Search for single vector-like quark production in the opposite-sign dilepton final state in proton-proton collisions at $\sqrt{s} = 13$ TeV

Dissertation

zur Erlangung des Doktorgrades

an der Fakultät für Mathematik,

Informatik und Naturwissenschaften

Fachbereich Physik

der Universität Hamburg

vorgelegt von

Di Wang

aus Beijing, China

Hamburg

2026

Gutachter/innen der Dissertation:	Dr. Rainer Mankel
	Prof. Dr. Elisabetta Gallo-Voss
Zusammensetzung der Prüfungskommission:	Prof. Dr. Gregor Kasieczka
	Prof. Dr. Kerstin Borras
	Prof. Dr. Geraldine Servant
	Dr. Rainer Mankel
	Prof. Dr. Elisabetta Gallo-Voss
Vorsitzende/r der Prüfungskommission:	Prof. Dr. Gregor Kasieczka
Datum der Disputation:	23.02.2026
Vorsitzender des Fach-Promotionsausschusses PHYSIK:	Prof. Dr. Johannes Haller
Leiter des Fachbereichs PHYSIK:	Prof. Dr. Markus Drescher
Dekan der Fakultät MIN:	Prof. Dr.-Ing. Norbert Ritter

Declaration on oath

I hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated.

If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.

Date 29.01.2026 **Signature of doctoral candidate** Di Wang

Abstract

Vector-like quarks are predicted in various scenarios in beyond the Standard Model (SM) theories. This work presents a search for a single vector-like top quark (T) decaying into a top quark and a SM Higgs boson (T→tH), in the T mass range from 600 GeV to 1200 GeV. The data used is from proton-proton collisions collected by the CMS experiment at the LHC in Run 2, corresponding to 138 fb^{-1} of integrated luminosity. The final state includes two opposite-sign leptons (muons or electrons), jets, and missing transverse momentum. The analysis uses cut-based selection criteria optimized for SM background suppression, along with a mass reconstruction algorithm developed under the neutrino kinematic assumptions. For signal extraction, the main observable is the reconstructed T mass, and the background models are data-driven. No excess in data above the standard model prediction is observed, thus upper limits are set on the production cross section of vector-like top quark times branching ratio. The analysis presents the first results for the opposite-sign dilepton final state, achieving a sensitivity similar to that of other channels with the same target.

Zusammenfassung

Vektorartige Quarks werden in verschiedenen Szenarien von Theorien jenseits des Standardmodells (SM) vorhergesagt. Diese Arbeit präsentiert eine Suche nach einem einzeln produzierten vektorartigen Top-Quark (T), das in ein Top-Quark und ein Higgs-Boson des SM zerfällt (T→tH), im T-Massenbereich von 600 GeV bis 1200 GeV. Die verwendeten Daten stammen aus Proton-Proton-Kollisionen, die vom CMS-Experiment am LHC während Run 2 aufgezeichnet wurden, und entsprechen einer integrierten Luminosität von 138 fb^{-1}. Der Endzustand umfasst zwei Leptonen, Myonen oder Elektronen, mit entgegengesetzter Ladung, Jets sowie fehlenden Transversalimpuls. Die Analyse verwendet schnittbasierte Selektionskriterien, die zur Unterdrückung der SM-Hintergründe optimiert sind, zusammen mit einem Algorithmus zur Massenrekonstruktion, der unter geeigneten Approximationen zur Neutrino-Kinematik entwickelt wurde. Zur Signalextraktion dient als Hauptobservable die rekonstruierte T-Masse, während der Hintergrund aus den Daten bestimmt wird. Es wird kein Überschuss der Daten gegenüber der Vorhersage des Standardmodells beobachtet; folglich werden obere Grenzen für den Produktionswirkungsquerschnitt des vektorartigen Top-Quarks multipliziert mit dem Verzweigungsverhältnis festgelegt. Die Analyse präsentiert die ersten Ergebnisse für den Endzustand mit einem Lepton-Paar entgegengesetzter Ladung und erreicht eine Empfindlichkeit, die mit der anderer Kanäle mit demselben Ziel vergleichbar ist.

Contents

CHAPTER 1

Introduction

> ”It is a miracle that curiosity survives formal education. ”
>
> Albert Einstein

Experimental discoveries usually lead to theoretical improvements, and theoretical predictions always inspire experimental developments. Since 1970, or even earlier, remarkable progress in particle physics has been made in both theory and experiment.

The Standard Model (SM) of particle physics is a very successful theory that has been validated by many experiments. The theory combines special relativity and quantum field theory, introducing seventeen elementary particles and their interactions.

