}$)
- exactly one light lepton and exactly two τ_{h} candidates ($1\ell2\tau_{\mathrm{h}}$)

Light leptons are denoted by ℓ, and hadronic tau candidates are denoted by τ_{h}. In the 4ℓ channel, only the leading four light leptons in p_{T} are used in the subsequent analysis. Likewise, in the $3\ell1\tau_{\mathrm{h}}$, $2\ell2\tau_{\mathrm{h}}$, and $1\ell3\tau_{\mathrm{h}}$ channels, only the leading one, two, and three τ_{h} are used, respectively.

Selected events in the seven channels are further categorized in a model-independent way, based on the characteristics of the SM backgrounds. The model-independent search regions are defined by splitting the channels into various regions based on the charge, flavor, invariant mass of lepton pairs, and kinematic properties of leptons, jets, and $p_{\mathrm{T}}^{\mathrm{miss}}$, as well as multiplicity of b-tagged jets. In this analysis, the DeepCSV b-tagging algorithm is used to determine the number of b-tagged jets in an event with a b-tagging efficiency of 70% at a 1% mistagging rate of light jets being misidentified as b-jets. The observable L_{T} is defined 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 four light leptons leading in p_{T}, while for the $3\ell1\tau_{\mathrm{h}}$ channel, it is calculated from the three light leptons and the leading τ_{h} candidate. The observable HT is defined as the scalar p_{T} sum of all selected jets in the event.

Events categorization based on the number of reconstructed Z boson candidates (with $|M_{\ell\ell} - M_Z| < 15$ GeV), the transverse mass of the light lepton (M_T^{ℓ}), and the light lepton p_{T} are used for this projection study. This categorization scheme yields 43 orthogonal event categories as shown in Figure 8.1. The L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ variable with optimized bin boundaries (200 GeV binning) constitutes the final signal region (SR) bins used to conduct the counting experiments in each category. A total of 156 L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ bins in the seven channels with events categorized by the earlier scheme are used to project the sensitivity in the HL-LHC scenario.

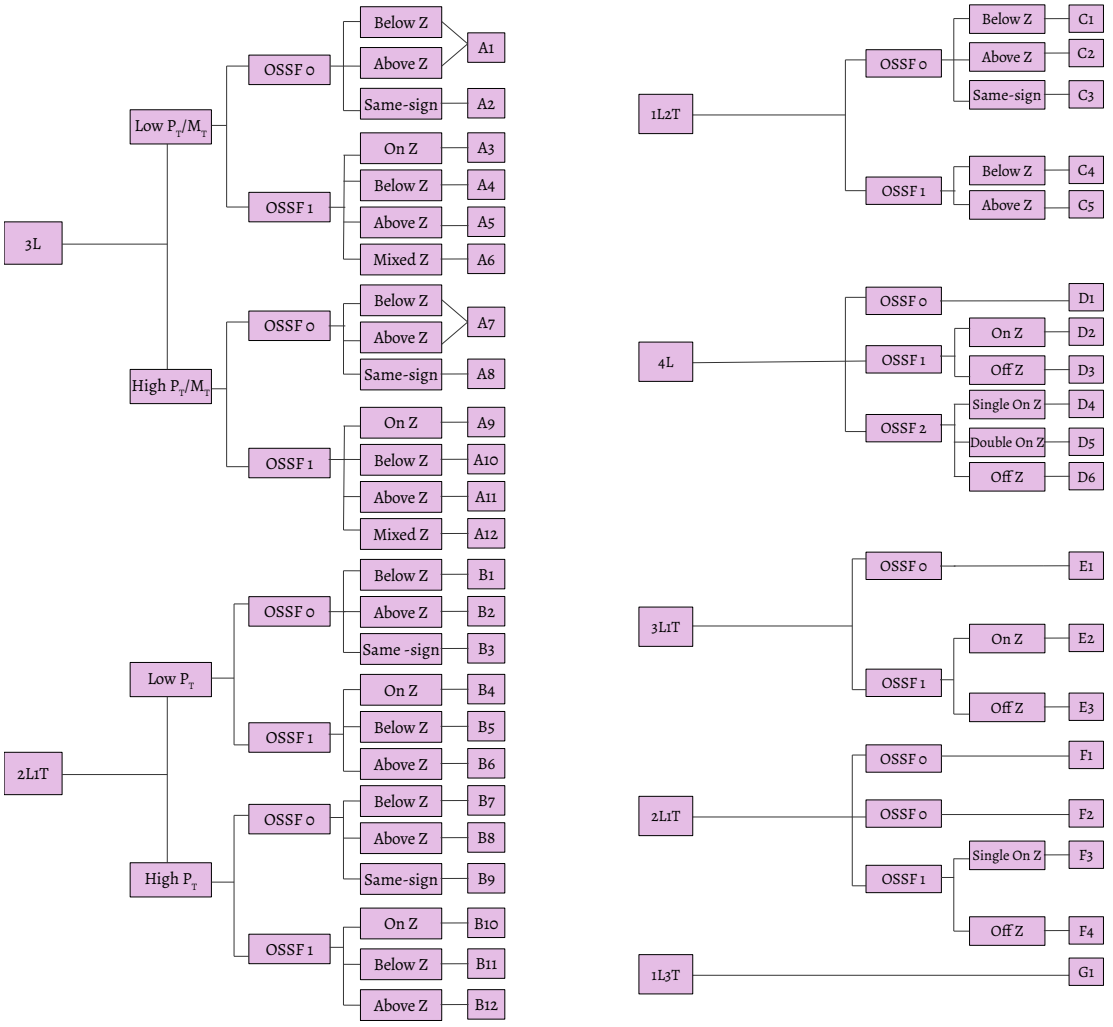


FIGURE 8.1: Event categorization as a function of lepton charge combinations and mass variables. The mass categorizations refer to masses of OSSF pairs if present, and of OSOF pairs otherwise.

The SM background processes, such as WZ, ZZ, ttZ, and ttW production, in which three or more reconstructed charged leptons originate from decays of SM bosons, contribute mainly to the irreducible background in various channels of this search. A smaller background contribution arises from ISR or FSR photons that convert asymmetrically, such that only one of the produced electrons is reconstructed in the detector, or from misidentifying on-shell photons as electrons. The dominant source of such backgrounds, collectively referred to as the conversion background, is DY events with an additional photon (Zγ). These backgrounds are estimated using simulation and normalized to observed data in the dedicated control regions. Another essential background component is the misidentified lepton background due to jets being misidentified as leptons, which is estimated using control samples in data via the matrix method.

No significant deviation from the SM background expectation was observed in these signal regions. Figure 8.2 shows the L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ bins in all the event categories per channel for the background-only hypothesis. The expected yield and uncertainty of individual backgrounds are used to compute the projected yields for the HL-LHC scenario.

8.2 Projection strategy

The number of events from a process satisfying a selection criteria in any final state for the entire Run-2 dataset can be written by the following formula,

$$N_{Run2} = \mathcal{L}_{Run2} \times \sigma_{13\ TeV} \times \mathcal{B} \times (\mathcal{A} \times \epsilon)_{Run2} \tag{8.1}$$

Where $\mathcal{L}$ is the total luminosity of collected data, σ is the theoretical production cross-section of the process, $\mathcal{B}$ is the branching ratio to any decay modes, $\mathcal{A}$ is the geometric acceptance of the detector and kinematic acceptance of the event selections, ϵ is the efficiency of some selection criteria. For Run-2, the cross-section corresponds to $\sqrt{s}$ =13 TeV p-p collision for a process, and $\mathcal{L}$ is 138 fb^{-1} of data. The branching fraction to different decay modes doesn't depend on the experimental scenarios. Given this key equation, we can write the projected yield for the HL-LHC case as follows:

$$N_{HL-LHC} = \mathcal{L}_{HL-LHC} \times \sigma_{14\ TeV} \times \mathcal{B} \times (\mathcal{A} \times \epsilon)_{HL-LHC} \tag{8.2}$$

The projected luminosity of the data to be taken at HL-LHC is 3000 fb^{-1} . We assume that the kinematic cuts, the efficiency of different reconstruction and identification algorithms of the physics objects used in HL-LHC analysis, should yield similar $\mathcal{A} \times \epsilon$ of a particular process under consideration. Indeed, the assumption is quite conservative since the

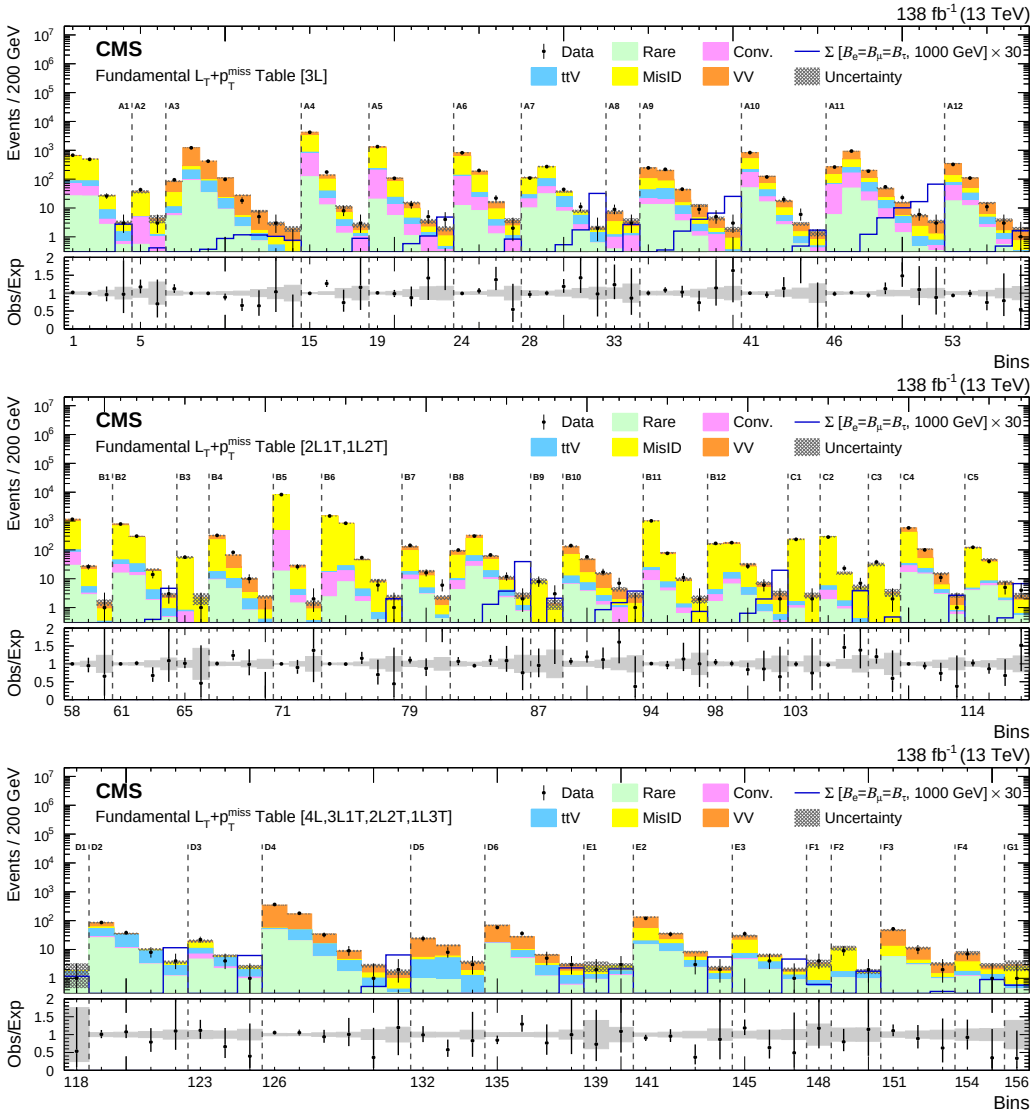


FIGURE 8.2: Model-independent L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ signal regions for the combined 2016-2018 dataset. L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ spectrum is shown for all event categories defined in the text. L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ bin boundaries are defined in Tables III to VI in Ref. [23]. 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. Individual background yields from these SR bins are taken, and projected background yields are calculated for the HL-LHC scenario. The figure is taken from the multilepton paper [23].

increase in tracker acceptance, advanced machine learning algorithms, and improved identification techniques may result in better signal acceptance and low backgrounds, improving the signal significance for potential early discovery of such BSM particles. Nonetheless, the assumption gives an idea of the minimum sensitivity reach for such BSM searches if everything goes according to the plan (no detector malfunction, etc). Although the following points need to be taken into account for a reasonable estimate of the background and signal events:

- **Irreducible backgrounds**: Cross-section enhancements of the prompt processes WZ, ZZ, ttZ, and ttW at $\sqrt{s} = 14$ TeV p-p collider. This enhancement factor is taken as 1.02415 for WZ or ZZ, and 1.04839 for the ttW or ttZ process.

- **Reducible background**: Estimating misidentified backgrounds from the MC samples is challenging, as different processes contribute to the total misidentified background in different channels considered here. Misidentified backgrounds are estimated in a data-driven way in the actual analysis. The yield of misidentified background in each SR bin can be scaled to the total data luminosity of HL-LHC, but as recommended the statistical uncertainty is scaled down as $1/\sqrt{\frac{\mathcal{L}_{HLLHC}}{\mathcal{L}_{Run2}}}$. The systematic uncertainty is taken conservatively as the maximum variation (in Run-2) of misidentification yield in each channel.

- **Signal acceptance**: Vector-like electrons and muons are studied for the first time in the context of such projection. The signal acceptance, including the vector-like taus, is estimated from the simulated samples with Run-2 detector conditions as used in the Run-2 analysis. However, to calculate the signal yield, the cross-section of the doublet and singlet model (independent of flavor coupling, and only dependent on mass) at NLO precision at $\sqrt{s}$ =14 TeV needs to be derived.

- An **independent framework** is established to estimate the signal yield for all flavor coupling and model scenarios. The backgrounds of all seven multilepton channels in each SR bin are preserved in the internal CMS formats (datacards and corresponding histograms). Hence, the Run-2 background yield is taken from such files. Since the datacard level histograms are present, we can do a full systematic study. But, before going forward, this independent framework must be validated against the present Run-2 results published for the vector-like lepton coupling to the third-generation SM leptons.

8.3 Closure with Run-2 results

To validate the framework and methodology, we have produced the Run-2 sensitivity (limits) for the Vector-like tau (doublet) model and compared it with the published results. Our study reproduces the published results quite well despite the following limitations,

- 2018 signal samples were used to estimate signal yield for the full Run 2 dataset in this method. In contrast, the analysis used separate MC signal samples, corresponding scale factors, etc, for the three-year data-taking period.
- Correlation and impact of the nuisance parameters are simplified to enable this HL-LHC projection study. This may impact the derived limits using the current method compared to the actual analysis strategy, where correlation and variation of each nuisance parameter were taken realistically.

Simplified Run-2 systematics for validation

Systematics due to intrinsic detector limitations are estimated from the Run-2 analysis in a simplified manner. To simplify the Run 2 systematics, we decided to vary the yield of the different background and signal processes by the upper boundary of the impact listed in the systematic table of the multilepton analysis [23] to validate our framework. Table 8.1 shows the sources and magnitudes of Run-2 systematic uncertainties in our simplified systematic approach considered in this study.

Closure

Given these differences, reasonable closure with the published analysis sensitivity is established for each signal mass hypothesis in Fig 8.4 for vector-like tau leptons. It is worth noting that, near the exclusion, our methods accurately reproduce the published limit. Near exclusion, the analysis was statistically limited than the systematic uncertainties. We have also reproduced the acceptance of the vector-like tau lepton (doublet) model for different VLL masses. Fig 8.3 shows a nice closure with the published signal acceptance (acceptance * efficiency is implicitly meant). This closure established that the tool can be applied to extrapolate sensitivity for these models at HL-LHC.

Uncertainty source	variation	processes
Statistical	1-100%	All MC samples
Luminosity	2.5%	All MC samples
Electron/Muon reco., ID and iso. efficiency	5%	All MC samples
Tau reco., ID and iso. efficiency	25%	All MC samples
Lepton displacement efficiency	5%	All MC samples
Trigger efficiency	3%	All MC samples
b tag efficiency	5%	All MC samples
pileup	3%	All MC samples
PDF, fact./renorm. scale	10%	All MC samples
Jet energy scale	5%	All MC samples
Unclustered energy scale	2%	All MC samples
Muon energy scale and resolution	5%	All MC samples
Electron energy scale and resolution	5%	All MC samples
Tau energy scale	5%	All MC samples
Jet energy scale	5%	All MC samples
Electron charge misidentification	25%	All MC samples
WZ normalization	5%	WZ
ZZ normalization	5%	ZZ
ttZ normalization	25%	$t\bar{t}Z$
Conv normalization	50%	$Z\gamma/Conv$
Rare normalization	50%	Rare
Prompt and misidentification rates	50%	MisID
DY-tt process dependency	25%	MisID
Diboson jet multiplicity modeling	30%	WZ/ZZ
Diboson PT modeling	15%	WZ/ZZ

TABLE 8.1: Simplified Run-2 systematics following the multilepton paper systematics table (Table IX) [23]

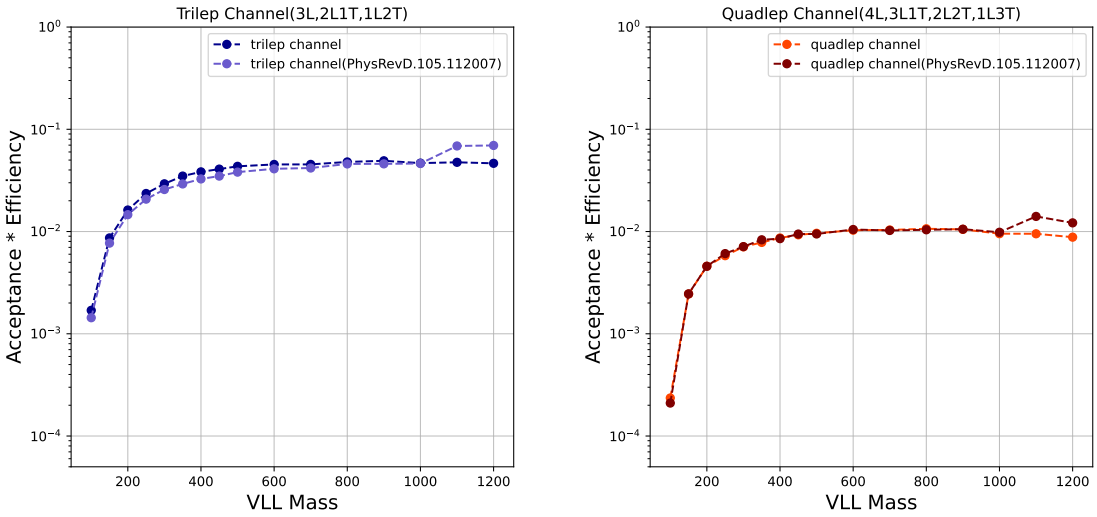


FIGURE 8.3: Comparison of our methods with the published analysis in the trilepton (3ℓ,$2\ell1\tau_h$,$1\ell2\tau_h$) and quadlepton (4ℓ,$3\ell1\tau_h$,$2\ell2\tau_h$,$1\ell3\tau_h$) channel acceptance*efficiency as a function of different mass hypothesis for the vector-like tau (doublet) model in Run-2 condition.

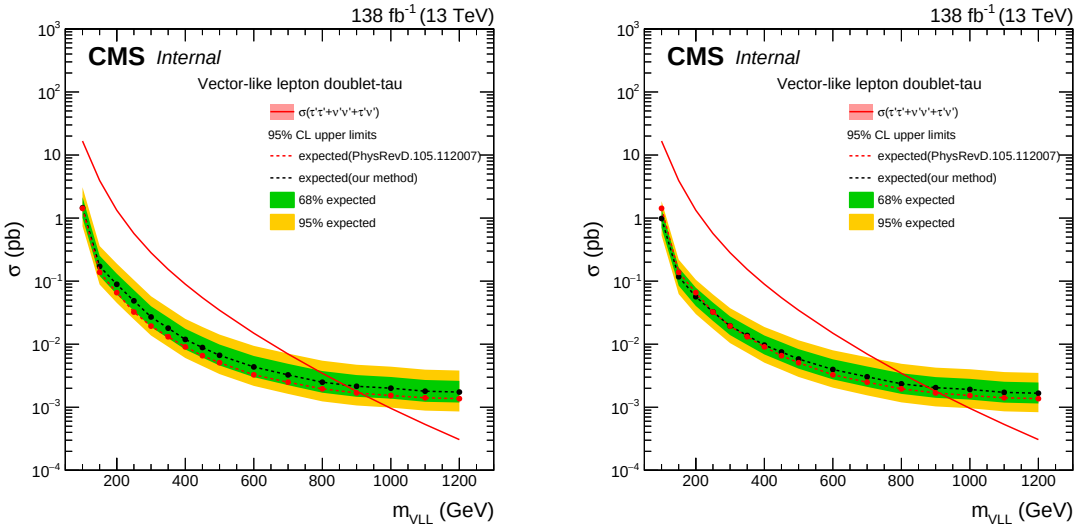


FIGURE 8.4: Comparison of the expected limit of the vector-like tau (doublet) model as a function of VLL mass in the published paper and our method using simplified Run-2 systematic(left) and stat. only(right) uncertainties.

8.4 VLL NLO cross-section at $\sqrt{s} = 14$ TeV

The last piece of our projection strategy is computing the NLO cross-section of the considered signal hypotheses at $\sqrt{s}$ =14 TeV. We consider the vector-like lepton (VLL) singlet and doublet model for the HL-LHC projection study. Depending on the coupling with the SM leptons, both models have three variants: VLL-electron, VLL-muon, or VLL-tau. Although the production cross-section is independent of VLL coupling with the SM leptons. To calculate the cross section for different mass hypotheses of a particular signal model at $\sqrt{s}$=14 TeV at NLO precision, we followed the following procedure,

- We used MadGraph to generate events, and the cross-section is taken for different signal mass points at $\sqrt{s}$=13 TeV (under the same assumption of PDF, Renormalization, and Factorization scale, etc.).
- Derived k-factor (σ_{NLO}/σ_{LO}) at 13 TeV from the ratio of NLO cross section(from the model Authors) and madgraph generated cross-section (LO).
- Similarly, to get the LO cross-section at $\sqrt{s}$=14 TeV, we generated events for signal mass points on the MADGRAPH by changing the center of mass energy on the MADGRAPH level cards.
- Finally, we multiply the k-factor(derived at 13 TeV) to get the NLO cross section at 14 TeV, i.e., $\sigma_{NLO}^{14TeV} = k_{13\ TeV} * \sigma_{LO}^{14\ TeV}$.

Table 8.2 and Table 8.3 show the cross sections and k-factor for VLL singlet and doublet models at $\sqrt{s}$=13 TeV and 14 TeV, respectively. Fig 8.5 and Fig 8.6 show the cross sections and k-factor for VLL singlet and doublet models at $\sqrt{s}$=13 TeV and 14 TeV, respectively.

8.5 Estimating HL-LHC yields and systematics

With the derived VLL cross-section at NLO precision, the signal acceptance is estimated from simulated samples with Run-2 detector conditions for all three coupling scenarios. Similarly, individual background yields of 138 fb^{-1} analysis in each SR bin are scaled to 3000 fb^{-1} of integrated luminosity. In addition, prompt backgrounds (WZ, ZZ, ttW, or ttZ) yields already take into account the enhancement due to the higher center-of-mass energy as described earlier.

Following the Yellow Report on BSM physics at the HL-LHC and High-Energy LHC [89], experimental uncertainties for signal and background yields were considered. For example, the theoretical uncertainties are assumed to be reduced by a factor of two with respect to the current estimate. All the uncertainties related to the limited number of simulated events

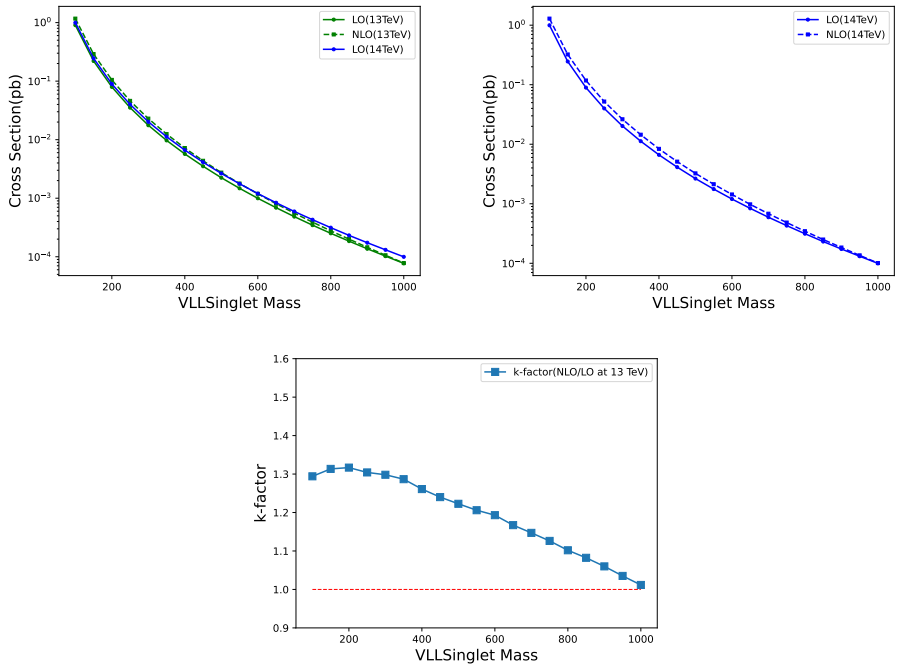


FIGURE 8.5: Figure shows VLL-singlet cross-section at LO and NLO precision for the $\sqrt{s}$= 13 TeV and 14 TeV (upper) and the ratio of σ_{NLO}/σ_{LO}(k-factor) at $\sqrt{s} = 13$ TeV (lower).

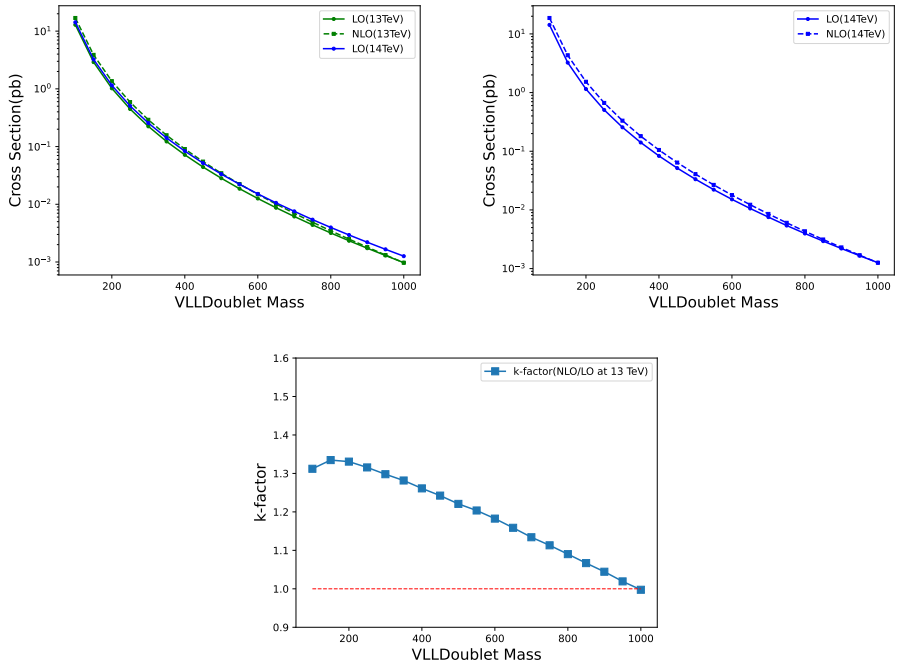


FIGURE 8.6: Figure shows VLL-doublet cross-section at LO and NLO precision for the $\sqrt{s}$= 13 TeV and 14 TeV (upper) and the ratio of σ_{NLO}/σ_{LO}(k-factor) at $\sqrt{s} = 13$ TeV (lower).

Mass	LO(13TeV)	NLO(13TeV)	k-factor	LO(14TeV)	NLO(14TeV)
100	0.9041	1.17	1.2941	0.9964	1.28945
150	0.2208	0.29	1.31341	0.2455	0.322441
200	0.07974	0.105	1.31678	0.0894	0.11772
250	0.03527	0.046	1.30422	0.03999	0.0521559
300	0.01764	0.0229	1.29819	0.02022	0.0262493
350	0.009715	0.0125	1.28667	0.01124	0.0144622
400	0.00571	0.0072	1.26095	0.006635	0.00836637
450	0.003516	0.00436	1.24005	0.004119	0.00510775
500	0.002241	0.00274	1.22267	0.002649	0.00323885
550	0.001476	0.00178	1.20596	0.001765	0.00212852
600	0.0009974	0.00119	1.1931	0.001201	0.00143292
650	0.0006897	0.000805	1.16717	0.0008371	0.000977041
700	0.0004847	0.000556	1.1471	0.0005936	0.000680919
750	0.0003463	0.00039	1.12619	0.0004294	0.000483586
800	0.0002514	0.000277	1.10183	0.000314	0.000345975
850	0.0001848	0.0002	1.08225	0.0002318	0.000250866
900	0.0001368	0.000145	1.05994	0.0001737	0.000184112
950	0.0001024	0.000106	1.03516	0.0001313	0.000135916
1000	7.71e-05	7.8e-05	1.01167	9.972e-05	0.000100884

TABLE 8.2: VLL Singlet model LO and NLO cross-section(in pb) at $\sqrt{s}$= 13 TeV and 14 TeV

are neglected, assuming that sufficiently large simulation samples will be available when the HL-LHC becomes operational. For all scenarios, the intrinsic statistical uncertainty in the measurement is reduced by a factor $1/\sqrt{L}$, where L is the projected integrated luminosity (3000 fb^{-1}) divided by that of the reference Run-2 analysis (138 fb^{-1}). For the HL-LHC scenario, all these experimental systematics uncertainties are scaled down by the square root of the integrated luminosity factor. In order to implement the Yellow Report systematics uncertainties, uncertainty related to luminosity measurement is reduced to 1%.

8.6 Results

We use a modified frequentist approach with the CLs criterion, with a test statistic based on the binned profile likelihood, to calculate the upper limits, in the asymptotic approximation. The upper limits are calculated at 95% CL. The systematic uncertainties are incorporated in the likelihood as nuisance parameters with log-normal probability density functions. The statistical uncertainties in the signal and background estimates are treated accordingly in the Higgs combine tool [87].

Figure 8.7 shows the expected upper limits on the cross-section at the HL-LHC for the production of vector-like leptons coupled to first-, second-, and third-generation SM leptons. Both the singlet ($\mathrm{E_i}$) and the doublet ($\mathrm{E_i}$, $\mathrm{N_i}$) scenarios are considered, where i = 1, 2,

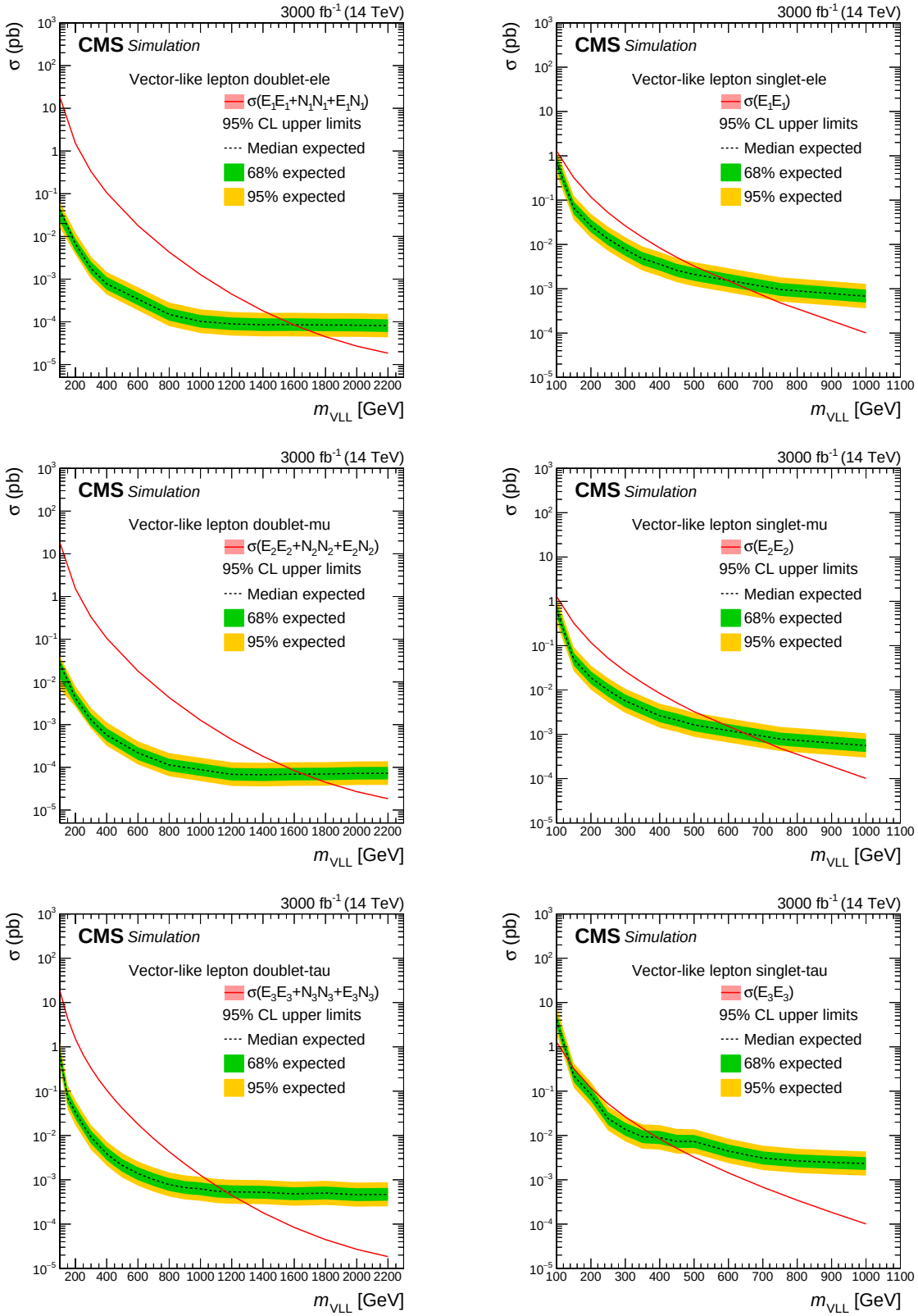


FIGURE 8.7: Expected HL-LHC exclusion limits for vector-like leptons coupled to first-generation (upper row), second-generation (middle row), and third-generation SM leptons (lower row) in the doublet model (left) and the singlet model (right). For both models, limits are calculated using L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ from the model-independent SRs for all masses.

Mass	LO(13TeV)	NLO(13TeV)	k-factor	LO(14TeV)	NLO(14TeV)
100	12.88	16.9	1.31211	14.24	18.6845
150	2.907	3.88	1.33471	3.24	4.32446
200	1.022	1.36	1.33072	1.149	1.529
250	0.4477	0.589	1.31561	0.5066	0.66649
300	0.2242	0.291	1.29795	0.2559	0.332145
350	0.1225	0.157	1.28163	0.1411	0.180838
400	0.07192	0.0907	1.26112	0.08338	0.105152
450	0.04419	0.0549	1.24236	0.05169	0.0642177
500	0.02826	0.0345	1.22081	0.03336	0.0407261
550	0.01861	0.0224	1.20365	0.02219	0.0267091
600	0.0126	0.0149	1.18254	0.01514	0.0179037
650	0.008717	0.0101	1.15866	0.01058	0.0122586
700	0.006146	0.00697	1.13407	0.007541	0.00855203
750	0.004393	0.00489	1.11313	0.005418	0.00603096
800	0.003183	0.00347	1.09017	0.003966	0.0043236
850	0.002334	0.00249	1.06684	0.002938	0.00313437
900	0.001733	0.00181	1.04443	0.002199	0.00229671
950	0.001295	0.00132	1.01931	0.001658	0.00169001
1000	0.0009734	0.000971	0.997534	0.001262	0.00125889

TABLE 8.3: VLL Doublet model LO and NLO cross-section (in pb) at $\sqrt{s}$= 13 TeV and 14 TeV

3 denotes VLLs coupled to first-, second-, and third-generation SM leptons, respectively. For the doublet model, the VLLs are expected to be excluded at 95% CL up to a mass of 1600 GeV (E_1, N_1), 1630 GeV (E_2, N_2), and 1150 GeV (E_3, N_3). The singlet VLLs are expected to be excluded up to a mass of 600 GeV (E_1), 640 GeV (E_2), and between a mass of 150 and 395 GeV (E_3).

The less stringent constraints observed in the singlet model arise from the notably lower cross-section of VLL pair production proceeds exclusively through the $pp \to Z/\gamma \to EE$ and involves a weaker gauge coupling strength compared to the doublet scenario. Additionally, the prevalent $E \to W\nu$ decay mode in the singlet model might not result in energetically charged leptons in the final state. Slightly better sensitivity for vector-like muons than vector-like electrons can be attributed to the higher reconstruction and identification efficiency of muons than electrons. The weakest limit is obtained in the vector-like tau leptons, which is expected as the reconstruction and identification efficiency of hadronically decaying taus is the least. Vector-like taus in the singlet model are the hardest signal phase space to be sensitive to for future experiments. New techniques must be exercised to improve the experimental reach to such BSM particles at the electroweak scale.

These projected sensitivities using model-independent SRs are better by a factor of approximately 2–3 in the upper limit of the cross-section compared to the Run-2 results. The Run-2 results used a BDT to enhance sensitivity. In contrast, the HL-LHC will provide a

much larger dataset and the opportunity for more advanced ML techniques and optimization. Thus, the eventual reach in terms of VLL mass is expected to be higher than the projected sensitivities here. Current best constraints and expected HL-LHC reach on the mass of vector-like leptons for various models and coupling scenarios are presented in Table 8.4 to give an overview of the current status of VLL searches in the minimal model toward discovering such BSM particles.

Models	Best LHC limit	HL-LHC sensitivity	Experiment (current best constraints)
Singlet VLL-e (E_1)	320 GeV	600 GeV	ATLAS [26]
Singlet VLL-μ (E_2)	400 GeV	640 GeV	ATLAS [26]
Singlet VLL-τ (E_3)	125-150 GeV	150-395 GeV	CMS [23]
Doublet VLL-e (E_1,N_1)	1200 GeV	1600 GeV	ATLAS [26]
Doublet VLL-μ (E_2, N_2)	1270 GeV	1630 GeV	ATLAS [26]
Doublet VLL-τ (E_3,N_3)	1045 GeV	1150 GeV	CMS [23]

TABLE 8.4: Current status of the minimal vector-like lepton extension models at the LHC experiments.

Chapter 9

Representation learning and generative networks

The quest for new physics has driven the high-energy physics (HEP) community to build new powerful accelerators and maintain the operation of the existing state-of-the-art collider facilities. While pushing the energy frontier to its extreme, it is important to acknowledge that unique experimental techniques that explore subtle new physics signatures are crucial for the discovery potential of the available resources. Needless to say, this great potential for discovery comes with significant data challenges. With the increasing complexity and growing volume of data taken by the current experiments, the monumental challenge to isolate potential BSM signatures from the known SM footprints is an active area of research in HEP. ML algorithms are appropriate for analyzing large amounts of data. They can find more intrinsic patterns in the multidimensional data, including their applications in efficient data processing at the hardware level [101, 102]. On the other hand, the large data volume also poses a unique challenge to our simulation software used to mimic the collision events. To ensure good statistical precision across the full parameter space of a high-energy process, simulated samples typically contain billions of events and correspond to an effective luminosity 5-10 times greater than that of the collected data. The simulation techniques in Section 4.1.2 describe that GEANT4 based Full Simulation techniques are quite slow and computationally expensive, but best for explaining data. Fast Simulation technique is much faster (at least 20 times) but comes with a cost of decreased accuracy. Future operations like high luminosity HL-LHC will take 20 times more data than what has been collected. Naturally, the need for a faster yet accurate version of the simulation chain is increasing [45]. Proposing a simulation chain using ML techniques has gained popularity in solving tasks that scale easily to huge data volumes.