Particle colliders are powerful tools for testing the SM and searching for new physics, leading to enormous important discoveries in particle physics. For instance, the gluon was observed at the electron-positron collider at DESY in 1979, and the W and Z bosons were observed at the proton-antiproton collider at CERN in 1983.

In 2012, the CMS and ATLAS experiments at the Large Hadron Collider (LHC) observed a neutral scalar boson at 125 GeV, which agrees with the Higgs boson prediction in the SM [1, 2]. The Higgs boson discovery was a major milestone, as all the particles predicted by the SM were found. The properties of the Higgs boson are measured after its discovery, and the results agree with SM predictions so far [3].

The SM is thus a great success, although many remaining questions can not be explained by it. For example, gravity and dark matter are not included in the SM, and the lightness of the Higgs boson is not explained. Various physics models beyond the SM are proposed to address these open questions, such as the little Higgs and composite Higgs models [4,5]. Many BSM models predict a new vector-like quark (VLQ) model, providing an explanation for the Higgs boson’s observed lightness and supporting the idea that the Higgs boson is a pseudo-Goldstone boson that introduces electroweak breaking [6–8].

The hypothetical VLQs typically include four types, namely T, B, X, and Y, and can be produced in pairs through the strong interaction or singly via the electroweak interaction [9]. Pair production dominates the VLQ mass region below 1 TeV and has a cross section that rapidly decreases with increasing T mass and is independent of the VLQ flavor. The single production cross section is constrained by the coupling strength between the VLQ and the SM particles and has a sizable contribution in the high VLQ mass region. The VLQs were extensively searched by CMS and ATLAS experiments at the LHC, and no evidence of their discovery has been found to date [10–12]. Upper limits are set on the VLQ production cross sections and on the coupling parameters under theoretical assumptions. To further increase the sensitivity of the VLQ searches and improve understanding of the VLQ models, more VLQ decay modes are being explored for future combinations.

The main topic of this thesis is to search for a singly produced vector-like top quark (T) using data collected by the CMS experiment in the LHC Run 2. The analysis focuses on the T → tH channel in the opposite-sign dilepton final state, a T decay mode first searched for in CMS. Chapter 2 introduces the theoretical framework, the SM of particle physics and the vector-like quark model. Chapter 3 introduces the experimental setup, including the LHC and the CMS detector, and the physics objects used by the physics analysis. The analysis strategy is presented in Chapter 4, introducing the cut-based selections and the signal reconstruction, followed by the signal and background modeling. The results are presented in Chapter 5, and are based on the statistical model described in Chapter 4, showing no evidence for the VLQ signal and thus presenting upper limit results.

The analysis is very close to publication as a CMS paper, and the results are ready to be presented at a conference. As the contact person of the analysis, I designed the analysis strategy, developed the software framework, and wrote the documents, including the analysis note and the paper draft. I also delivered the pre-approval and approval presentations to the CMS Beyond-Two-Generations (B2G) subgroup and gave a parallel talk titled ”Searches for vector-like quarks and leptons at CMS” at the EPS-HEP conference in Marseille in July 2025 on behalf of the CMS collaboration.

CHAPTER 2

Theoretical framework

2.1 Standard Model Physics of Particle Physics

The Standard Model (SM) of particle physics introduces elementary particles and describe their interactions. Figure 2.1 illustrates the elementary SM particles, which can be roughly divided into the matter particles, the force-carrier particles, and the Higgs boson.

The matter particles are fermions with spin 1/2, which can be categorized into leptons and quarks; each type consists of six particles in three generations. The first generation has the lightest and most stable particles, while the second and third generations have heavier, less stable particles. The first, second, and third generations of leptons are electrons (e) and electron neutrinos (ν_e), muons (μ) and muon neutrinos (ν_μ), and taus (τ) and tau neutrinos (ν_τ), respectively. Similarly, the three generations of quarks are: up (u) and down (d) quarks, charm (c) and strange (s) quarks, and top (t) and bottom (b) quarks.

There are four types of fundamental interactions in nature: strong, electromagnetic, weak, and gravitational. Each of the four interactions is described by a quantum field theory (QFT), and the first three interactions result from the exchange of force-carrier particles with spin 1, known as mediators. The strong force is described by quantum chromodynamics, and is carried by gluons (g). The electromagnetic force arises from the exchange of photons (γ) and can be described by quantum electrodynamics. The weak force is carried by W bosons (W^+/W^-) and Z bosons (Z), and can be unified into the electroweak theory with the electromagnetic force. The gravitational force is described with relativistic geometrodynamics, with no corresponding mediator particle found to date.

The Higgs boson is the only element SM particle with spin 0, providing mass to other massive SM particles through the Higgs field via the Brout-Englert-Higgs mechanism [13–16]. The Higgs boson was discovered by the CMS and ATLAS collaborations at the LHC in 2012 [1, 2].