In this chapter, a few techniques for discovering the hidden patterns in data and exploiting the multidimensionality of the dataset will be presented. Next, we will discuss deep neural network-based generative techniques to simulate p-p collision events in the context of a few physical processes.

9.1 Representation Learning: low-dimensional embedding of a multidimensional dataset

We described the application of supervised ML techniques to isolate the VLL signal from the overwhelming SM backgrounds in Chapter 7. This represents an event-level task, where properties of the event and its constituent objects were exploited to distinguish signal from background. An example of an object-level task was presented in Section 4.6, where ML is used to differentiate light leptons originating from prompt sources (such as W, Z, H, τ leptons, or BSM processes) from those arising from b-hadron decays.

One common aspect of those tasks is that the data space is multidimensional. For example, many light-lepton properties were exploited, or the signal and background events were assigned a list of properties. We have also enjoyed the availability of training examples and a fairly balanced dataset with mostly equal numbers of class examples to train the ML algorithms. However, these complex algorithms often have suboptimal performance in more challenging scenarios when training data is limited and one class is represented more than the other in the training dataset. The questions we want to ask are the following:

- Is it possible to draw decision boundaries using simplified and faster algorithms such that the signal and background (noise) can be optimally separated?
- The above task assumes that we know the signal and background beforehand. In other words, class labels are available before training. What if the class labels are not available or are contaminated? Can we extract any information from the data available?
- Do we need multiple data features to attain optimal performance? The structure of the data might be preserved in a low-dimensional representation. One direction could be dropping irrelevant variables pertinent to the classification problem from training. In this way, we can reduce the data dimension, but would it be possible to construct a low-dimensional representation that captures the most important local and global data features?

It is often possible to reduce the dimension of the multidimensional dataset considerably while still retaining much of the information in the original dataset. Principal component analysis (PCA) is probably the best-known and most widely used dimension-reducing technique for accomplishing the task. Also, the Uniform Manifold Approximation and Projection (UMAP) is a nonlinear dimension reduction technique to visualize high-dimensional data in a lower-dimensional latent space [103]. Figure 9.1 illustrates the low-dimensional representation, denoted by $\vec{x}_m$, of the original feature vector ($\vec{x}_n$), where $n >> m$. A

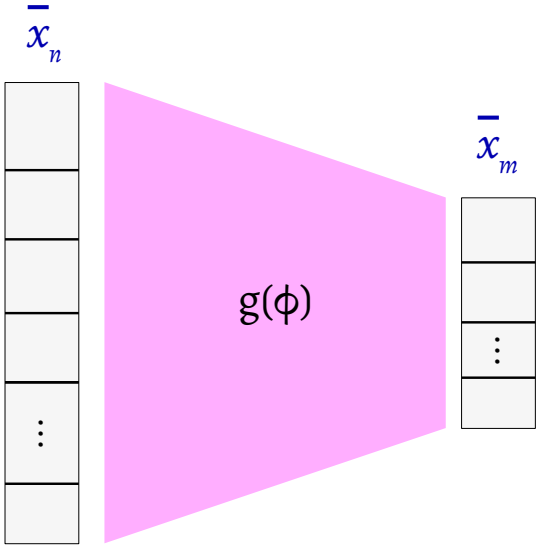


FIGURE 9.1: Basic idea of representation learning, denoted by $g(\phi)$, a linear or non-linear function that maps high dimensional feature vector ($\vec{x}_n$) to a low dimensional latent vector ($\vec{x}_m$), where $m << n$. The low-dimensional embedding space is devised to preserve the local and global structure of multidimensional data.

linear or non-linear mapping, denoted by $g(\phi)$, is a trainable function that maps the high-dimensional feature space to a low-dimensional latent space. In the literature, these unsupervised algorithms are primarily conceived as dimension-reduction techniques or as a data pre-processing step to transform correlated variables into various independent variables before feeding them to an ML algorithm. They are mainly used outside HEP in multiple studies, from searching for never-seen gravitational wave sources [104] to learning the features of symmetry-protected topological phase transitions [105]. In high energy physics, these algorithms were applied in studying initial state structures [106], collective flow in relativistic heavy-ion collisions [107], and anomaly detection [108]. PCA and UMAP preserve the global structure of the multidimensional data in a lower-dimensional latent space. The lower-dimensional embedding inherits the high-dimensional connection among the data points (density of points) belonging to the same class. It could be useful to discriminate between different classes available in the data (which we are unaware of). This motivates us to explore the possibility of using these unsupervised algorithms to analyze the data space to find intrinsic patterns.

The following dimension reduction models are used in the context of the prompt-fake lepton dataset as described in Section 4.6. This dataset is extracted from a CERN open dataset on $t\bar{t}$ MC samples applying the same event selection. The lepton properties are discussed later in the text.

9.1.1 Algorithms

Principal Component Analysis

Principal component analysis is a statistical method to analyze big data, which uses orthogonal transformations to transform a set of correlated variables into various independent variables. Thus, PCA is most widely used to reduce the dimension of a multidimensional dataset, keeping the crucial features intact to describe the global structure of data. Suppose we have N measurements on a vector x of n random variables, and we want to reduce the dimension from n to m, where m is typically much smaller than n ($m << n$). PCA does this by constructing a covariance matrix(S_X) of the original dataset($X_{N\times n}$) and then calculating the eigenvalues and eigenvectors of S_X. These eigenvectors form a basis to represent the old data, called principal components. Eigenvalues give the variances of their respective principal components, and the component with the highest eigenvalue is called the First principal component. To reduce the dimension of the original dataset, m principal components that correspond to the m largest eigenvalues can be chosen. Thus a projection matrix ($W_{n\times m}$) can be formed from the diagonalized covariance matrix S_X, and can be applied on original dataset $X_{N\times n}$ to reduce it to a new dataset $Y_{N\times m}$ as shown in the Equation 9.1

$$Y_{N\times m} = (W^T_{m\times n} X^T_{n\times N})^T \tag{9.1}$$

Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) is a nonlinear dimension reduction technique to visualize multidimensional data in a lower-dimensional space. UMAP constructs a fuzzy simplicial set representation using local manifold approximation to find the topological representations of the high-dimensional data. The embedding is found by searching for a low-dimensional projection of the data that has the closest possible equivalent fuzzy topological structure. Hyperparameters that need to be optimized in this algorithm are:

- **n_neighbours**: maximum number of local data-points considered to construct the high-dimensional graph.
- **min_dist**: minimum distance between two data-points to be connected in constructing the graph.
- **n_components**: dimension of the latent space (low-dimensional embedding)
- **metric**: distance metric in the hyperspace (eg, Euclidean distance).

t-distributed stochastic neighbor algorithm (t-SNE)

The t-SNE algorithm is another nonlinear dimension reduction technique [109]. It measures the similarity between data points in the high-dimensional space using a Gaussian kernel and constructs a low-dimensional representation using a Student t-distribution. The algorithm is then trained to learn the appropriate low-dimensional embedding that minimizes the difference between the similarities of points in high-dimensional and low-dimensional representations. The following hyperparameters are crucial for optimal performance of this algorithm,

- **Perplexity:** measure the effective number of neighbors for each point. The value of perplexity affects the balance between local and global aspects of your data. A small perplexity emphasizes local structure, while a larger perplexity brings more global structure.
- **n_components:** dimension of the latent space (low-dimensional embedding)

PCA, UMAP, or TSNE are categorized as unsupervised techniques that reduce the dimensionality of high-dimensional datasets while preserving the original structure and relationships inherent to the original dataset.

9.1.2 Low-dimensional visualization of prompt-fake leptons

Higher dimensional representation of a prompt or fake (non-prompt) lepton is defined by nine properties: SIP2D, SIPD$_Z$, charge and neutral EM energy fraction, charge and neutral hadronic energy fraction, number of neutral and charged particles in the mother jet, and the ratio of Lepton p_{T} and mother jet p_{T}. Figure 9.2 shows the input variable distribution for the prompt and non-prompt leptons used in this study. We can think of each lepton as an instance of a 9-dimensional data space, meaning each can be identified by a set of 9 real numbers ($\vec{x}_9$). The goal is to construct a low-dimensional embedding space where each lepton can be identified by two numbers (in case of a 2D latent space, $\vec{x}_2^{latent}$) with the expectation that the same type of object will be grouped to form a cluster.

PCA and UMAP models were implemented using the scikit-learn package and deployed to reduce the 9-dimensional data space to a 2-dimensional latent (or embedded) data space. It was found that normalizing the data between -1 to 1 gave the best performance for all the models considered. These reduced variables are not physically meaningful, but rather abstract variables made of the nine variables (either based on maximum variances in PCA or preserving the local connection in higher-dimensional data space). The population of prompt and fake leptons in the low-dimensional latent space is plotted as 2D scatter plots that can be easily visualized.

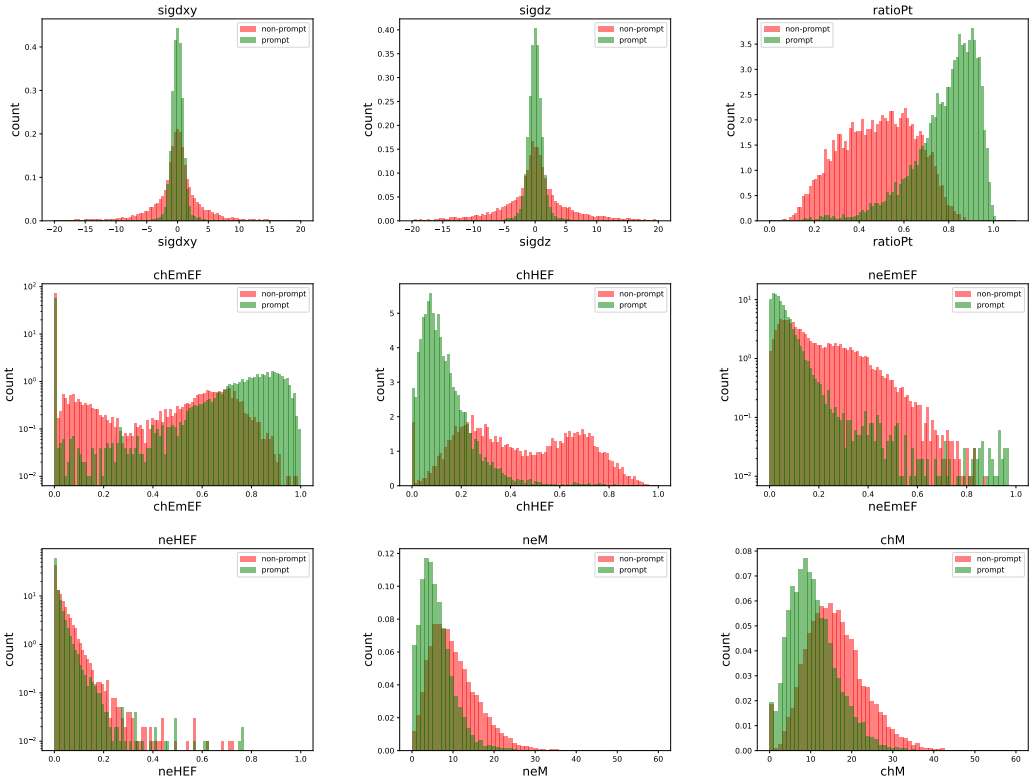


FIGURE 9.2: Input variables used in this dimension reduction study. These variables are extracted from the CERN Open dataset of $t\bar{t}$ MC sample.

The PCA algorithm diagonalizes the covariance matrix and finds the corresponding eigenvectors in the direction of maximum variance. In the vector space formed by the PCA eigenvectors, the direction of the maximum variance is defined as the first principal component. Similarly, the other components are also defined in the descending order of variance computed in the multidimensional space. Figure 9.3 illustrates the low-dimensional distribution of prompt and fake leptons in a 2D plane composed of the first four principal components. It is worth noting that the algorithm clusters the points based on the given properties. The label of each point is not provided during training. The plots are made by accessing the label at the testing level to show how these algorithms can cluster populations of the same class in a low-dimensional embedding. It can be seen that the third and fourth PCA components have the lowest variance of the data. It means that the first three components are enough to capture the important global and local information of the 9D space of the prompt and non-prompt leptons. This is illustrated in the 3D distribution of the first 4 PCA components in Figure 9.4.

Figure 9.5 demonstrates the low-dimensional embedding constructed by the PCA, UMAP, and TSNE algorithms. Note that the PCA is a linear algorithm, but the UMAP and TSNE are non-linear algorithms by construction. Latent representations produced by the UMAP or TSNE capture more features, and the population of different types of leptons is clustered

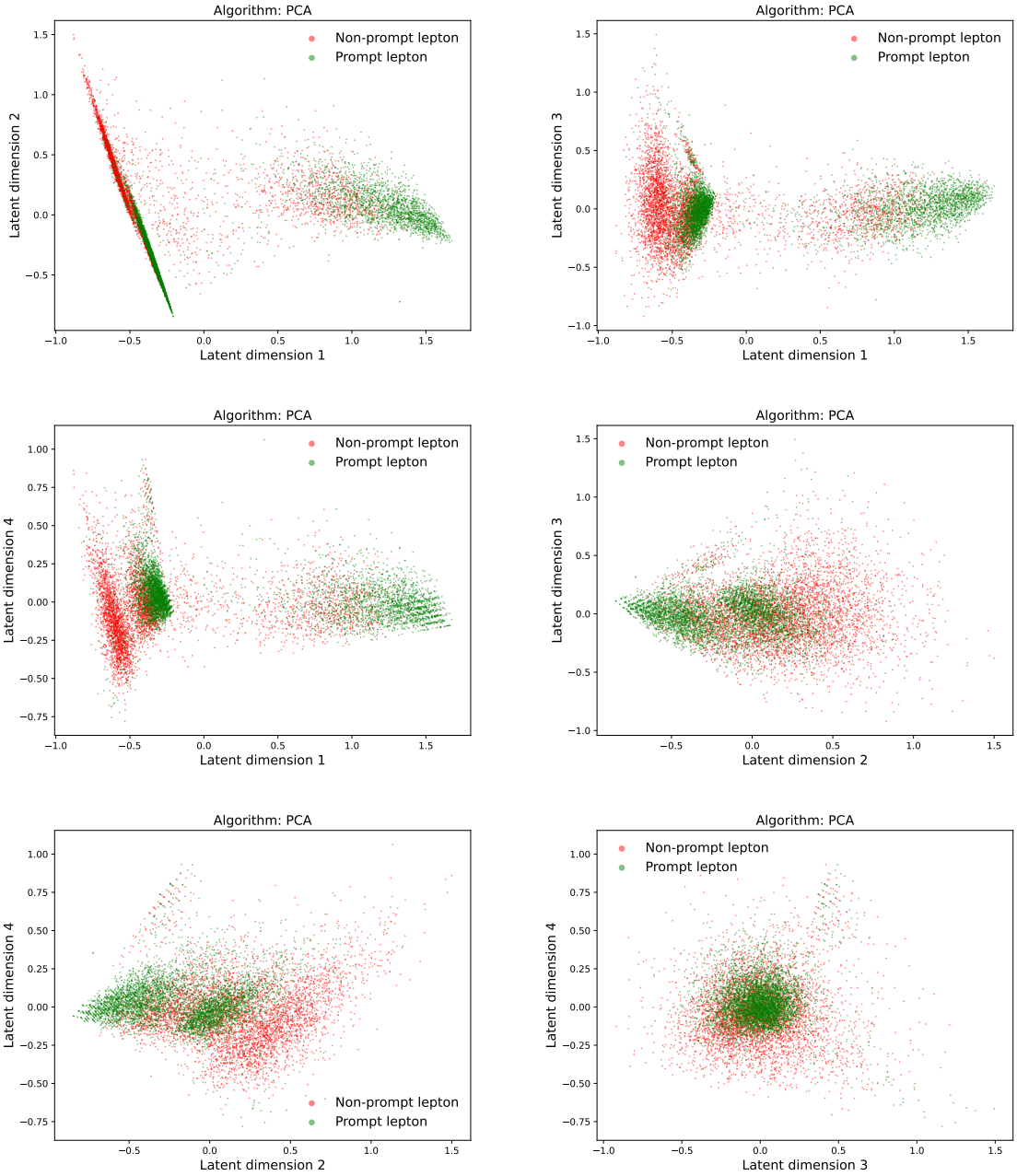


FIGURE 9.3: Low-dimensional (2D) embedding of the high-dimensional (9D) prompt and non-prompt leptons in the first 4 PCA components.

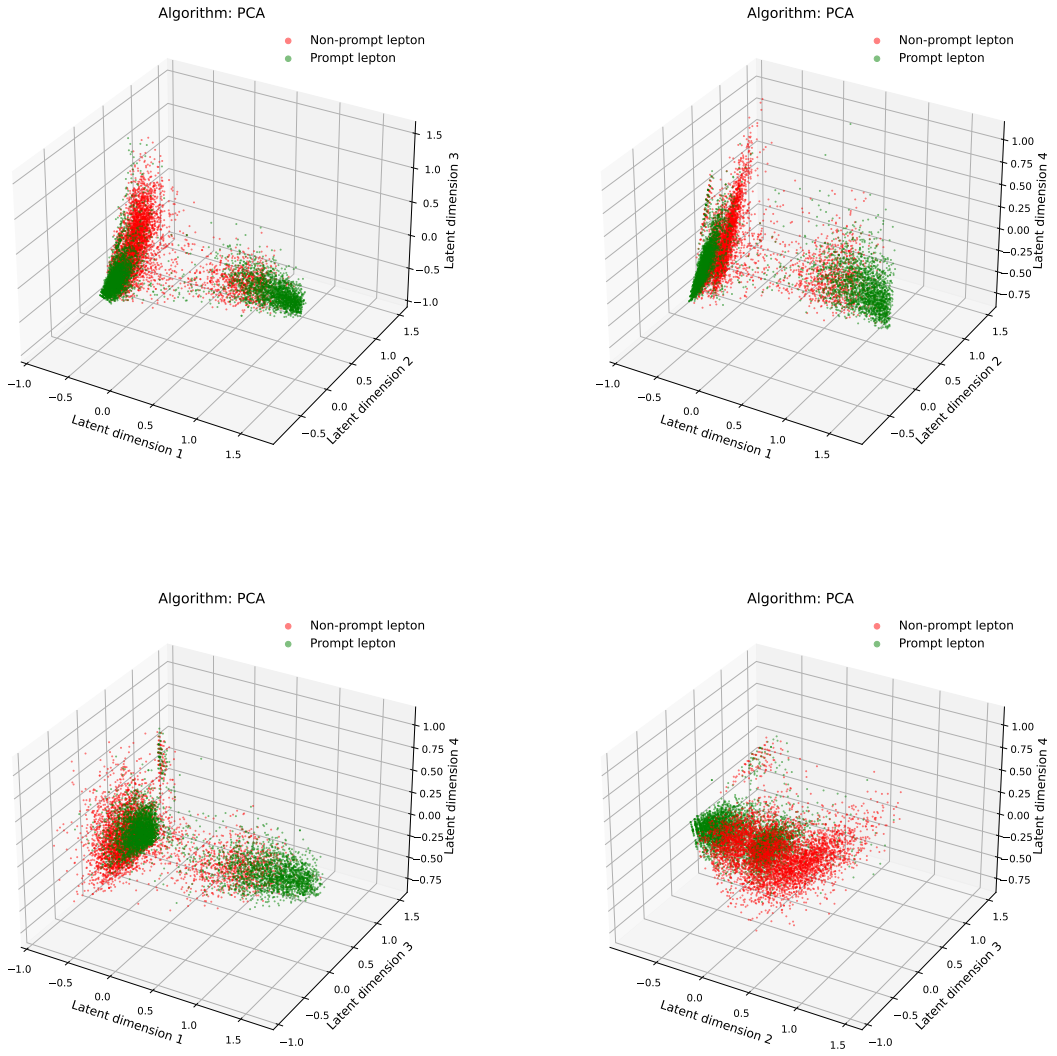


FIGURE 9.4: Low-dimensional (3D) embedding of the high-dimensional (9D) prompt and non-prompt leptons constructed using the first 4 PCA components.

together in an interesting pattern.

To understand the underlying lepton properties of the various clusters, we investigated the prompt and fake lepton clusters in low-dimensional embedding in terms of multiple properties that were used to construct the embedding. Figure 9.6 demonstrates the isolation value of the latent space for prompt and non-prompt embedding.

The plots indicate two kinds of non-prompt lepton clusters populated with isolated and non-isolated types. The UMAP and TSNE algorithms produce interesting patterns by grouping these non-isolated non-prompt leptons in a single cluster, and isolated non-prompt leptons are clustered in a different phase space of low-dimensional latent space, as illustrated by the middle (UMAP) and lower (TSNE) plots of Figure 9.6. Note that the isolated non-prompt leptons are most difficult to distinguish as they look like prompt-isolated leptons. However, this low-dimensional embedding gives discriminatory power to remove this fake, apart from the usual isolation variable, which we can choose as a smaller value to remove non-isolated fakes. Another important property is the promptness of the leptons. Figure 9.7 shows the distribution of the SIP3D parameter as a measure of promptness for prompt and non-prompt leptons in the latent space. Highly displaced leptons are clustered in a small phase space of low-dimensional TSNE embedding, which gives an intuitive idea of how the high-dimensional feature space is structured. UMAP latent space also captures the large displacement of non-prompt leptons, but the low embedding space overlaps significantly with the prompt leptons.

The UMAP or TSNE algorithm demonstrates three kinds of population clusters in the latent space, as seen in three distinct lobes. In each lobe, the characteristics of non-prompt and prompt leptons are different from each other. Low-dimensional representation provides an intuitive idea of how a complex neural network learns to draw the non-linear decision boundary. PCA, being a linear decomposition of the data hyperspace, is unable to capture the richness of the dataset. It is possible to train a neural network downstream using the low-dimensional embedded features for a complex problem or where computing time is a concern. Dimension reduction algorithms with (or without) sophisticated deep neural network downstream can be explored to find anomalous regions of data or events (jets) in various HEP tasks. One disadvantage of UMAP or TSNE algorithms is that the clustering process is sensitive to the number of features (hyperspace), the distance metric in hyperspace (Euclidean metric might be too simple to measure distance in a hyperspace), and the starting point of the clustering procedure in hyperspace. These algorithms may produce different low-dimensional embeddings in each run, but the hyperparameters could be optimized to construct a fairly consistent low-dimensional embedding. This is an active research field in the ML community in the context of the unsupervised learning paradigm.

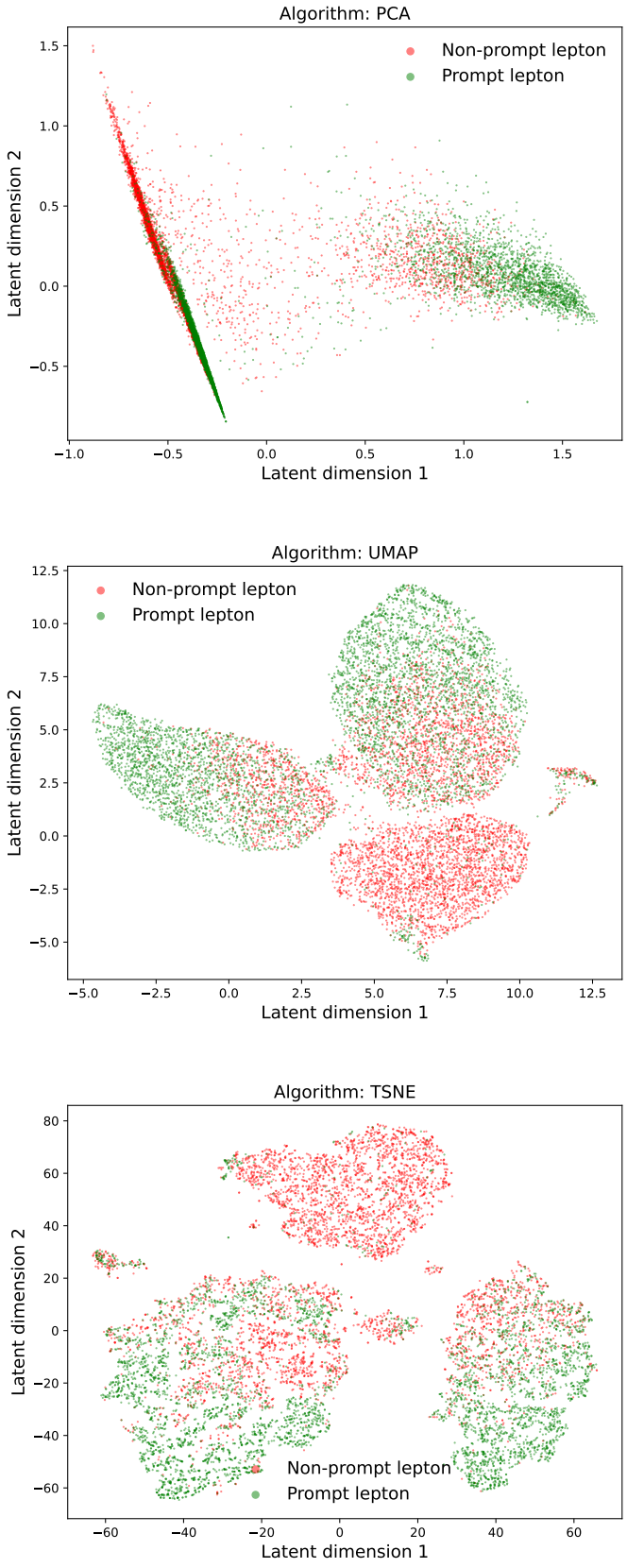


FIGURE 9.5: Low-dimensional (2D) embedding of the high-dimensional (9D) prompt and non-prompt leptons constructed using PCA, UMAP, and TSNE.

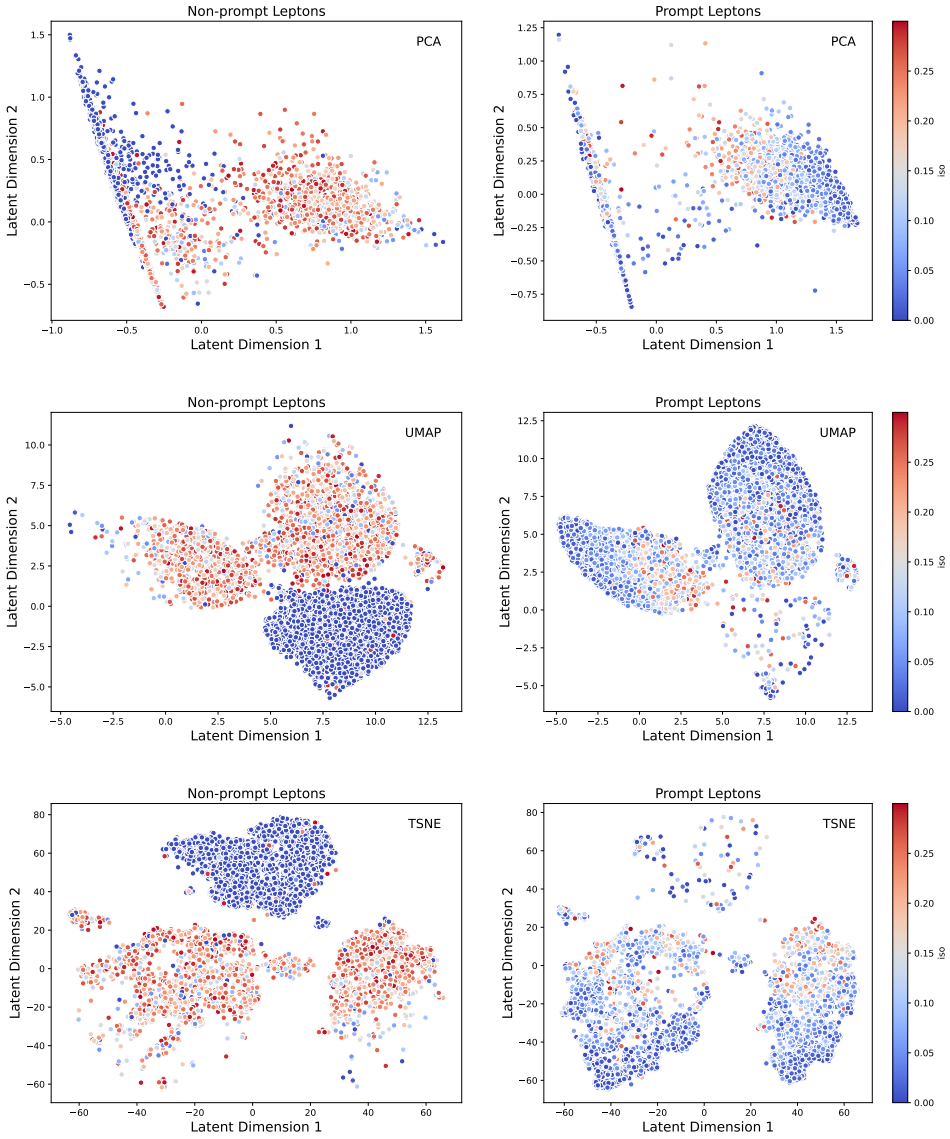


FIGURE 9.6: Low-dimensional (2D) embedding of prompt and non-prompt leptons constructed using PCA, UMAP, and TSNE as a function of isolation variable, which is not explicitly used in constructing the latent space.

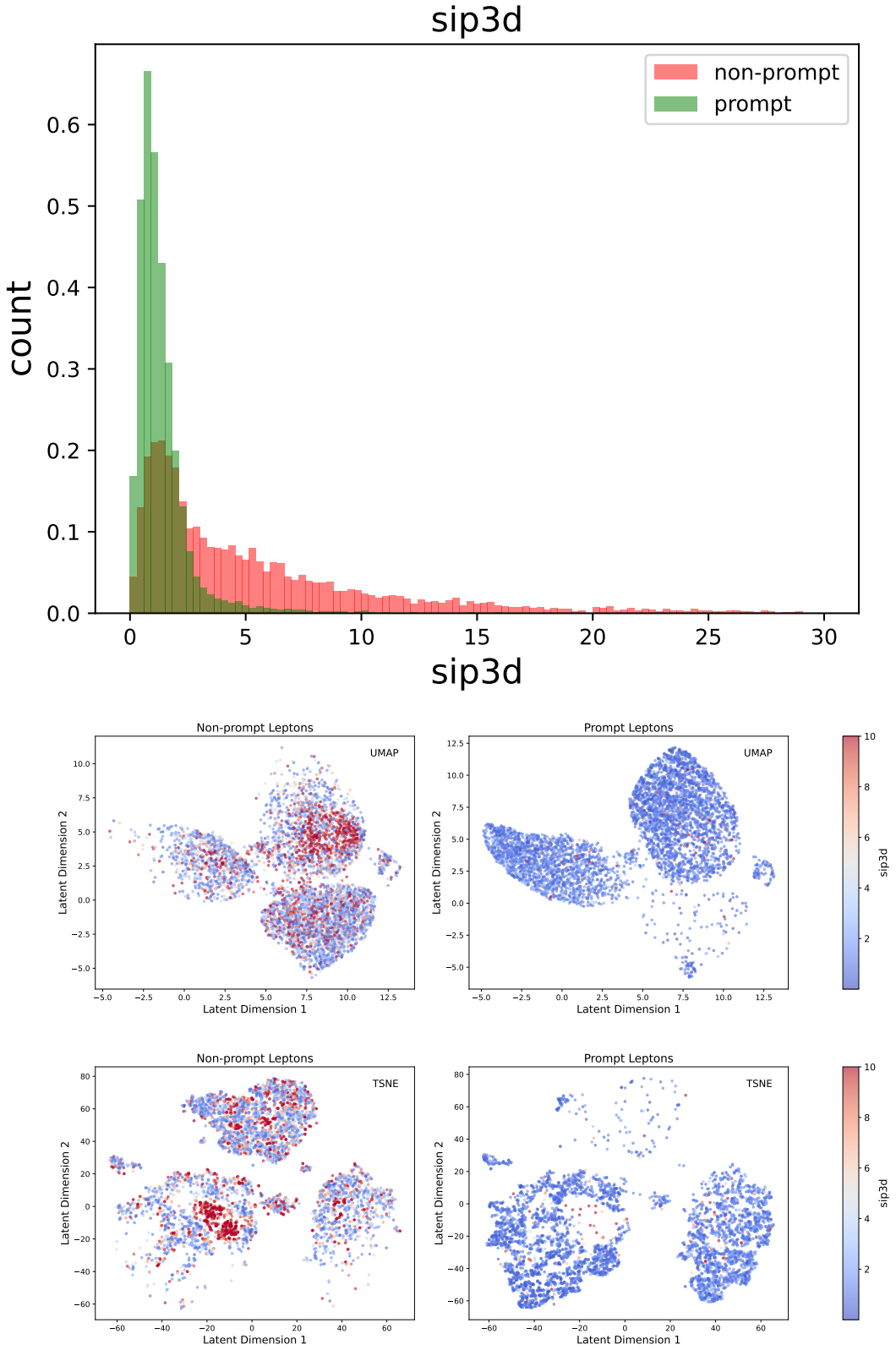


FIGURE 9.7: Distribution of SIP3D (top) and low-dimensional (2D) embedding of prompt and non-prompt leptons constructed using UMAP (middle), and TSNE (bottom) as a function of the SIP3D variable. Highly displaced leptons are clustered in a small phase space of low-dimensional TSNE embedding, which gives an intuitive idea of how the high-dimensional feature space is structured.

9.1.3 Classification using dimension reduction algorithms

The low-dimensional embedding using the PCA algorithm is used to construct a classifier to explore the potential of such algorithms to distinguish prompt and non-prompt leptons. The PCA algorithm is a linear model and computationally much cheaper than a deep neural network-based classifier. Sophisticated DNN is more complex, computationally costly to train, and may not perform optimally in unbalanced and low-statistics training scenarios. This led us to construct a PCA-classifier. PCA is an unsupervised algorithm, trained with unlabeled data, and we accessed the label later to identify prompt and non-prompt lepton low-dimensional embedding to construct a classifier.

To use PCA as a classifier, the trained PCA model is saved and used further to transform test data into the reduced 2D embedding. We exploit the clustering property of prompt and non-prompt lepton populations by defining a centroid for both clusters (at the training level), and then define a distance metric between any test data points and the cluster's centroid on the latent 2D plane to predict if the test data belongs to a prompt or fake cluster. The prompt centroid and non-prompt centroid are the mean position of all prompt and non-prompt leptons in the 2D latent space. It could be possible to compute the centroid for each cluster with other features, such as isolation or SIP3D as weights (equivalent to energy weighted sum), but this possibility is left for future studies. An extensive set of distance metrics has been studied to determine the best distance metric on this 2D latent space. Cosine similarity or cosine distance metric is found to be the best in terms of discriminating power. We also presented the results using the Euclidean metric for discrimination. For any test data point, the distance is computed from both clusters' centroids. The test data point assigned a label (prompt or non-prompt), whichever distance is smaller. Figure 9.8 shows the cosine and Euclidean distance metric for the prompt and non-prompt leptons test dataset. It can be seen that the cosine distance metric maximally separates the prompt and non-prompt leptons from the Euclidean metric, which might be susceptible to outliers.

To assess the performance of such a PCA-classifier, its performance is compared with that of a DNN-based classifier optimized using the same set of variables. A special training and testing strategy is carried out to evaluate the performance of the available models for the classification task. Training is categorized by the number of training leptons available for each class (prompt or non-prompt). The strategy is summarized in Table 9.1

ROC curves for the PCA and DNN binary classifier for all training scenarios are presented in Figure 9.10– 9.12. For completeness, the UMAP embedded space is also used as a classifier in the same way as the PCA classifier. Although the performance of UMAP is not optimized, it is left for future studies.

There are four ROC curves in each of the PCA and UMAP categories based on the choice of distance metric (cosine or Euclidean) and the centroids (prompt or non-prompt

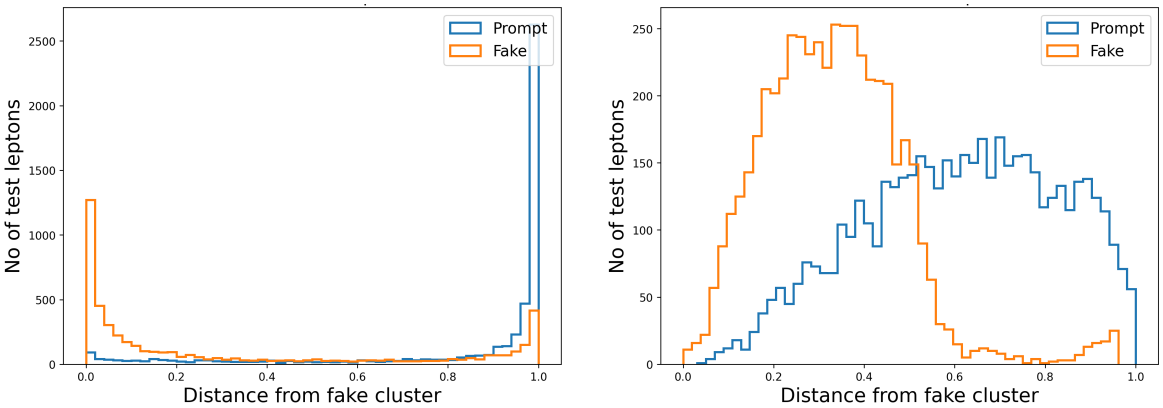


FIGURE 9.8: Cosine and Euclidean distance from prompt and non-prompt (fake) cluster's centroid for the test dataset. Based on a threshold in the distance score from the prompt or non-prompt centroid, a classifier can be built by predicting a test data point's prompt or non-prompt class. This PCA algorithm-aided classifier is simple and computationally cheaper than a deep neural network.

Type	N^{train}_{prompt}	$N^{train}_{non-prompt}$	N^{test}_{prompt}	$N^{test}_{non-prompt}$
High stat symmetric training	5000	5000	5000	5000
High stat asymmetric training	5000	2500	5000	5000
Low stat symmetric training	500	500	5000	5000
Low stat asymmetric training	500	250	5000	5000

TABLE 9.1: Training strategy based on the number of prompt and non-prompt leptons

cluster) used to calculate the distance score. Only the best-performing Euclidean and cosine distance scores of each category are presented based on the associated AUC values. DNN-based classifier performs better in all training scenarios than PCA and UMAP. It is not surprising since the PCA and UMAP are unsupervised learning techniques, while the DNN-based binary classifier is a supervised technique. On the other hand, a low-dimensional embedding after transforming the test dataset using PCA and UMAP is used to classify prompts from non-prompt leptons. PCA and UMAP show comparable discriminating power at low statistics training, the regime where DNN-based classifier training is not robust.

9.2 Generating events using Variational Auto Encoder

In this section, a deep neural network-based simulation chain is described using a special type of network called a variational autoencoder (VAE) [110]. This type of network can be used to generate simulated events of targeted physics processes in some predefined parameter space relevant to BSM searches. Many ML algorithms, such as VAE, Generative Adversarial Networks (GAN), or Normalizing Flow (NF) networks, can achieve such a high fidelity simulation chain to produce events. These networks can learn the transformation function of parton-level distribution (generator output) to the final physics observables. They can later be used to unfold the data (observables) to parton-level quantities. The unfolded distributions are crucial to tune the event generators to produce reliable and robust predictions of the particle collisions. The most popular usage is generating billions of events with less computing resources compared to the GEANT4 based full simulation. Not only physics variables, but these generative models can predict shower shape in calorimeters, hence can replace some part of the event reconstruction chain, resulting in a significant gain in computing time.