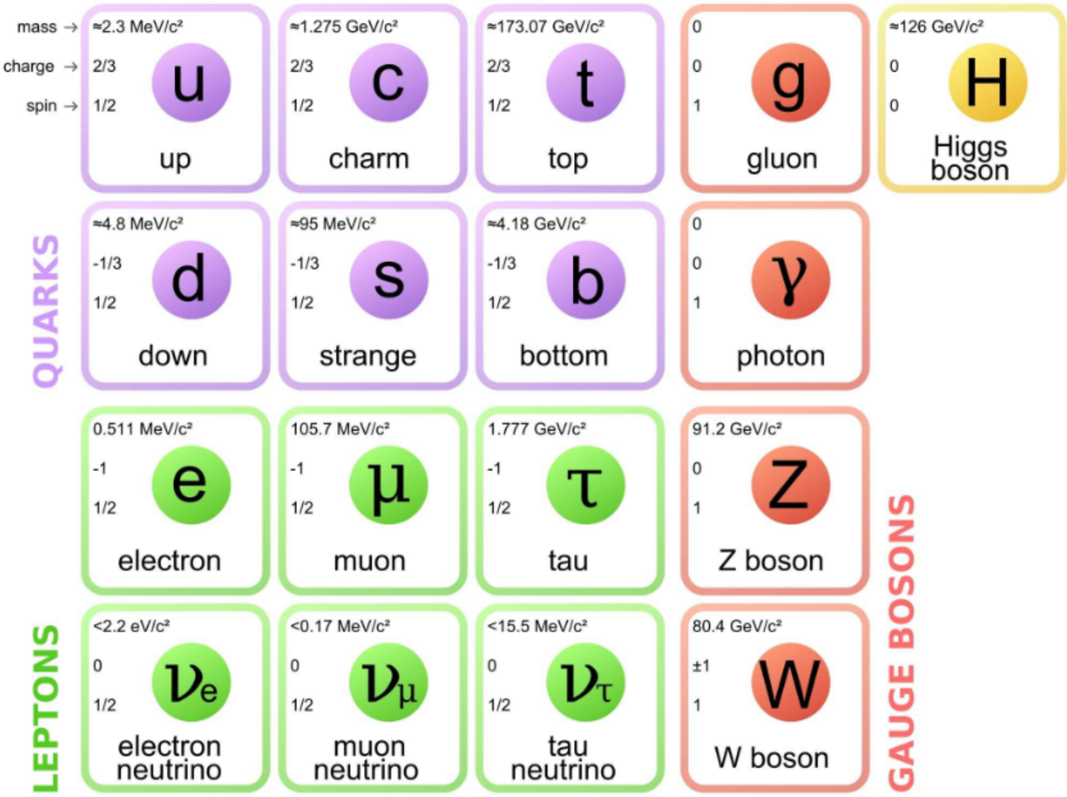

Figure 2.1: Summary of SM particles [17].

2.1.1 Lagrangian and Symmetries

The theory to describe the SM physics is the relativistic quantum field theory (QFT), based on the Lagrangian density ($\mathscr{L}$) as a function of the particle fields ($\Phi(x)$) and their derivatives ($\partial_\mu \Phi(x)$), where the x is a four-vector in the space-time coordinate (ct, x, y, z), and μ is the index of the four vector. In a physics system, the Lagrangian density contains all the information, and the action is defined by the integral of $\mathscr{L}$ over space-time:

$$S = \int \mathscr{L} d^4x \tag{2.1}$$

The motion equations can be derived from the principle of least action by minimizing the action function below.

$$\delta S = \int d^4x \sum_i (\frac{\partial \mathscr{L}}{\partial \Phi_i}\delta\Phi_i + \frac{\partial \mathscr{L}}{\partial(\partial_\mu \Phi_i)}\delta(\partial_\mu \Phi_i)) = \int d^4x \sum_i (\frac{\partial \mathscr{L}}{\partial \Phi_i} - \partial_\mu \frac{\partial \mathscr{L}}{\partial(\partial_\mu \Phi_i)})\delta\Phi_i = 0 \tag{2.2}$$

where S is the action, and $\delta\Phi_i$ is a small variation of the field. The first term can be transferred to the second term by integrating by parts, assuming the fields vanish on the boundary. Thus, the equation of motion, known as the Euler-Lagrange equation, can be obtained as shown in Equation 2.3.

$$\frac{\partial \mathscr{L}}{\partial \Phi_i} - \partial_\mu \frac{\partial \mathscr{L}}{\partial(\partial_\mu \Phi_i)} = 0 \tag{2.3}$$