The VAE architecture is used to model the W+jets phase space in the VLL search described earlier in this thesis. The VAE model is trained to learn simulated (CMS full simulation) event properties, and later used to generate simulated events in the phase space used for training. This can be thought of as boosting sample statistics by producing more events in a particular phase space. Instead of using a comparatively slow GEANT4 based full simulation, the same can be achieved using VAE. The key performance metric is the accuracy of predicting such variables and their correlation. A complex network can learn the correlation and produce the multidimensional phase space of event properties.

Figure 9.14 illustrates the different components of a variational autoencoder network. It learns to encode input data into a probabilistic latent space and then decode from it to reconstruct the data. VAE architecture can be described in the following points:

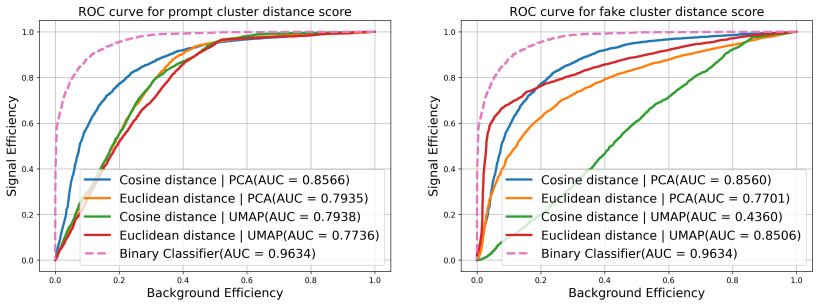


FIGURE 9.9: High stat symmetric training case

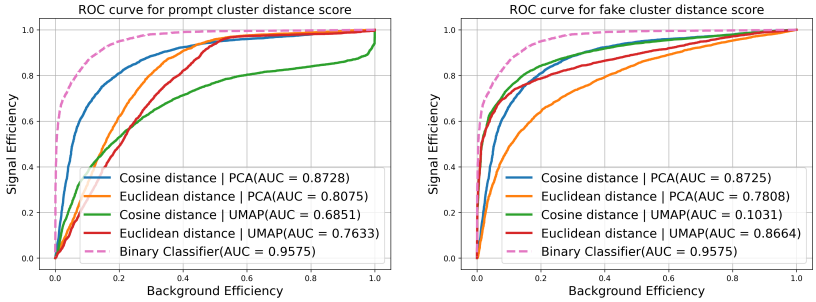


FIGURE 9.10: High stat asymmetric training case

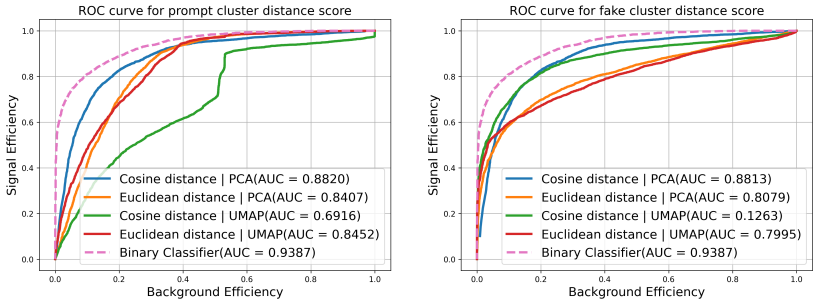


FIGURE 9.11: Low stat symmetric training case

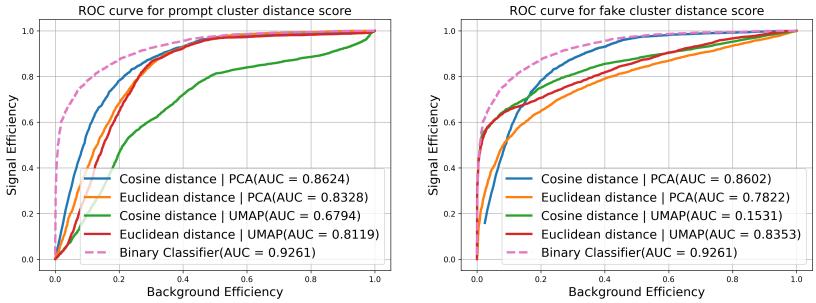


FIGURE 9.12: Low stat asymmetric training case

FIGURE 9.13: Comparison of the ROC curves for PCA and DNN-based binary classifier. The AUC value for the best-performing distance metric (Euclidean or cosine) and the choice of centroids (prompt or non-prompt cluster's centroid) are shown here. A classifier performance using the UMAP embedding space is also shown for completeness.

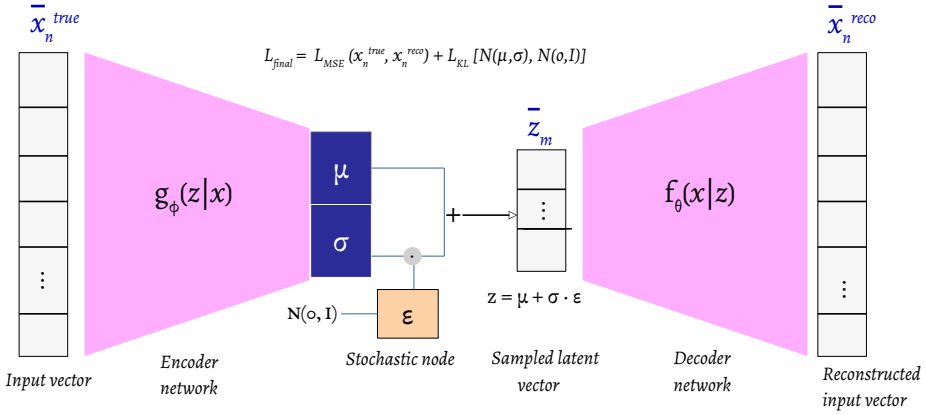


FIGURE 9.14: Neural network architecture of Variational Autoencoder (VAE)

- **Encoder**: This network mapped the input vector ($\vec{x}^{true}$) to the parameters of a probability distribution. Usually, the probability distribution is Gaussian: the encoder outputs a mean vector (μ) and a variance vector (σ) of the latent dimension m. This forms a posterior approximation:

$$g_\phi(z|x) = \mathcal{N}(z; \mu_\phi(\vec{x}, \sigma_\phi(\vec{x})I)) \tag{9.2}$$

  where z is the latent variable, and ϕ are the parameters (weights) of the encoder neural network.

  This is a crucial difference between an autoencoder and a VAE architecture. An autoencoder network suffers from a sparse, irregular latent space due to the deterministic mapping of an input vector to the latent space. Hence, the generative ability of an autoencoder is quite limited, as the quality of generated samples is inaccurate if the latent vector is sampled from outside of the distribution (OOD). The probabilistic approach in VAE encodes each input into a distribution rather than a single point, adding a layer of variability and uncertainty.

- **Latent space sampling**: To enable backpropagation through stochastic sampling, VAE uses the reparameterization trick:

$$z = \mu + \sigma.\epsilon, \epsilon \approx \mathcal{N}(0, I) \tag{9.3}$$

  This allows gradients to flow through the network during training by changing the stochastic node to a stochastic sampling from the unit multivariate Gaussian.

- **Decoder (Generative model)**: The decoder maps samples from the latent space back to the data space, trying to reconstruct the original input vector ($\vec{x}^{reco}$). The generative model can be written as $f_\theta(x|z)$, which transforms a latent vector into the

reconstructed input vector.

- **Regularization**: Regularization of the latent space defined by the multivariate Gaussian is achieved by forcing them to approximate a multivariate unit Gaussian whose mean and standard deviation are known. This is achieved by using a Kullback-Leibler (KL) divergence term. KL divergence is a measure of how one probability distribution (P(x)) differs from another, reference distribution (Q(x)).

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} P(x) \log \left(\frac{P(x)}{Q(x)} \right) dx \tag{9.4}$$

- **Loss objective**: One part of the loss measures how well the output of the decoder matches the input. It's often the log-likelihood of the data given the latent variable. Another one is the regularization term that forces the learned posterior $g_\phi(z|x)$ to be close to the prior $p(z) = \mathcal{N}(0, I)$.

$$\mathcal{L}_{\mathrm{VAE}}(x) = -\mathbb{E}_{f_\theta(z|x)}[\log q_\phi(x|z)] + D_{\mathrm{KL}}(q_\phi(z|x) \parallel p(z)) \tag{9.5}$$

Assuming that the decoder outputs a Gaussian distribution, the log-likelihood is reduced to Mean Squared Error (MSE) loss. The VAE loss function is minimized in the training to learn the parameters θ and ϕ of the two neural networks. We used $\beta-$VAE, where the KL divergence term can be weighted by the factor β in contrast to the MSE loss.

The VAE is trained to generate W+jets events selected by predefined criteria compatible with the analysis selection. The following variables are used for generation purposes:

- Lepton M_T: Transverse mass of the lepton
- M_{jj}: Invariant mass of the leading and subleading jets
- M_T^{jj}: Transverse mass of the dijet system
- $p_{\mathrm{T}}^{\mathrm{miss}}$: $p_{\mathrm{T}}^{\mathrm{miss}}$ of the event
- HT: scalar sum of all the jet p_{T} in the event

These input variables are normalized to a standard Gaussian before being fed into the network to have similar importance in calculating the MSE loss. The encoder network is composed of three hidden layers of 128 neurons each and four dense layers of 64, 32, 16, and 8 neurons, respectively. The encoding dimension (z) is optimized to be 3. ReLU activation functions were used in dense layers. Decoder network is the same as the encoder,

with the order of the layers reversed. The training is conducted in 1000 epochs with a batch size of 1024.

Figure 9.15 shows the agreement between generated and (CMS) simulated W+jets events (labeled as truth) for all the variables considered for this study. Although the bulk of the distribution is well modeled, the generated events struggle to model the tails of the distributions.

Figure 9.16 shows the pairwise correlation (using kernel density estimator) of the variables. It can be seen that the correlation is well captured in the generated samples, but fine-tuning of the VAE model is needed to improve the performance, specifically in the tails.

9.3 Future direction

The main task of generative modeling is defined by the challenge of approximating complex, high-dimensional probability distributions. VAE takes a probabilistic approach for learning latent representations and generating new data samples. VAEs learn a mapping from data to a latent space using an encoder, and from the latent space back to the data space using a decoder. The trained decoder model can be used to generate new samples given a sampled latent vector. VAE improves latent space embedding by regularizing it, enhancing the generation quality. Still, the samples tend to be blurry or lack fine-grained details, a limitation attributed to the element-wise or pixel-wise reconstruction loss and the Gaussian assumptions made in the model.

Another popular alternative generative model is the Generative Adversarial Network (GAN) [111]. GAN uses explicit likelihood formulation and operates in an adversarial training regime. A GAN consists of two neural networks: a generator G and a discriminator D, engaged in a two-player minmax game. The generator takes random noise $z \approx p_z(z)$ from a simple prior distribution (typically a multivariate Gaussian or uniform distribution). It maps it to the data space, aiming to produce realistic samples ($\vec{x}_{fake}$). The discriminator, on the other hand, receives either real data ($\vec{x_{real}}$) or generated data ($\vec{x}_{fake}$) and tries to distinguish between the two. The idea is that the generator learns to fool the discriminator, and the discriminator learns to improve its ability to differentiate real from fake. The loss objective is:

$$\min_G \max_D \; \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{9.6}$$

This represents the minmax game between the generator G and discriminator D, where $p_{data}(\vec{x})$ is the real data distribution, $p_z\vec{z}$ is the prior over the input noise variables. The log-likelihood that the discriminator will accurately categorize real data is represented by $log\ D(x)$, and correctly categorize generated samples as fake is represented by $log(1 - D(G(z)))$. Figure 9.17 illustrates the architecture of GAN. The generator network can be

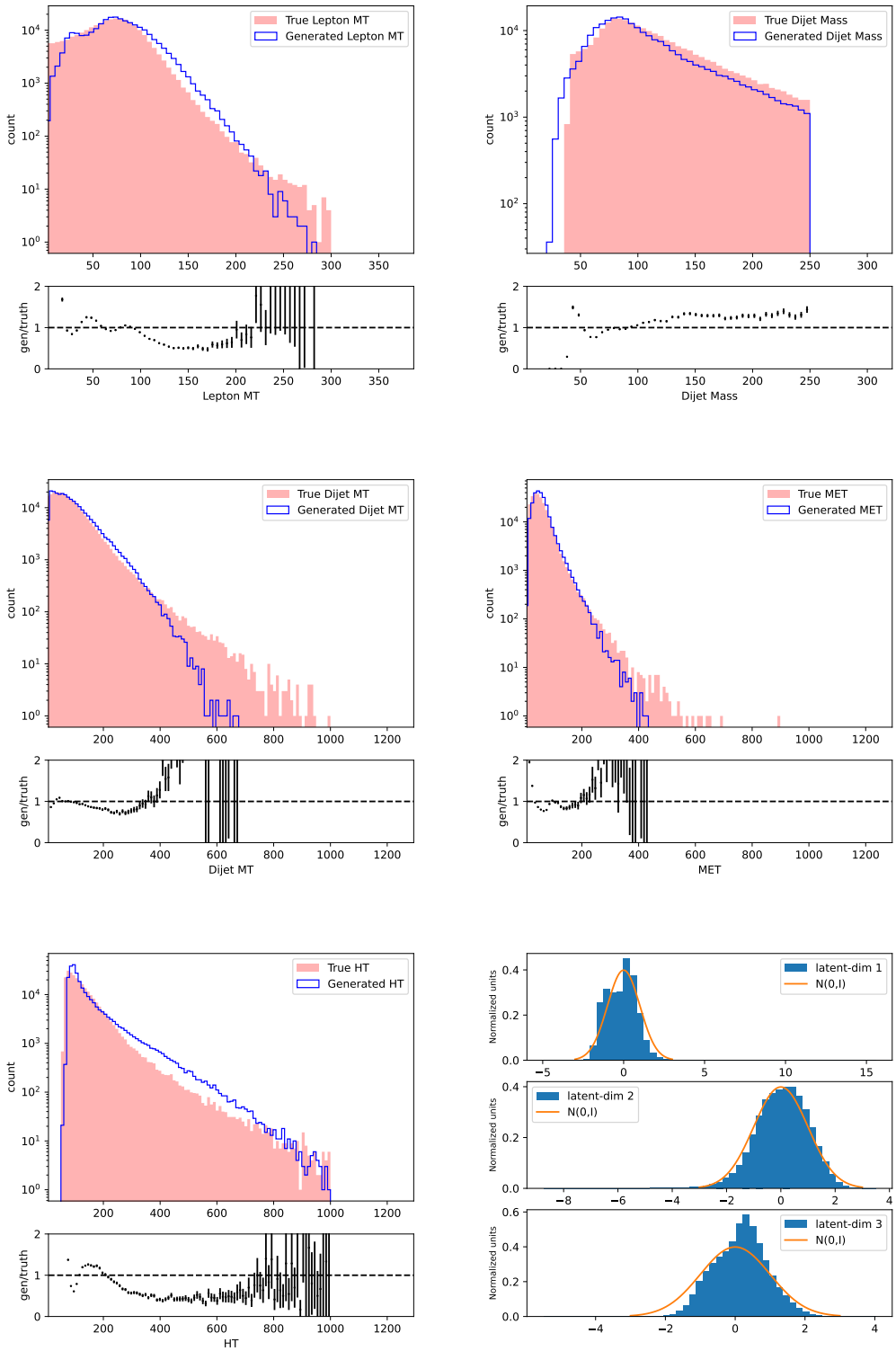


FIGURE 9.15: Agreement between the (CMS) simulated W+jets samples (truth) and VAE generated samples in different event or object properties. The ratio panel shows the ratio of gen and truth. Bottom right plot shows the modeling of 3-dimensional latent space wrt the standard Gaussian.

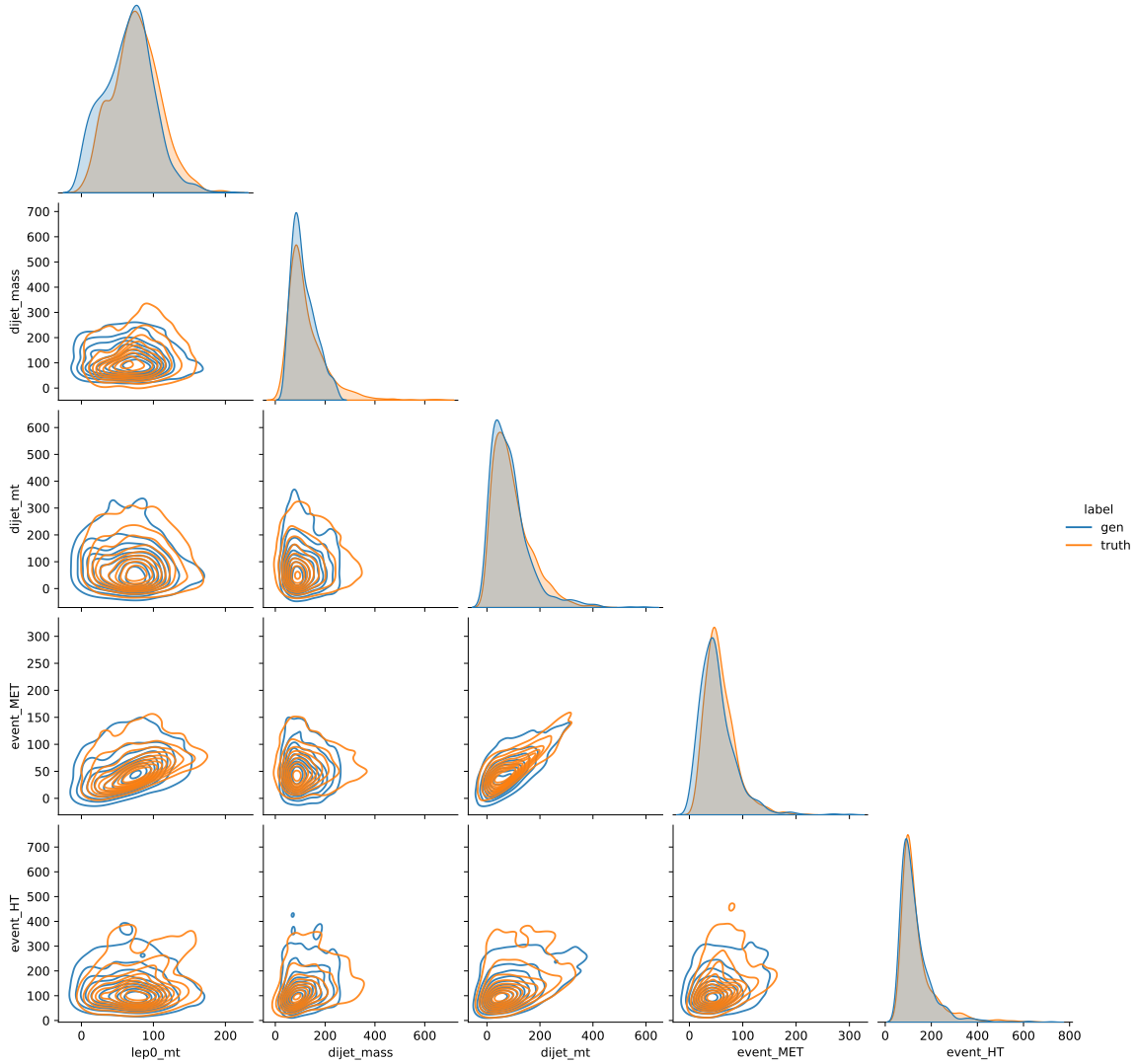


FIGURE 9.16: Pairwise correlation plot of the variables between generated and (CMS) simulated events. The generative model can capture the correlation between variables.

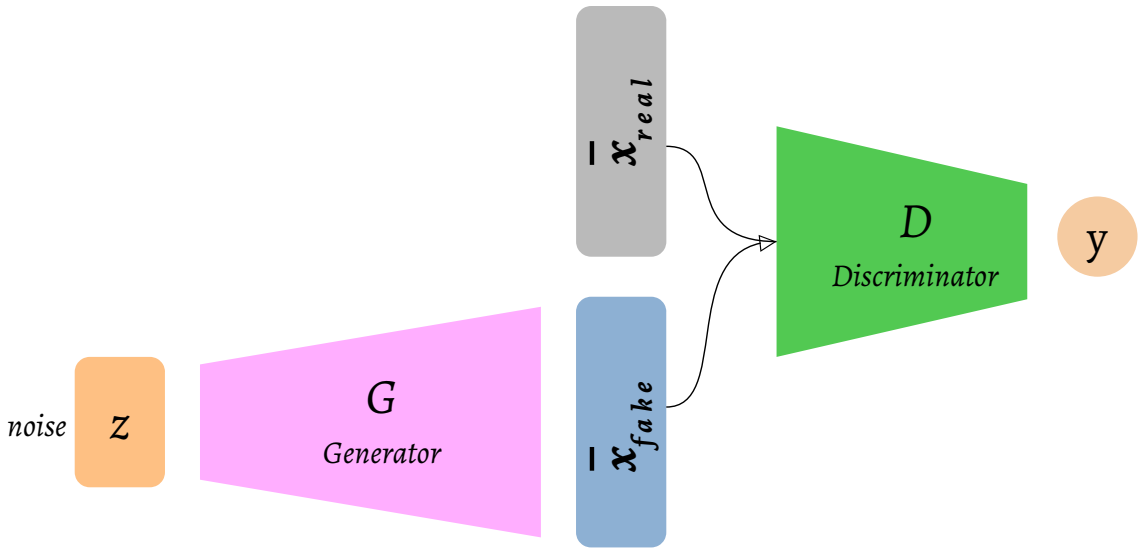


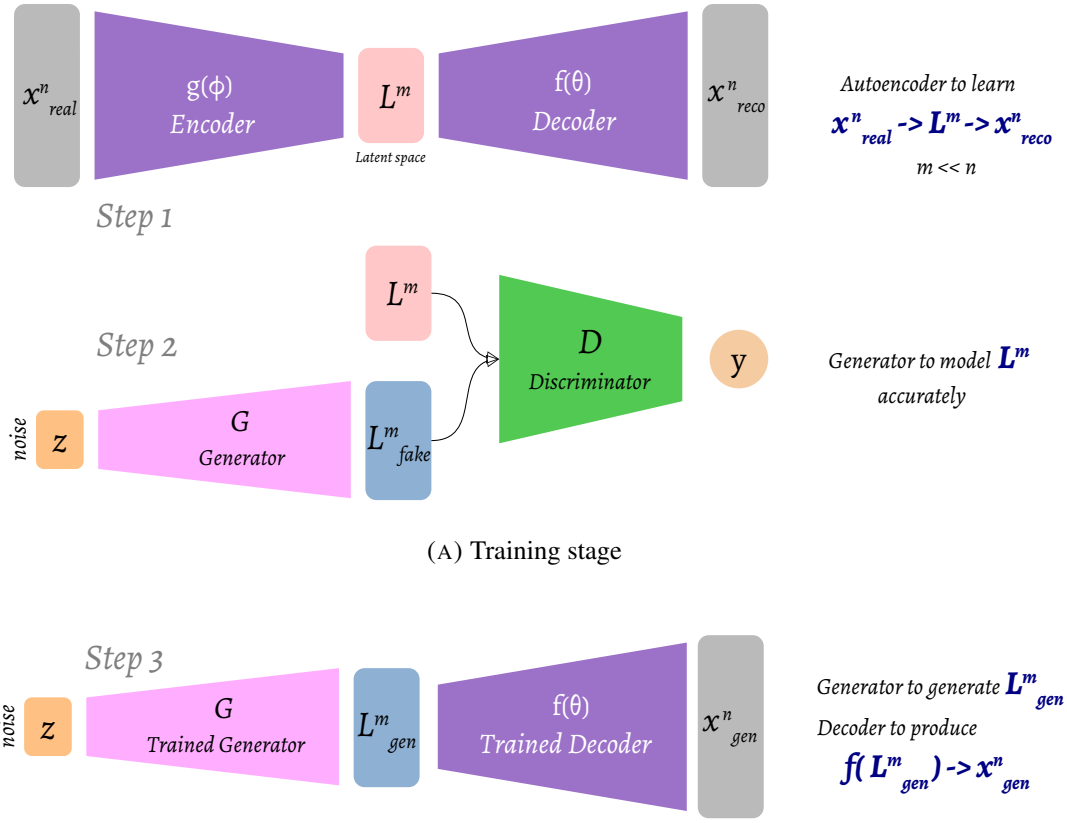
FIGURE 9.17: Neural network architecture of Generative Adversarial Networks (GANs). The generator (G) learns to fool the discriminator (D), and the discriminator tries to improve its ability to differentiate real from fake data during the training.

used to generate samples from the noise vector ($\vec{z}$) after the training is complete.

GANs do not require an explicit density function to model the data distribution; instead, they learn to generate samples that match the real data distribution through adversarial learning. GANs are extremely difficult to train due to the adversarial nature of the optimization. Common issues include mode collapse (where the generator produces a limited variety), vanishing gradients, and instability during training. One possible bottleneck is that implementing a GAN that gives usable output requires large, complicated networks and a high amount of computing, as it takes a large number of epochs to achieve reasonable generation quality.

A possible future direction to solve the bottleneck is to combine dimensionality reduction algorithms with GANs to construct an example pipeline for faster ML-based simulation. It can be done in three stages:

- Algorithms such as PCA, UMAP, or autoencoder are used to obtain a latent dimensional (m) representation of the original kinematic variables (n). (Of course, $m << n$).
- Algorithms such as VAE or GAN are used to generate the m variables with high accuracy.
- The generated m variables are used to obtain the n "generated" kinematic variables.



(A) Training stage

(B) Evaluation stage

FIGURE 9.18: Neural network architecture of dimension reduction GAN (DR-GAN). I) First step is to train an autoencoder architecture to encode the real data ($\vec{x}^n$) to a latent space ($\vec{L}^m$) and reconstruct the original data from the latent space. II) The Second step of the training process is to train a GAN architecture to generate the latent space accurately. III) Finally, the trained generator can be used to generate the latent space ($\vec{L}^m_{gen}$) and the trained decoder transforms the generated latent space to original data space ($\vec{x}^n_{gen}$) to have the generated samples. Given that the decoder performance is optimum, this procedure generates a much lower-dimensional latent space faster and uses less computing resources.

One advantage of this approach is that generating a smaller subset of numbers is an easier task for GANs and requires a more straightforward loss function. Additionally, less computing resources will be required to generate the m numbers (latent dimensions) than the n numbers (dimension of data). For example, we shall test the generation quality by comparing with a classifier trained using the kinematic variables as inputs. Figure 9.18 illustrates the three-step procedure of the dimension reduction GAN (DR-GAN). Step 1 involves an autoencoder architecture to encode the real data into a latent space and reconstruct the original data from the latent space. Step 2 of the training process is to train a GAN architecture to generate the latent space accurately. In step 3, the trained generator can be used to generate the latent space, and the trained decoder transforms the generated latent space into the original data space to produce the generated samples.

Chapter 10

Summary

This thesis presents a search for vector-like leptons coupling to the second and third generation SM leptons using a muon and at least two jets in the final state using proton-proton collision 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} . Vector-like leptons, if they exist, can solve some of the outstanding questions of nature, such as the stability of the Higgs mass at 125 GeV (hierarchy problem), mass hierarchies between different fermion flavor families. They also appear as a dark matter candidate in a few model-specific scenarios. This search is designed to cover a large model parameter space by targeting the VLLs coupling to second- and third-generation SM leptons. This is the first CMS search that probed the VLLs coupling to second-generation SM leptons.

The analysis probed minimal model of VLL extension to standard model in singlet scenarios which was particularly difficult to probe in the prior searches using multiple charged leptons at the final state due to their dominant decay modes to $E \to W\nu_\ell$ at low mass, and extremely small cross-section at high mass. By requiring less number of leptons and including hadronic activity from the gauge bosons in the decay modes, this channel selected more signal, but at the same time allowed huge SM backgrounds. The most dominant background in this search is the production of the W boson with associated jets in the event. The huge production cross-section of the W boson at the LHC, and overlapping signal characteristics, was a difficult challenge to suppress this background. Other irreducible and reducible such as the QCD multijet backgrounds, were suppressed by the event and object level selections. The backgrounds were estimated and validated in dedicated control and validation data samples before looking at the final signal regions, where a counting experiment was conducted. Deep neural network-based classifiers are used in a unique combination to suppress each background to enhance signal sensitivity. Unfortunately, no significant deviations from the SM expectations were observed. While this search did not have enough sensitivity to exclude phase space for the considered models, it exercised the power of good analysis to improve signal-to-background significantly. Harnessing advanced ML algorithms (like GNN or transformer) and exploring event shape variables to maximize the sensitivity of

such searches can be developed in the future. Combining Run-2 and Run-3 data may open the possibility to probe the high mass regime where these BSM particles are produced at a very low rate (EWK production). In another aspect, extending the searches to new event topologies to uncover more parameter space (model-independent or targeted) is also an exciting opportunity. For example, at high VLL mass, decay products of the gauge bosons (coming from VLLs) may be collimated and give rise to a fatjet signature at the collider. This is an interesting topology to probe new physics beyond the SM.

The discovery potential of such heavy leptons is studied at the HL-LHC scenario with $\sqrt{s}$ =14 TeV and 3000 fb^{-1} of data. Model-independent L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ signal regions of a published CMS analysis (138 fb^{-1}) that performed a search for new BSM phenomena in multilepton final states are utilized to extrapolate the sensitivity at HL-LHC. A comprehensive study is carried out to project the HL-LHC discovery reach for vector-like leptons coupling to first-, second-, and third-generation SM leptons in both singlet and doublet scenarios. For the doublet model, the VLLs are expected to be excluded at 95% CL up to a mass of 1600 GeV (E_1, N_1), 1630 GeV (E_2, N_2), and 1150 GeV (E_3, N_3). The singlet VLLs are expected to be excluded up to a mass of 600 GeV (E_1), 640 GeV (E_2), and between a mass of 150 and 395 GeV (E_3). The weakest limit is obtained in the vector-like tau leptons, which is expected as the reconstruction and identification efficiency of hadronically decaying taus is the least. Vector-like taus in the singlet model are the hardest signal phase-space to be sensitive to for future experiments, and novel techniques must be exercised to improve the experimental reach to such BSM particles at the electroweak scale.

The need for a faster yet accurate version of the simulation chain is increasing to compete with the large volume of data to be collected at HL-LHC. CMS Fast Simulation is a first principle based simulation pipeline with a few simplified assumptions that make the event simulation faster but with a loss of accuracy to explain data in a wider phase-space. ML-based generative models, such as Variational Auto Encoder, are studied in this thesis in the context of simulating the W+jets process in p-p collisions in some particular phase-space pertinent to the VLL search. The generative model reproduced the shape of the bulk of the distribution quite well with learning the correlation among different event properties, but it is hard to train to model the tail features well. Albeit more complicated networks may be needed, but the computational overhead should be taken into account in the performance metric. Dimensionality reduction techniques, which are very effective in visualizing high-dimensional data in a low-dimensional embedded latent space and unravel how a neural network learns, can come to the rescue. A GAN-like architecture could generate a low-dimensional embedding easily, and the generated latent space could be reconstructed back to the original data space. A more dedicated study on the scalability and possibility of such dimensionality-reduced generative models needs to be carried out.

The work described in this thesis forms part of the following references: [1]* [2] [3] [4]

1. CMS VLL team, Search for vector-like leptons in final states with muon and jets using $\sqrt{s}$ =13 TeV CMS data (Thesis Approved). CMS AN-24-158: CMSAN-24-158 (2025).
2. CMS Collaboration, Review of searches for vector-like quarks, vector-like leptons, and heavy neutral leptons in proton-proton collisions at $\sqrt{s}$=13TeV at the CMS experiment. *Phys. Rept.* **1115,** 570–677. arXiv: 2405.17605 [hep-ex] (2025).
3. CMS Collaboration, Inclusive nonresonant multilepton probes of new phenomena at $\sqrt{s}$=13 TeV. *Phys. Rev. D* **105,** 112007. arXiv: 2202.08676 [hep-ex] (2022).
4. CMS Collaboration, Search for a scalar or pseudoscalar dilepton resonance produced in association with a massive vector boson or top quark-antiquark pair in multilepton events at $\sqrt{s}$=13 TeV. *Phys. Rev. D* **110,** 012013. arXiv: 2402.11098 [hep-ex] (2024).

$*$: *Thesis approved by the collaboration*

The full list of publications as a CMS author can be obtained from iNSPIRE HEP (https://inspirehep.net/authors/1714683).

Bibliography

[1] Fernando Quevedo and Andreas Schachner. "Cambridge Lectures on The Standard Model". In: (). arXiv: 2409.09211 [hep-th].

[2] F. Halzen and Alan D. Martin. *Quarks And Leptons: An Introductory Course In Modern Particle Physics*. 1984. ISBN: 978-0-471-88741-6.

[3] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to quantum field theory*. Reading, USA: Addison-Wesley, 1995. ISBN: 978-0-201-50397-5, 978-0-429-50355-9, 978-0-429-49417-8. DOI: 10.1201/9780429503559.

[4] Y. Fukuda et al. "Evidence for oscillation of atmospheric neutrinos". In: *Phys. Rev. Lett.* 81 (1998), pp. 1562–1567. DOI: 10.1103/PhysRevLett.81.1562. arXiv: hep-ex/9807003.

[5] Q. R. Ahmad et al. "Direct evidence for neutrino flavor transformation from neutral current interactions in the Sudbury Neutrino Observatory". In: *Phys. Rev. Lett.* 89 (2002), p. 011301. DOI: 10.1103/PhysRevLett.89.011301. arXiv: nucl-ex/0204008.

[6] Q. R. Ahmad et al. "Measurement of day and night neutrino energy spectra at SNO and constraints on neutrino mixing parameters". In: *Phys. Rev. Lett.* 89 (2002), p. 011302. DOI: 10.1103/PhysRevLett.89.011302. arXiv: nucl-ex/0204009.

[7] M. N. Rebelo. "On quasidegeneracy of Majorana neutrinos and the observed pattern of Leptonic mixing". In: *17th Lomonosov Conference on Elementary Particle Physics*. 2017, pp. 126–132. DOI: 10.1142/9789813224568_0019. arXiv: 1603.01210 [hep-ph].

[8] Katherine Garrett and Gintaras Duda. "Dark Matter: A Primer". In: *Adv. Astron.* 2011 (2011), p. 968283. DOI: 10.1155/2011/968283. arXiv: 1006.2483 [hep-ph].

[9] T. S. van Albada et al. "The Distribution of Dark Matter in the Spiral Galaxy NGC-3198". In: *Astrophys. J.* 295 (1985), pp. 305–313. DOI: 10.1086/163375.

[10] M. Tanabashi et al. "Review of Particle Physics". In: *Phys. Rev. D* 98.3 (2018), p. 030001. DOI: 10.1103/PhysRevD.98.030001.

[11] CMS Collaboration, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC". In: *Phys. Lett. B* 716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[12] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Phys. Lett. B* 716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[13] Stephen P. Martin. "Extra vector-like matter and the lightest Higgs scalar boson mass in low-energy supersymmetry". In: *Phys. Rev. D* 81 (2010), p. 035004. DOI: 10.1103/PhysRevD.81.035004. arXiv: 0910.2732 [hep-ph].

[14] James Halverson, Nicholas Orlofsky, and Aaron Pierce. "Vectorlike Leptons as the Tip of the Dark Matter Iceberg". In: *Phys. Rev. D* 90.1 (2014), p. 015002. DOI: 10.1103/PhysRevD.90.015002. arXiv: 1403.1592 [hep-ph].

[15] Radovan Dermisek and Aditi Raval. "Explanation of the Muon g-2 Anomaly with Vectorlike Leptons and its Implications for Higgs Decays". In: *Phys. Rev. D* 88 (2013), p. 013017. DOI: 10.1103/PhysRevD.88.013017. arXiv: 1305.3522 [hep-ph].

[16] Motoi Endo et al. "Higgs Mass and Muon Anomalous Magnetic Moment in Supersymmetric Models with Vector-Like Matters". In: *Phys. Rev. D* 84 (2011), p. 075017. DOI: 10.1103/PhysRevD.84.075017. arXiv: 1108.3071 [hep-ph].

[17] Gudrun Hiller et al. "Model Building from Asymptotic Safety with Higgs and Flavor Portals". In: *Phys. Rev. D* 102.9 (2020), p. 095023. DOI: 10.1103/PhysRevD.102.095023. arXiv: 2008.08606 [hep-ph].

[18] Felipe F. Freitas et al. "Phenomenology of vector-like leptons with Deep Learning at the Large Hadron Collider". In: (Oct. 2020). arXiv: 2010.01307 [hep-ph].

[19] Radovan Dermisek, Aditi Raval, and Seodong Shin. "Effects of vectorlike leptons on $h \to 4\ell$ and the connection to the muon g-2 anomaly". In: *Phys. Rev. D* 90.3 (2014), p. 034023. DOI: 10.1103/PhysRevD.90.034023. arXiv: 1406.7018 [hep-ph].

[20] Radovan Dermisek et al. "Limits on Vectorlike Leptons from Searches for Anomalous Production of Multi-Lepton Events". In: *JHEP* 12 (2014), p. 013. DOI: 10.1007/JHEP12(2014)013. arXiv: 1408.3123 [hep-ph].

[21] P. Achard et al. "Search for heavy neutral and charged leptons in e^+e^- annihilation at LEP". In: *Phys. Lett. B* 517 (2001), pp. 75–85. DOI: 10.1016/S0370-2693(01)01005-X. arXiv: hep-ex/0107015.

[22] CMS Collaboration, "Search for vector-like leptons in multilepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Phys. Rev. D* 100 (2019), p. 052003. DOI: 10.1103/PhysRevD.100.052003. arXiv: 1905.10853 [hep-ex].

[23] CMS Collaboration, "Inclusive nonresonant multilepton probes of new phenomena at $\sqrt{s} = 13$ TeV". In: *Phys. Rev. D* 105 (2022), p. 112007. DOI: 10.1103/PhysRevD.105.112007. arXiv: 2202.08676 [hep-ex].

[24] ATLAS Collaboration, "Search for heavy lepton resonances decaying to a Z boson and a lepton in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector". In: *JHEP* 09 (2015), p. 108. DOI: 10.1007/JHEP09(2015)108. arXiv: 1506.01291 [hep-ex].

[25] ATLAS Collaboration, "Search for third-generation vector-like leptons in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *JHEP* 07 (2023), p. 118. DOI: 10.1007/JHEP07(2023)118. arXiv: 2303.05441 [hep-ex].

[26] ATLAS Collaboration, "Search for vector-like leptons coupling to first- and second-generation Standard Model leptons in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: (Nov. 2024). arXiv: 2411.07143 [hep-ex].

[27] "LHC Machine". In: *JINST* 3 (2008). Ed. by Lyndon Evans and Philip Bryant, S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[28] "Simulation of the Silicon Strip Tracker pre-amplifier in early 2016 data". In: (2020). URL: http://cds.cern.ch/record/2740688.

[29] CMS Collaboration, "The CMS Experiment at the CERN LHC". In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[30] CMS Collaboration, "CMS technical design report, volume II: Physics performance". In: *J. Phys. G* 34.6 (2007), pp. 995–1579. DOI: 10.1088/0954-3899/34/6/S01.

[31] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker". In: *JINST* 9.10 (2014), P10009. DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569 [physics.ins-det].

[32] CMS Collaboration, "Strategies and performance of the CMS silicon tracker alignment during LHC Run 2". In: *Nucl. Instrum. Meth. A* 1037 (2022), p. 166795. DOI: 10.1016/j.nima.2022.166795. arXiv: 2111.08757 [physics.ins-det].

[33] CMS Collaboration, "Energy Calibration and Resolution of the CMS Electromagnetic Calorimeter in pp Collisions at $\sqrt{s} = 7$ TeV". In: *JINST* 8 (2013), P09009. DOI: 10.1088/1748-0221/8/09/P09009. arXiv: 1306.2016 [hep-ex].