Given a Lagrangian density for a physics system, the corresponding equation of motion can be obtained via Equation 2.3. For example, the Klein-Gordon equation 2.5 for a spin-0 particle can be obtained from the Lagrangian density in Equation 2.4.

$$\mathscr{L} = \frac{1}{2}(\partial_\mu \phi)(\partial^\mu \phi) - \frac{1}{2}m^2\Phi^2 \tag{2.4}$$

$$\partial_\mu \partial^\mu \Phi + m^2 \Phi = 0 \tag{2.5}$$

Symmetries play an important role in particle physics, since each symmetry yields a conservation law, as Noether's Theorem [18] shows. In classical physics, the invariance of time translation, space translation, and rotation leads to the conservation of energy, momentum, and angular momentum, respectively. Group theory, a mathematical tool, is widely used to study the symmetry of a physical system. For example, the symmetry of the global gauge transformation, which contains a constant phase parameter, yields charge conservation. It corresponds to a set of one-parameter transformations that characterizes the unitary group of order one (U(1)). The following sections in this chapter discuss the establishment of the SM based on gauge invariance, where the Lagrangian density is invariant under local gauge transformations. Each type of interaction follows the gauge symmetry of the corresponding group.

2.1.2 Quantum Electrodynamics

Quantum electrodynamics (QED) describes the electromagnetic interaction between elementary particles and arises from a local gauge symmetry under a U(1) transformation. The local gauge transformation can be written as $\psi \rightarrow \psi e^{iq\alpha(x)}$, where q is a real constant and $\alpha(x)$ is a continuous function of the space-time four-vector. The Dirac equation for spin $\frac{1}{2}$ fermions can be derived from the Lagrangian density in Equation 2.6,

$$\mathscr{L} = \bar{\psi} i\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi \tag{2.6}$$

where ψ is the Dirac spinor, γ^{μ} represents the 4×4 gamma matrices, and $\bar{\psi}$ is the adjoint Dirac spinor defined as $\bar{\psi} = \psi^{\dagger}\gamma^{0}$. The first term is the kinetic term, and the second term introduces the mass. Under the local gauge transformation, the derivative term in Equation 2.6 violates the invariance of the Lagrangian density. To recover the invariance, a covariant derivation D_{μ}, is introduced to replace the ∂_{μ} in the Lagrangian density. The new derivation is defined as $D_{\mu}\psi = \partial_{\mu} + iqA_{\mu}(x)$, where $A_{\mu}(x)$ is a new vector field and transform as $A'_{\mu}(x) = A_{\mu}(x) - \partial_{\mu}\alpha(x)$ under the local gauge transformation. Therefore, the Lagrangian density becomes Equation 2.12, with a new term that adds a vector field representing the interaction with photons.

$$\mathscr{L} = \bar{\psi} i\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi - q\bar{\psi}\gamma^{\mu}\psi A_{\mu} \tag{2.7}$$

The mass term brought by $A_{\mu}(x)$ is zero to ensure the local gauge symmetry, which agrees with the experimental fact that the photons are massless particles. The kinetic term for $A_{\mu}(x)$ has the form $\mathscr{L}_{A} = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu}$, where $F_{\mu\nu}$ is a gauge invariant field tensor defined as $F_{\mu\nu} = \partial_{\mu}A_{\nu} - \partial_{\nu}A_{\mu}$. Thus, the complete Lagrangian density for QED has the following form:

$$\mathscr{L}_{\text{QED}} = \bar{\psi} i\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi - q\bar{\psi}\gamma^{\mu}\psi A_{\mu} - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} \tag{2.8}$$

The field equation for A_{μ} and ψ can be derived from the Lagrangian density via the Euler-Lagrange equation, resulting in $\partial_{\mu}F^{\mu\nu} = q\bar{\psi}\gamma^{\nu}\psi$. One can thus define the charge current density as $j^{\nu} = q\bar{\psi}\gamma^{\nu}\psi$, where the coupling strength q is the electric charge of the Dirac particle.