[34] CMS Collaboration, "Performance of CMS Muon Reconstruction in pp Collision Events at $\sqrt{s} = 7$ TeV". In: *JINST* 7 (2012), P10002. DOI: 10.1088/1748-0221/7/10/P10002. arXiv: 1206.4071 [physics.ins-det].

[35] Abhijith Gandrakota. "Realtime Anomaly Detection at the L1 Trigger of CMS Experiment". In: *PoS* ICHEP2024 (2025), p. 1025. DOI: 10.22323/1.476.1025. arXiv: 2411.19506 [hep-ex].

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

[37] Prudhvi N. Bhattiprolu and Stephen P. Martin. "Prospects for vectorlike leptons at future proton-proton colliders". In: *Phys. Rev. D* 100.1 (2019), p. 015033. DOI: 10.1103/PhysRevD.100.015033. arXiv: 1905.00498 [hep-ph].

[38] Richard D. Ball et al. "Parton distributions for the LHC Run II". In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

[39] Richard D. Ball et al. "Parton distributions from high-precision collider data". In: *Eur. Phys. J. C* 77 (2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428 [hep-ph].

[40] Torbjörn Sjöstrand et al. "An Introduction to PYTHIA 8.2". In: *Comput. Phys. Commun.* 191 (2015), p. 159. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph].

[41] Albert M Sirunyan et al. "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements". Submitted to *Eur. Phys. J. C*. 2019. arXiv: 1903.12179 [hep-ex].

[42] Rikkert Frederix and Stefano Frixione. "Merging meets matching in MC@NLO". In: *JHEP* 12 (2012), p. 061. DOI: 10.1007/JHEP12(2012)061. arXiv: 1209.6215 [hep-ph].

[43] Stefan Hoeche et al. "Matching parton showers and matrix elements". In: *HERA and the LHC: A Workshop on the Implications of HERA for LHC Physics: CERN - DESY Workshop 2004/2005 (Midterm Meeting, CERN, 11-13 October 2004; Final Meeting, DESY, 17-21 January 2005)*. 2005, pp. 288–289. DOI: 10.5170/CERN-2005-014.288. arXiv: hep-ph/0602031.

[44] S. Agostinelli et al. "—a simulation toolkit". In: *Nucl. Instrum. Meth. A* 506 (2003), p. 250. DOI: 10.1016/S0168-9002(03)01368-8.

[45] Natascha Krammer. "The challenges of the HL-LHC for the Full and Fast Simulation of the CMS Experiment". In: *PoS* LHCP2024 (2025), p. 280. DOI: 10.22323/1.478.0280.

[46] Sezen Sekmen. "Recent Developments in CMS Fast Simulation". In: *PoS* ICHEP2016 (2016), p. 181. DOI: 10.22323/1.282.0181. arXiv: 1701.03850 [physics.ins-det].

[47] Rahmat Rahmat, Rob Kroeger, and Andrea Giammanco. "The fast simulation of the CMS experiment". In: *J. Phys. Conf. Ser.* 396 (2012). Ed. by Michael Ernst et al., p. 062016. DOI: 10.1088/1742-6596/396/6/062016.

[48] Andrea Giammanco. "The Fast Simulation of the CMS Experiment". In: *J. Phys. Conf. Ser.* 513 (2014). Ed. by D. L. Groep and D. Bonacorsi, p. 022012. DOI: 10.1088/1742-6596/513/2/022012.

[49] Guenter Grindhammer, M. Rudowicz, and S. Peters. "The Fast Simulation of Electromagnetic and Hadronic Showers". In: *Nucl. Instrum. Meth. A* 290 (1990), p. 469. DOI: 10.1016/0168-9002(90)90566-O.

[50] Rahmat Rahmat and Rob Kroeger. *HF GFlash*. Tech. rep. Geneva: CERN, 2012. DOI: 10.1016/j.phpro.2012.02.385. URL: https://cds.cern.ch/record/1395461.

[51] Samuel Bein et al. "Refining fast simulation using machine learning". In: *EPJ Web Conf.* 295 (2024), p. 09032. DOI: 10.1051/epjconf/202429509032. arXiv: 2309.12919 [physics.ins-det].

[52] CMS Collaboration, *The Phase-2 Upgrade of the CMS Tracker*. Tech. rep. Geneva: CERN, 2017. DOI: 10.17181/CERN.QZ28.FLHW. URL: https://cds.cern.ch/record/2272264.

[53] CMS Collaboration, *Technical proposal for the Phase-II upgrade of the Compact Muon Solenoid*. CMS Technical Proposal CERN-LHCC-2015-010, CMS-TDR-15-02. 2015. URL: https://cds.cern.ch/record/2020886.

[54] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det].

[55] R. Fruhwirth, W. Waltenberger, and P. Vanlaer. "Adaptive vertex fitting". In: *J. Phys. G* 34 (2007), N343. DOI: 10.1088/0954-3899/34/12/N01.

[56] CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JINST* 13.06 (2018), P06015. DOI: 10.1088/1748-0221/13/06/P06015. arXiv: 1804.04528 [physics.ins-det].

[57] CMS Collaboration, "Performance of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s}$ = 13 TeV". In: *JINST* 19.09 (2024), P09004. DOI: 10.1088/1748-0221/19/09/P09004. arXiv: 2403.15518 [physics.ins-det].

[58] CMS Collaboration, "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC". In: *JINST* 16.05 (2021), P05014. DOI: 10.1088/1748-0221/16/05/P05014. arXiv: 2012.06888 [hep-ex].

[59] Gavin P. Salam and Gregory Soyez. "A Practical Seedless Infrared-Safe Cone jet algorithm". In: *JHEP* 05 (2007), p. 086. DOI: 10.1088/1126-6708/2007/05/086. arXiv: 0704.0292 [hep-ph].

[60] Yuri L. Dokshitzer et al. "Better jet clustering algorithms". In: *JHEP* 08 (1997), p. 001. DOI: 10.1088/1126-6708/1997/08/001. arXiv: hep-ph/9707323.

[61] S. Catani et al. "Longitudinally invariant K_t clustering algorithms for hadron hadron collisions". In: *Nucl. Phys. B* 406 (1993), pp. 187–224. DOI: 10.1016/0550-3213(93)90166-M.

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

[63] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *JINST* 12.02 (2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014. arXiv: 1607.03663 [hep-ex].

[64] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "FastJet User Manual". In: *Eur. Phys. J. C* 72 (2012), p. 1896. DOI: 10.1140/epjc/s10052-012-1896-2. arXiv: 1111.6097 [hep-ph].

[65] The 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.

[66] CMS Collaboration, *Pileup Removal Algorithms*. CMS Physics Analysis Summary CMS-PAS-JME-14-001. 2014. URL: http://cds.cern.ch/record/1751454.

[67] A. Perloff. "Pileup measurement and mitigation techniques in CMS". In: *J. Phys. Conf. Ser.* 404 (2012). Ed. by Nural Akchurin, p. 012045. DOI: 10.1088/1742-6596/404/1/012045.

[68] CMS Collaboration, "Jet energy scale and resolution measurement with Run 2 Legacy Data Collected by CMS at 13 TeV". In: (2021). URL: http://cds.cern.ch/record/2792322.

[69] CMS Collaboration, "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector". In: *JINST* 14.07 (2019), P07004. DOI: 10.1088/1748-0221/14/07/P07004. arXiv: 1903.06078 [hep-ex].

[70] CMS Collaboration, "Mitigation of anomalous missing transverse momentum measurements in data collected by CMS at sqrt(s) = 13 TeV during the LHC Run 2". In: (2020). URL: https://cds.cern.ch/record/2714938.

[71] CMS Collaboration, "Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV". In: *JINST* 13.10 (2018), P10005. DOI: 10.1088/1748-0221/13/10/P10005. arXiv: 1809.02816 [hep-ex].

[72] CMS Collaboration, "Identification of hadronic tau lepton decays using a deep neural network". In: *JINST* 17 (2022), P07023. DOI: 10.1088/1748-0221/17/07/P07023. arXiv: 2201.08458 [hep-ex].

[73] Emil Bols et al. "Jet Flavour Classification Using DeepJet". In: *JINST* 15.12 (2020), P12012. DOI: 10.1088/1748-0221/15/12/P12012. arXiv: 2008.10519 [hep-ex].

[74] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *JINST* 13 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158 [physics.ins-det].

[75] CMS Collaboration, "Performance summary of AK4 jet b tagging with data from proton-proton collisions at 13 TeV with the CMS detector". In: (2023). URL: https://cds.cern.ch/record/2854609.

[76] François Chollet. *Keras*. https://github.com/fchollet/keras. 2015.

[77] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[78] CMS Collaboration, "Measurement of the differential cross sections for the associated production of a W boson and jets in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Phys. Rev. D* 96.7 (2017), p. 072005. DOI: 10.1103/PhysRevD.96.072005. arXiv: 1707.05979 [hep-ex].

[79] CMS Collaboration, "Measurements of differential cross sections for associated production of a W boson and jets in proton-proton collisions at $\sqrt{s} = 8$ TeV". In: *Phys. Rev. D* 95 (2017), p. 052002. DOI: 10.1103/PhysRevD.95.052002. arXiv: 1610.04222 [hep-ex].

[80] CMS Collaboration, "Measurement of differential cross sections for the production of a Z boson in association with jets in proton-proton collisions at $\sqrt{s}$ = 13 TeV". In: *Phys. Rev. D* 108 (2023), p. 052004. DOI: 10.1103/PhysRevD.108.052004. arXiv: 2205.02872 [hep-ex].

[81] HEP ML Community. *A Living Review of Machine Learning for Particle Physics*. URL: https://iml-wg.github.io/HEPML-LivingReview/.

[82] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *Eur. Phys. J. C* 71 (2011). [Erratum: 10.1140/epjc/s10052-013-2501-z], p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0. arXiv: 1007.1727 [physics.data-an].

[83] Luca Lista. "Practical Statistics for Particle Physicists". In: *2016 European School of High-Energy Physics*. 2017, pp. 213–258. DOI: 10.23730/CYRSP-2017-005.213. arXiv: 1609.04150 [physics.data-an].

[84] R. L. Workman et al. "Review of Particle Physics". In: *PTEP* 2022 (2022), p. 083C01. DOI: 10.1093/ptep/ptac097.

[85] Eilam Gross and Ofer Vitells. "Trial factors for the look elsewhere effect in high energy physics". In: *Eur. Phys. J. C* 70 (2010), p. 525. DOI: 10.1140/epjc/s10052-010-1470-8. arXiv: 1005.1891 [physics.data-an].

[86] Alexander L. Read. "Presentation of search results: The CL_s technique". In: *J. Phys. G* 28 (2002), p. 2693. DOI: 10.1088/0954-3899/28/10/313.

[87] CMS Collaboration, "The CMS Statistical Analysis and Combination Tool: Combine". In: *Comput. Softw. Big Sci.* 8.1 (2024), p. 19. DOI: 10.1007/s41781-024-00121-4. arXiv: 2404.06614 [physics.data-an].

[88] Gabriella PÃąsztor. "The Phase-2 Upgrade of the CMS Detector". In: *PoS* LHCP2022 (2023), p. 045. DOI: 10.22323/1.422.0045. URL: https://cds.cern.ch/record/2880161.

[89] X. Cid Vidal and other. *Report from working group 3: Beyond the standard model physics at the HL-LHC and HE-LHC*. CERN Report CERN-LPCC-2018-05. 2019. DOI: 10.23731/CYRM-2019-007.585. arXiv: 1812.07831 [hep-ph].

[90] Oliver Brüning and Lucio Rossi, eds. *The High Luminosity Large Hadron Collider*. World Scientific, Mar. 2024. ISBN: 978-981-12-7894-5. DOI: 10.1142/13487.

[91] CMS Collaboration, *The Phase-2 upgrade of the CMS trigger*. CMS Technical Proposal CERN-LHCC-2020-004, CMS-TDR-021. 2020. URL: https://cds.cern.ch/record/2714892.

[92] CMS Collaboration, *The Phase-2 upgrade of the CMS data acquisition and high level trigger*. CMS Technical Proposal CERN-LHCC-2021-007, CMS-TDR-022. 2021. URL: https://cds.cern.ch/record/2759072.

[93] CMS Offline Software and Computing. *CMS Phase-2 Computing Model: Update Document*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2815292.

[94] Kevin Nash. "The Phase-2 upgrade of the CMS outer tracker". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1058 (2024), p. 168788. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2023.168788. URL: https://www.sciencedirect.com/science/article/pii/S0168900223007799.

[95] CMS Collaboration, *Expected performance of the physics objects with the upgraded CMS detector at the HL-LHC*. CMS Note CMS-NOTE-2018-006. 2018. URL: https://cds.cern.ch/record/2650976.

[96] CMS Collaboration, *The Phase-2 upgrade of the CMS endcap calorimeter*. CMS Technical Proposal CERN-LHCC-2017-023, CMS-TDR-019. 2017. URL: https://cds.cern.ch/record/2293646.

[97] CMS Collaboration, *The Phase-2 upgrade of the CMS barrel calorimeters*. CMS Technical Proposal CERN-LHCC-2017-011, CMS-TDR-015. 2017. URL: https://cds.cern.ch/record/2283187.

[98] CMS Collaboration, *The Phase-2 upgrade of the CMS muon detectors*. CMS Technical Proposal CERN-LHCC-2017-012, CMS-TDR-016. 2017. URL: https://cds.cern.ch/record/2283189.

[99] CMS Collaboration, *A MIP timing detector for the CMS Phase-2 upgrade*. CMS Technical Proposal CERN-LHCC-2019-003, CMS-TDR-020. 2019. URL: https://cds.cern.ch/record/2667167.

[100] CMS Collaboration, "Review of searches for vector-like quarks, vector-like leptons, and heavy neutral leptons in proton–proton collisions at s=13TeV at the CMS experiment". In: *Phys. Rept.* 1115 (2025), pp. 570–677. DOI: 10.1016/j.physrep.2024.09.012. arXiv: 2405.17605 [hep-ex].

[101] Georgia Karagiorgi et al. "Machine Learning in the Search for New Fundamental Physics". In: (Dec. 2021). arXiv: 2112.03769 [hep-ph].

[102] Matthew Feickert and Benjamin Nachman. "A Living Review of Machine Learning for Particle Physics". In: (Feb. 2021). arXiv: 2102.02770 [hep-ph].

[103] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML]. URL: https://arxiv.org/abs/1802.03426.

[104] Tom Marianer, Dovi Poznanski, and J Xavier Prochaska. "A semisupervised machine learning search for never-seen gravitational-wave sources". In: *Monthly Notices of the Royal Astronomical Society* 500.4 (Nov. 2020), 5408âĂŞ5419. ISSN: 1365-2966. DOI: 10.1093/mnras/staa3550. URL: http://dx.doi.org/10.1093/mnras/staa3550.

[105] En-Jui Kuo and Hossein Dehghani. "Unsupervised learning of interacting topological and symmetry-breaking phase transitions". In: *Phys. Rev. B* 105.23 (2022), p. 235136. DOI: 10.1103/PhysRevB.105.235136. arXiv: 2111.08747 [cond-mat.str-el].

[106] Shreyasi Acharya and Subhasis Chattopadhyay. "Estimation of initial-state structures in high-energy heavy-ion collisions using principal component analysis". In: *Phys. Rev. C* 103.3 (2021), p. 034909. DOI: 10.1103/PhysRevC.103.034909. arXiv: 2103.12380 [nucl-th].

[107] Ziming Liu, Wenbin Zhao, and Huichao Song. "Principal Component Analysis of collective flow in Relativistic Heavy-Ion Collisions". In: *Eur. Phys. J. C* 79.10 (2019), p. 870. DOI: 10.1140/epjc/s10052-019-7379-y. arXiv: 1903.09833 [nucl-th].

[108] Florencia Canelli et al. "Autoencoders for semivisible jet detection". In: *JHEP* 02 (2022), p. 074. DOI: 10.1007/JHEP02(2022)074. arXiv: 2112.02864 [hep-ph].

[109] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[110] Diederik P. Kingma and Max Welling. "An Introduction to Variational Autoencoders". In: *Foundations and TrendsÂő in Machine Learning* 12.4 (2019), 307âĂŞ392. ISSN: 1935-8245. DOI: 10.1561/2200000056. URL: http://dx.doi.org/10.1561/2200000056.

[111] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

Chapter 11

Appendix

11.1 Trigger, b-tagging, and custom lepton identification efficiency measurements

11.1.1 Single muon trigger efficiency

Single muon trigger efficiency measurements are conducted using the tag and probe method in $\mathrm{Z} \to \mu\mu$ enriched data and DY MC samples. Di-muon events are selected with a single muon trigger. The tag object must pass our muon object selection criteria described in Section 4.5 and match a HLT-level object. Probe object also passes the analysis muon object definition, is opposite in charge with the Tag and lies outside the cone of radius ΔR = 0.4 centered around the Tag object. Probe matching to another HLT-level object gives the desired trigger efficiency. For matching purposes, $\Delta R < 0.2$ is used. Figure 11.1 shows the single muon trigger efficiency as a function of probe muon p_{T} in the barrel and endcap regions of the detector.

11.1.2 Muon custom identification efficiency

The choice of impact parameter cuts ($\mathrm{d_{xy}}, \mathrm{d_z}$), SIP3D, and LeptonDeepJet score has different efficiency in data and MC samples used in this analysis. Figure 11.2 shows the efficiency of all these choices in data and MC as a function of muon p_{T} and η. As can be seen, the scale factors are close to unity and, thus, not applied to correct the simulation in this analysis.

11.1.3 MC b-tagging efficiency of the DEEPJET tagger

B-tagging efficiency, mistagging efficiency for light and c jets measured in MC for DEEPJET medium WP used in this analysis to derive the b-tagging scale factors are demonstrated in Figure 11.3–11.4.

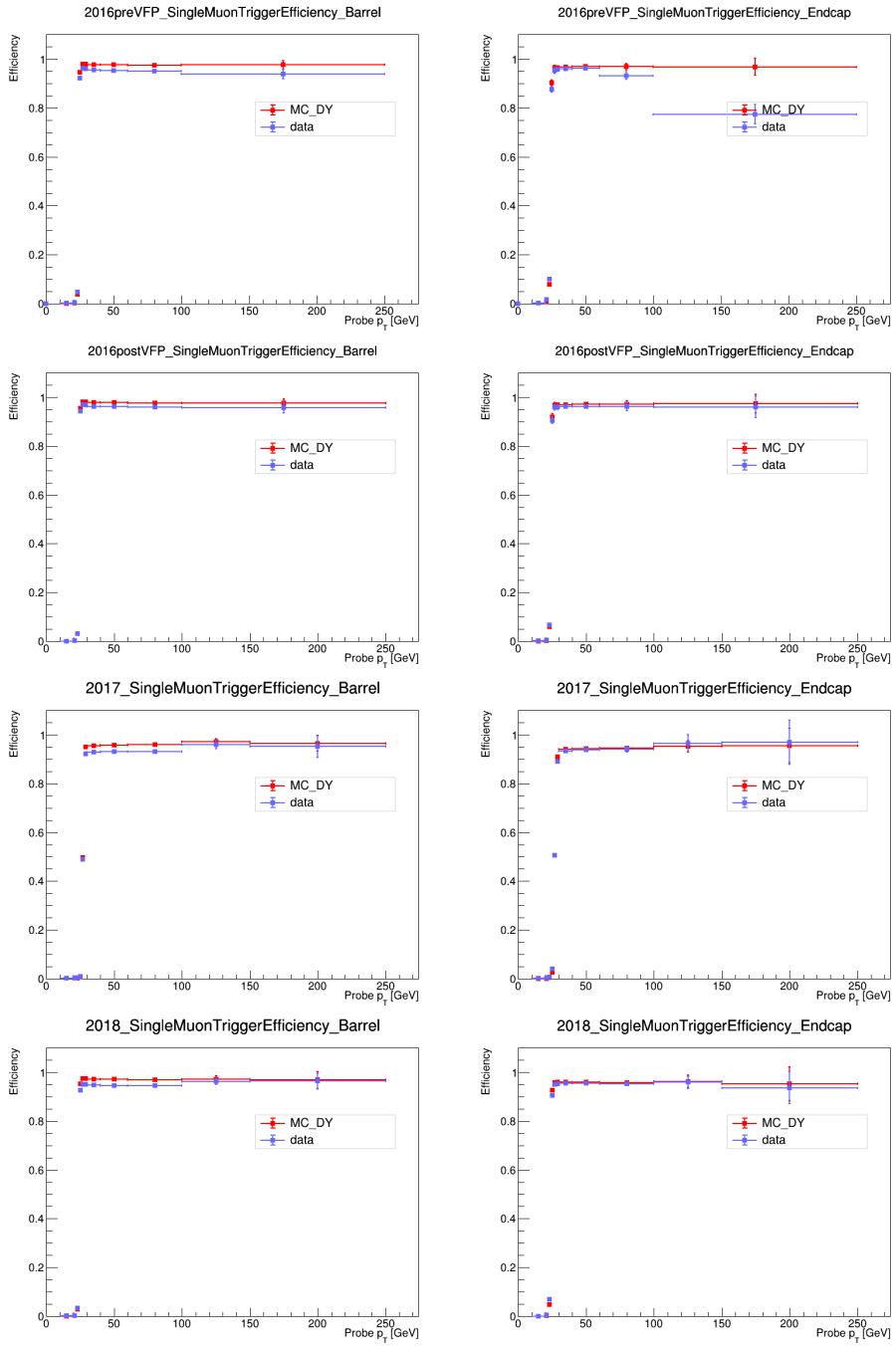


FIGURE 11.1: Single isolated muon trigger efficiencies for barrel (left) and endcap (right) in 2016preVFP (first row), 2016postVFP (second row), 2017 (third row), and 2018 (fourth row) as measured by the tag-and-probe method in $Z \to \mu\mu$ enriched data and DY MC samples. Finally, we fit a function in each year separately for data and MC.

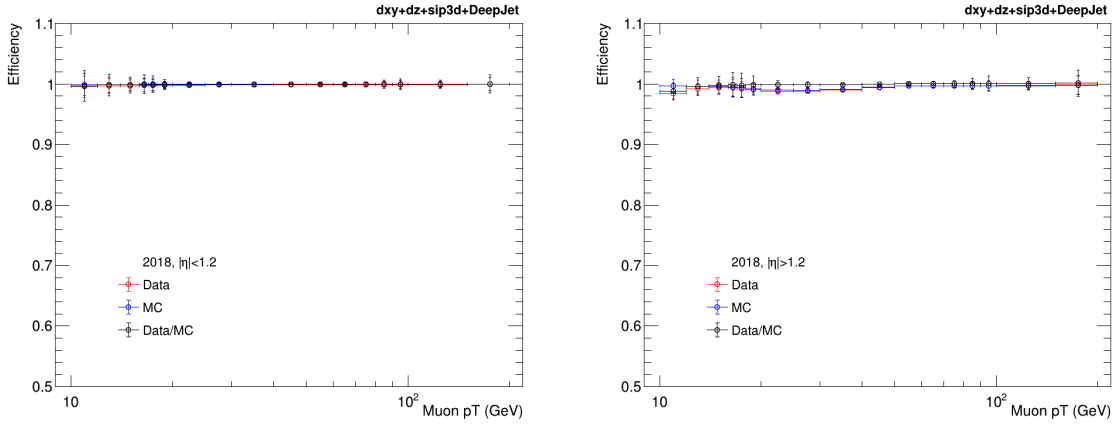


FIGURE 11.2: Custom muon ID efficiencies and efficiency scale factors (data/MC) in 2018 for barrel (left) and right (endcap) as measured by the tag-and-probe method in Z $\rightarrow \mu\mu$ enriched OSSF 2L OnZ events in data and DY MC samples. The custom ID requirements refer to the d_{xy}, d_z, SIP 3D, and DEEPJET 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.

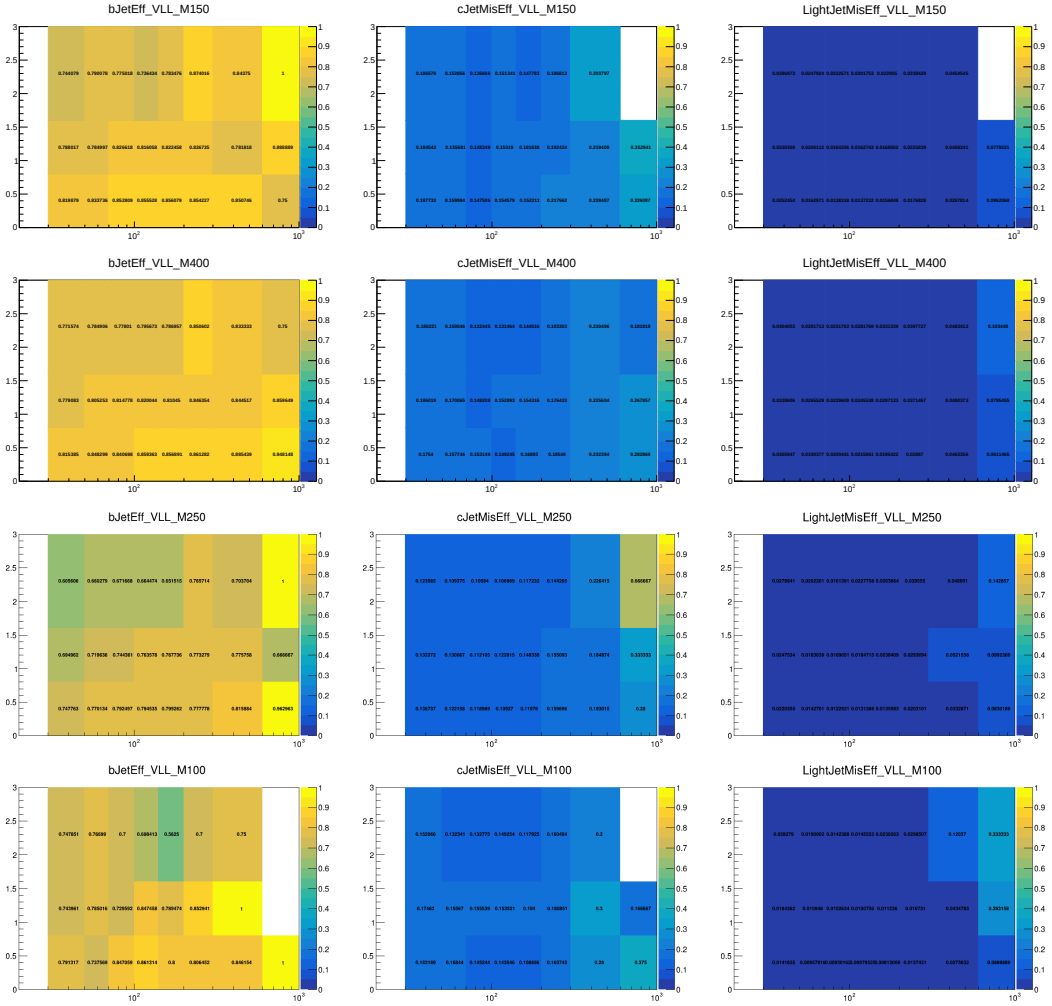


FIGURE 11.3: DEEPJET medium WP b-tagging efficiency for b-jets(left) and mistagging efficiency for c-jets (middle), and light-jets(right) in a few representative signal samples for Mass 150 in 2018 (first row), Mass 400 in 2017 (second row), Mass 250 in 2016preVFP (third row), and Mass 100 in 2016postVFP (fourth row).

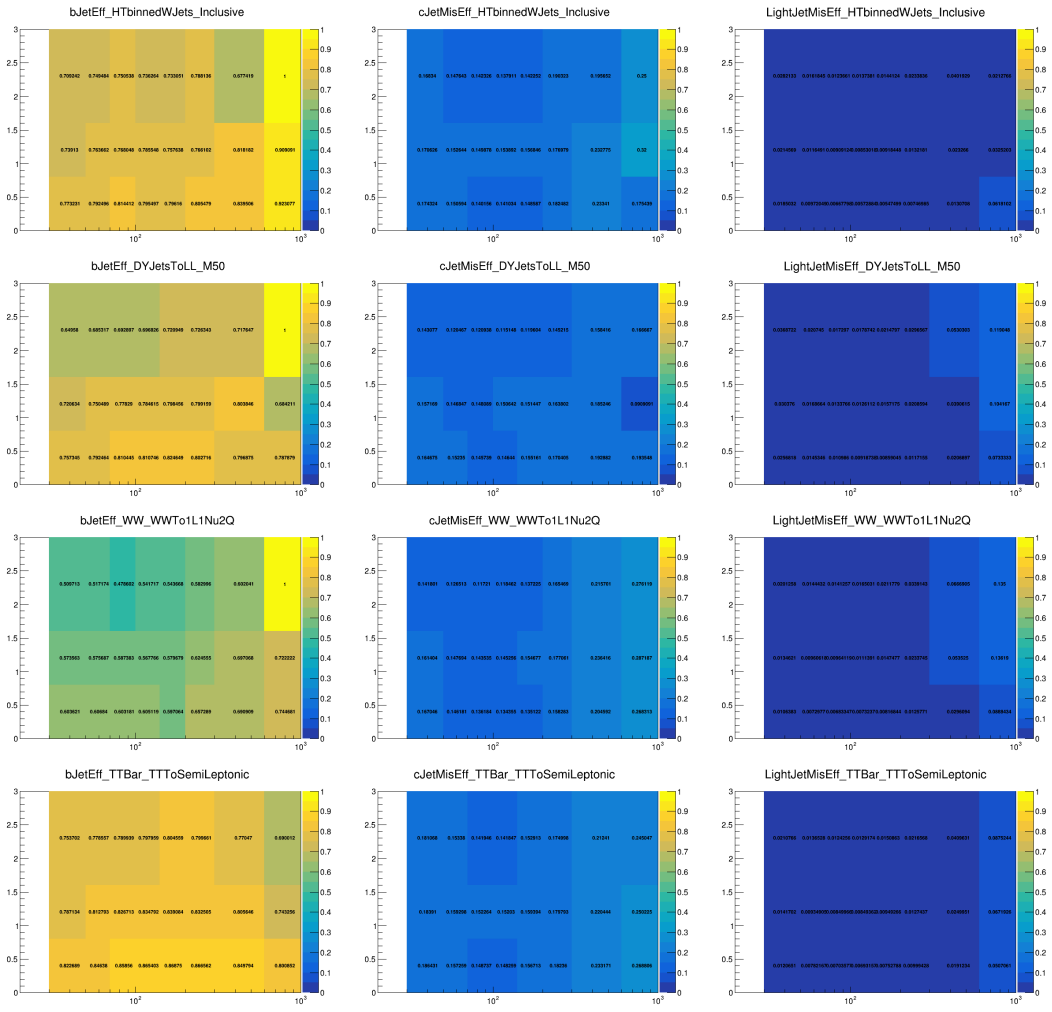


FIGURE 11.4: DEEPJET medium WP b-tagging efficiency for b-jets(left) and mistagging efficiency for c-jets (middle), and light-jets(right) in a few representative background samples for w+jets in 2018 (first row), z+jets in 2017 (second row), diboson (WW) in 2016preVFP (third row), and $t\bar{t}$+jets in 2016postVFP (fourth row).

11.2 W+jets study with NLO W-p_{T} binned samples

This study was performed to check the feasibility of using NLO samples to estimate the W+jets background in the analysis phase space. An inclusive WJets NLO sample covering the full W p_{T} phase space is prepared from the following W p_{T} binned samples.

- WJets NLO Inclusive: Inclusive in W-boson p_{T}.
- WJets NLO WPt_100-250: W-boson p_{T} between 100 and 250 GeV at LHE level.
- WJets NLO WPt_250-400: W-boson p_{T} between 250 and 400 GeV at LHE level.
- WJets NLO WPt_400-600: W-boson p_{T} between 400 and 600 GeV at LHE level.
- WJets NLO WPt_600-Inf: W-boson p_{T} with 600 GeV and above at LHE level.

The WJets NLO inclusive sample overlaps with the explicit p_{T}-binned samples. We use the LHE_VPt (vector boson p_{T} at generator level) to remove the overlap. Events from the inclusive sample are only selected if $\mathrm{LHE\_VPt} < 100$ GeV in the event. Thus, the inclusive sample is made orthogonal to other explicit p_{T}-binned samples and can be used together. Events with $\mathrm{LHE\_VPt} > 100$ GeV in the inclusive sample are much less than those of exclusive p_{T}-binned samples, thus not selected to simplify the stitching procedure without weighting the overlap events with relevant cross-sections. It is worth noting that this simplified stitching procedure has a negligible impact on the available statistics of stitched WJets NLO samples.

Figure 11.5 and 11.6 demonstrate the data mc agreement in a few key kinematic and event variables used in this analysis in the W+jets control region and validation region, respectively. We observed that the NLO samples are better at predicting the $p_{\mathrm{T}}^{\mathrm{miss}}$, Muon M_T, and HT shape compared to the LO samples (without applying NJets-HT and Muon p_{T} or M_T based corrections), and also the normalization factor is closer to 1, as expected. However, the high jet multiplicity region in jet multiplicity and low $\Delta R(j_0, j_1)$ are poorly modeled in NLO. LO W+jets samples are found to be better for modeling these variables. Muon M_{T} has modeling issues at high M_{T} region, because the off-shell W contribution in high M_{T}muon phase space is missing from these p_{T} binned NLO samples, which is crucial for this analysis.

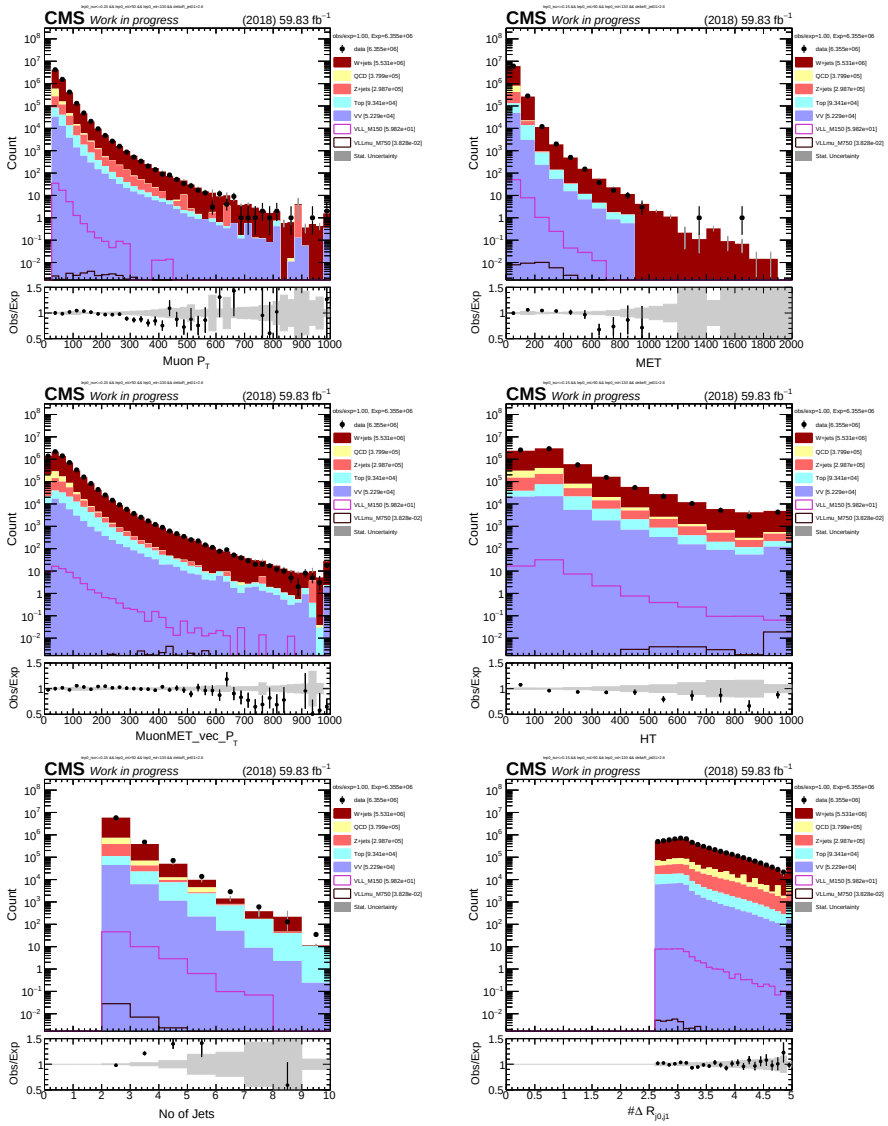


FIGURE 11.5: Data-mc agreement in the W+jets control region for 2018 using W+jets NLO W-p_{T} binned samples. Statistical uncertainties only.

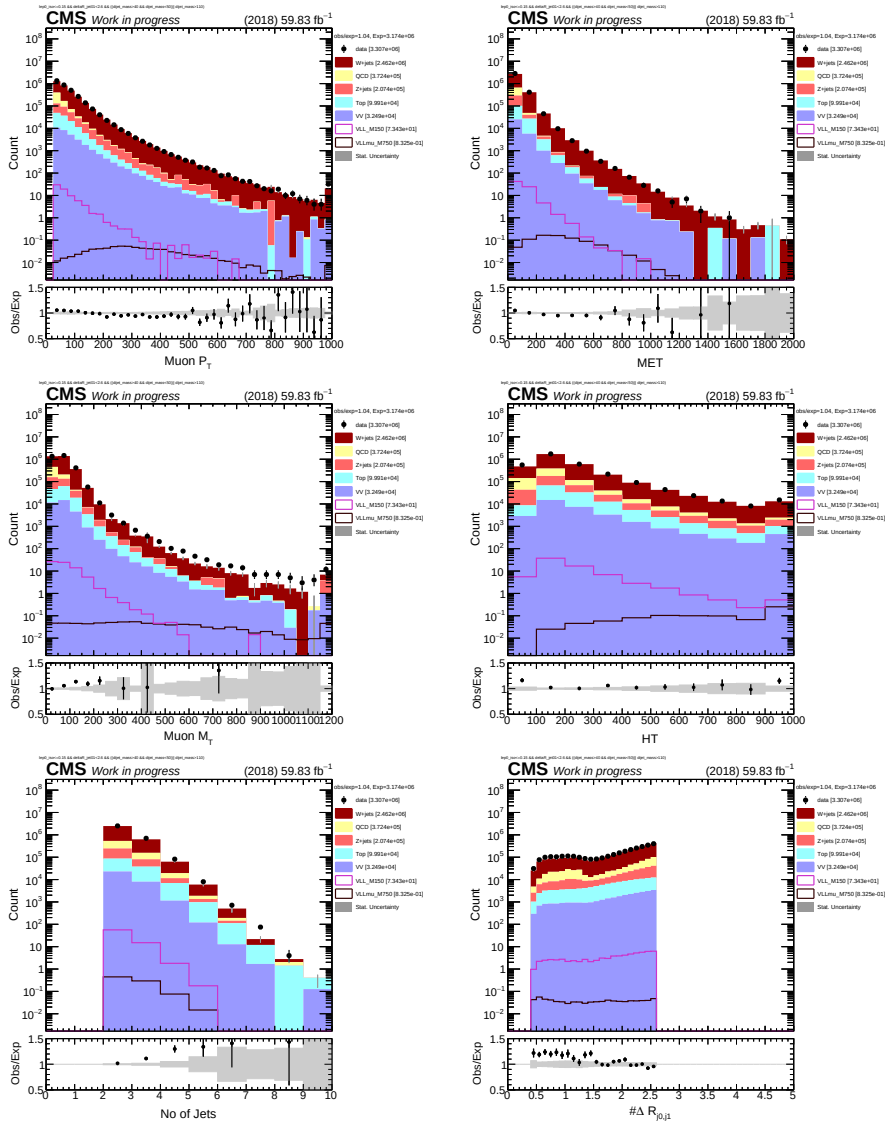


FIGURE 11.6: Data-mc agreement in the W+jets validation region for 2018 using W+jets NLO W-p_{T} binned samples. Statistical uncertainties only.