2.1.3 Quantum Chromodynamics

Quantum chromodynamics (QCD) describes the strong interaction between color-charged particles and is derived by requiring a local gauge symmetry of a special unitary group of order three (SU(3)). There are three color charges in nature: red, green, and blue, leading to a three-dimensional color space. The SU(3) group is non-abelian, and its structure is defined by 8 linearly independent generators, which are indexed by a in this section. Its transformation can be written as a 3×3 matrix defined as $e^{ig_s\alpha^{a}(x)\cdot T^{a}}$, where T^{a} are the generators in 3 × 3 matrices in the color space, $\alpha^a(x)$ are real functions of the space-time vector, and g_s is a constant coupling strength parameter. The Lagrangian density can be written in the same format as in Equation 2.6, where ψ now becomes a three-component column vector:

$$\psi = \begin{pmatrix} \psi_r \\ \psi_b \\ \psi_g \end{pmatrix}, \ \bar{\psi} = (\bar{\psi}_r, \ \bar{\psi}_b, \ \bar{\psi}_g) \tag{2.9}$$

Analogous to the situation discussed in the QED section, the ∂_μ term is not gauge invariant since $\alpha^a(x)$ varies with space and time. A covariant derivative D_μ, defined as $D_\mu = \partial_\mu + ig_sT^aG^a_\mu$, replaces ∂_μ and introduces eight new vector fields G^a_μ corresponding to eight gluons. The Lagrangian density thus has the form:

$$\mathscr{L} = \bar{\psi}i\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi - g_s\bar{\psi}(\gamma^\mu T^a G^a_\mu)\psi \tag{2.10}$$

The last term represents the interaction between gluons and the elementary particles carrying color charge. Under the SU(3) symmetry, the G^a_μ follows the transformation:

$$G'^a_\mu = G^a_\mu - \partial_\mu\alpha^a(x) - g_s f^{abc}\alpha^b(x)G^c_\mu \tag{2.11}$$

where f^{abc} are the structure constants of SU(3), arising from the commutation relations of the generators T^a. To ensure local gauge invariance, the force carriers are assumed to be massless, which is confirmed by the gluon mass measurement. Gluons carry colour charge, so they can interact with each other. A kinetic term for G^a_μ is added to the Lagrangian density, leading to the final form of QCD:

$$\mathscr{L}_{QCD} = \bar{\psi}i\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi - g_s\bar{\psi}(\gamma^\mu T^a G^a_\mu)\psi - \frac{1}{4}G^a_{\mu\nu}G^{a\mu\nu} \tag{2.12}$$

2.1.4 Weak Interaction and Electroweak Unification

The weak interaction proceeds by exchanging W (W^+ or W^-) bosons and Z bosons, which are massive particles. Its theory was firstly developed to explain the neutron β decay, $n \to p + e^- + \bar{\nu}_e$, where a neutrino appears in the final state. In 1956, Wu's cobalt decay experiment indicated that parity (P) is not conserved by the weak interaction. The spins of the cobalt nucleus are polarized by a strong magnetic field, and the electrons from the process $^{60}\text{Co} \to {}^{60}\text{Ni} + e^- + \bar{\nu}_e$ are emitted in one direction instead of both directions, providing direct evidence of violating parity conservation. As a solution to explain the phenomenon, leptons and quarks are divided into right-handed and left-handed particles. The right-handed particles are considered to be isospin singlets and do not participate in the weak interaction. The left-handed particles participate in the weak interaction and are arranged into the following isospin doublets,

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L, \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L, \begin{pmatrix} u \\ d' \end{pmatrix}_L, \begin{pmatrix} c \\ s' \end{pmatrix}_L, \begin{pmatrix} t \\ b' \end{pmatrix}_L \tag{2.13}$$

Further experiments, such as kaon decay, show that the weak interaction also violates the conservation of the combined charge conjugation and parity (CP). One example is kaon decay. Assuming CP conservation, there are two types of kaons as two CP eigenstates: The first type of kaons decays into two pions, and has a short lifetime of 0.9×10^{-10} s; Another type decays into three pions, and has a long lifetime of 5×10^{-8} s. In 1964, the experiment by Fitch and Cronin observed a small fraction of long-lived kaons decaying into two pions, establishing CP violation in weak interactions.

The CP violation and the quark mixing mechanism can be well described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix. The matrix is a 3×3 unitary matrix, and is defined as

$$V_{\mathrm{CKM}} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \tag{2.14}$$

where each element (V_{ij}) represents the coupling between the two quarks (i and j). The matrix can be parameterized by three mixing angles and the KM phase parameters that introduce CP violation. The precision measurement of CKM matrix parameters is an important topic for completing the SM and a powerful tool for discovering new physics beyond the standard model (BSM). An important feature of the SM is the unitarity of the CKM matrix, written as $\sum_i V_{ij}V_{ik}^* = \delta_{jk}$ and $\sum_j V_{ij}V_{kj}^* = \delta_{ik}$, which is extensively tested by many experiments [19].

Electroweak Unification

The weak interaction and the electromagnetic interaction can be unified into the electroweak (EW) theory under the $\mathrm{SU(2)_L} \times \mathrm{U(1)_Y}$ symmetry, where the index L indicates left-handed and the index Y represents weak hypercharge. The weak hypercharge is referred to as Y=$2(Q - I_3)$, where Q is the electron charge and I_3 is the third component of the weak isospin.

The Lagrangian density should stay invariant under the global gauge transformation, where the right-handed and left-handed components behave differently, leading to the forbidding of the mass term. Therefore, the Lagrangian density of a free Dirac particle in EW becomes

$$\mathscr{L} = \bar{\psi}(i\gamma^\mu \partial_\mu)\psi \tag{2.15}$$

where ψ is equal to $\psi_R + \psi_L$, consisting of both right- and left-handed components.

The SU(2) group is non-abelian and has three linearly independent generators. The SU(2) transformation can be written as $e^{ig_W \alpha^a(x) T^a}$, where T^a are generators typically represented by 2×2 matrices in the isospin space, $\alpha^a(x)$ are functions of the space-time vector, and g_W is the coupling strength for gauge bosons in weak interaction. Since only the left-handed particles participate in the weak interaction, the Lagrangian density can be written as $\mathscr{L} = \bar{\psi}_L(i\gamma^\mu\partial_\mu)\psi_L$, where ψ_L is an isospin doublet

$$\psi_L = \begin{pmatrix}\psi_1 \\ \psi_2\end{pmatrix}_L \tag{2.16}$$

To ensure the invariance of Lagrangian density under the SU(2) local gauge transformation, a covariant derivative term D_μ is applied to replace the original derivative ∂_μ. Analogous to QCD, D_μ is defined as $\partial_\mu + ig_W T^a W^a_\mu$, where W^a_μ are three new vector fields corresponding to the gauge bosons. A gauge invariant kinetic term is added to the Lagrangian density, written as $-\frac{1}{4}W^{a\mu\nu}W^a_{\mu\nu}$.

In order to consider both right- and left-handed components, the $\mathrm{U(1)_Y}$ symmetry is added. In this case, the transformation of U(1) is written as $e^{ig'\frac{\mathrm{Y}}{2}\beta(x)}$, where Y is the hypercharge, g' is the coupling strength, and $\beta(x)$ is a function of the space-time vector. The covariant derivative then becomes $D_\mu = \partial_\mu + ig'\frac{\mathrm{Y}}{2}B_\mu(x)$, introducing a new vector field $B_\mu(x)$. The Lagrangian density arising from the $\mathrm{U(1)_Y}$ symmetry thus has the following form:

$$\mathscr{L}_\mathrm{Y} = \bar{\psi}(i\gamma^\mu\partial_\mu)\psi - g'\frac{\mathrm{Y}_R}{2}\bar{\psi}_R(\gamma^\mu B_\mu)\psi_R - g'\frac{\mathrm{Y}_L}{2}\bar{\psi}_L(\gamma^\mu B_\mu)\psi_L - \frac{1}{4}B^{\mu\nu}B_{\mu\nu} \tag{2.17}$$

The second and third term represent interactions for the right- and left-handed components, as the hypercharges of them are different. The last term shows the kinetic term of the new vector field.

By combining the Lagrangian density terms obtained from $\mathrm{SU(2)_L}$ and $\mathrm{U(1)_Y}$ symmetry, the Lagrangian density expression for electroweak interaction becomes:

$$\mathscr{L}_\mathrm{EW} = \bar{\psi}_L(i\gamma^\mu D_\mu)\psi_L + \bar{\psi}_R(i\gamma^\mu D_\mu)\psi_R - \frac{1}{4}B^{\mu\nu}B_{\mu\nu} - \frac{1}{4}W^{a,\mu\nu}W^a_{\mu\nu} \tag{2.18}$$

The covariant derivative is given by the sum of the two discussed earlier in this section, as shown in Equation 2.19. In the first line, the second term is a 2 × 2 matrix obtained from $\mathrm{SU(2)_L}$ symmetry, while the third term comes from $\mathrm{U(1)_Y}$ symmetry. By bringing in the explicit expressions for the three generators of $\mathrm{SU(2)_L}$, and by defining $W^\pm = \frac{1}{\sqrt{2}}(W^1 \mp iW^2)$ and $T^\pm = \frac{1}{\sqrt{2}}(T^1 \pm iT^2)$, D_μ can be separated into three terms, as shown in the last line of Equation 2.19.

$$\begin{aligned}
D_\mu &= \partial_\mu + ig_W T^a W^a_\mu + ig'\frac{Y}{2}B_\mu \\
&= \partial_\mu + ig_W \left\{ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}_\mu W^1_\mu + \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}_\mu W^2_\mu + \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}_\mu W^3_\mu \right\} + ig'\frac{Y}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} B_\mu \\
&= \partial_\mu + i\frac{g_W}{2}\begin{pmatrix} 0 & W^+ \\ W^- & 0 \end{pmatrix}_\mu + \frac{i}{2}\begin{pmatrix} g_W W^3 + g'YB & 0 \\ 0 & -g_W W^3 + g'YB \end{pmatrix}_\mu \\
&= \partial_\mu + ig_W(T^+W^+ + T^-W^-)_\mu + (ig_W T^3 W^3 + ig'\frac{Y}{2}B)_\mu \\
&= \partial_\mu + D^W_\mu + D^{\gamma,Z}_\mu
\end{aligned} \tag{2.19}$$

In the D_μ expression, the second term D^W_μ corresponds to the interaction from exchanging the W bosons, and the third term $D^{\gamma,Z}_\mu$ shows the interaction mediated by the mixing of photons and Z bosons. The vector field A_μ in QED (Equation 2.8) and a new vector field Z_μ for Z boson interaction can be derived from $D^{\gamma,Z}_\mu$ by introducing a mixing angle θ_W. The mixing matrix between the vector fields is written as:

$$\begin{pmatrix} W^3_\mu \\ B_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} \tag{2.20}$$

Following the discussion of D_μ, the Lagrangian density for a neutrino-electron doublet (Equation 2.22) in electroweak interaction can be written in the following form to show the interactions of four gauge bosons (W^+, W^-, γ, and Z) separately:

$$\begin{aligned}
\mathscr{L}'_{\text{EW}} &= \sum_{j=e,\nu_e} (\bar{\psi}_j i\gamma^\mu \partial_\mu \psi_j) - \frac{1}{\sqrt{2}} g_W (\bar{\nu}_{eL}\gamma^\mu W^+_\mu e_L + \bar{e}_L \gamma^\mu W^-_\mu \nu_{eL}) \\
- \sum_{j=e,\nu_e} &(\bar{\psi}_j \gamma^\mu (g_W I_3 \sin\theta_W + g'\frac{Y}{2}\cos\theta_W) A_\mu \psi_j + \bar{\psi}_j \gamma^\mu (g_W I_3 \cos\theta_W - g'\frac{Y}{2}\sin\theta_W) Z_\mu \psi_j) \\
&-\frac{1}{4}B^{\mu\nu}B_{\mu\nu} - \frac{1}{4}W^{a,\mu\nu}W^a_{\mu\nu}
\end{aligned} \tag{2.21}$$

$$\psi_L = \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L \tag{2.22}$$

Here, ψ includes both left- and right-handed components of the fermions. The first term shows the kinetic energy of the electron and neutrino. The second term describes the W boson interaction, which involves only left-handed particles. The third term describes interactions mediated by photons and Z bosons. The last two terms give the kinetic energy of the newly introduced vector fields. No classic mass term is allowed in the Lagrangian density to ensure gauge invariance. The problem raised by the absence of the mass term is addressed by the Brout-Englert-Higgs mechanism, which will be discussed in the following section.

2.1.5 The Brout-Englert-Higgs Mechanism

The SM is established under the $\mathrm{SU(2)_L \times U(1)_Y \times SU(3)}$ symmetry, where the Lagrangian density stays invariant under gauge transformation, thus no classical mass term is allowed. The fact is that many gauge bosons are massive, such as the W boson and the Z boson. The Brout-Englert-Higgs Mechanism is thus developed to endow the SM Lagrangian with mass by introducing a new scalar field, ϕ. The field ϕ has spin 0 and hypercharge 1, and is a $\mathrm{SU(2)_L}$ complex doublet with four degrees of freedom:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{2.23}$$

The Lagrangian density of the spin-0 particle ϕ is derived from the Klein-Gordon equation, and can be written as:

$$\mathscr{L}_\phi = (D^\mu \phi)^\dagger (D_\mu \phi) - V(\phi) \tag{2.24}$$

where D_μ is the covariant derivative in Equation 2.19 from the $\mathrm{SU(2)_L \times U(1)_Y}$ symmetry, and $V(\phi)$ is a postulated potential term, referred to as the Higgs potential. The potential should be invariant under the rotation of ϕ^+ and ϕ^0, thus is assumed to be:

$$V(\phi) = \mu^2 |\phi|^2 + \lambda |\phi|^4 \tag{2.25}$$

where μ and λ are constant parameters. The λ is required to be positive to ensure $V(\phi) > 0$ in the large ϕ area. The spontaneously broken symmetry requires μ^2 to be negative. Therefore, the Higgs potential has a Mexican-hat shape, as illustrated in Figure 2.2. The minimum of the potential corresponding to the ground state is at $|\phi| = v = \sqrt{\frac{-\mu^2}{\lambda}}$. The variable v is known as the Higgs vacuum expectation value, and is measured to be 246 GeV. One of the common choices for the ground-state Higgs field reads:

$$\phi_{\text{ground state}} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \tag{2.26}$$

Consider a small perturbation to the ground-state field,

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ v + \phi_3 + i\phi_4 \end{pmatrix} \tag{2.27}$$

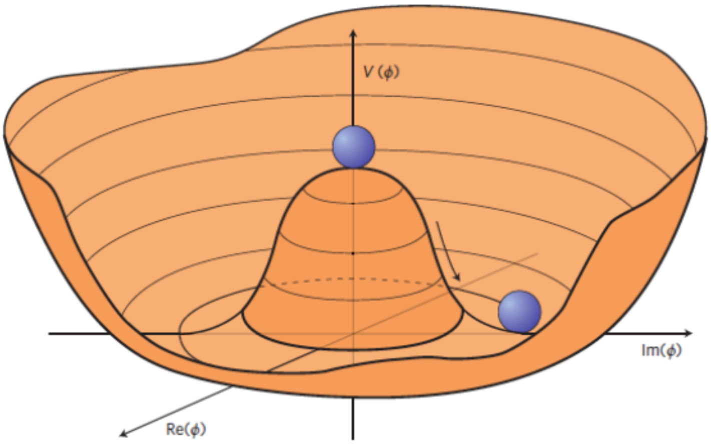

Figure 2.2: Higgs potential when $\mu^2 < 0$ [20].

This field equation should stay invariant when applying a $\mathrm{SU(2)_L}$ transformation $e^{ig_W\alpha^a(x)T^a}$. Therefore, ϕ_1, ϕ_2 and ϕ_4 in Equation 2.27 are required to be zero. The potential then becomes

$$\begin{aligned} V(\phi) &= \frac{1}{2}\mu^2(v+\phi_3)^2 + \frac{1}{4}\lambda(v+\phi_3)^4 \\ &= -\mu^2\phi_3^2 + \lambda v\phi_3^3 + \frac{1}{4}\lambda\phi_3^4 \end{aligned} \tag{2.28}$$

where the first term represents the Higgs boson mass, the second and third terms describe the self-interaction of the Higgs boson. By comparing the first term with the Klein-Gordon Lagrangian density, the Higgs boson mass is defined as $m_H = \sqrt{-2\mu^2}$, which is a free parameter measured to be 125 GeV. Hereby, ϕ_3 is usually referred to as H, representing the Higgs boson. In this case, the Higgs field ϕ can be written as

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v+H \end{pmatrix} \tag{2.29}$$

The interaction between the Higgs boson and the EW mediator bosons occurs via the covariant derivative term D_μ. With the D_μ defined in Equation 2.19, the first term in the Higgs boson Lagrangian density becomes:

$$\begin{aligned} (D^\mu\phi)^\dagger(D_\mu\phi) = \frac{1}{2}(\partial^\mu H)(\partial_\mu H) \\ +\frac{1}{4}g_W^2(v+H)^2W^{+,\mu}W_\mu^- + \frac{1}{8}(v+H)^2(W^{3,\mu}\ B^\mu)\begin{pmatrix} g_W^2 & -g_Wg' \\ -g_Wg' & g'^2 \end{pmatrix}\begin{pmatrix} W_\mu^3 \\ B_\mu \end{pmatrix} \end{aligned} \tag{2.30}$$

where the first term shows the kinetic energy of the Higgs field, the second term describes the interaction between the W boson and the Higgs field, and the third term shows the interaction mixing the Z boson and the photon. The third term can be rewritten to split the interactions with the Z boson and the photon. Using the mass eigenvalues of photon ($\tilde{m}_A = 0$) and Z boson ($\tilde{m}_Z = \frac{1}{2}\sqrt{g_W^2 + g'^2}$), and including the Higgs field potential in Equation 2.28, the Lagrange density of Higgs reads:

$$\begin{aligned}\mathscr{L}_\phi &= \frac{1}{2}(\partial^\mu H)(\partial_\mu H) + \frac{1}{4}g_W^2(v+H)^2 W^{+,\mu}W_\mu^- + (v+H)^2(\frac{1}{2}\tilde{m}_Z^2 Z^\mu Z_\mu + \frac{1}{2}\tilde{m}_A^2 A^\mu A_\mu) - V(\phi) \\ &= \frac{1}{2}(\partial^\mu H)(\partial_\mu H) + \frac{1}{4}g_W^2(v+H)^2 W^{+,\mu}W_\mu^- + \frac{1}{8}(v+H)^2(g_W^2 + g'^2)Z^\mu Z_\mu - V(\phi)\end{aligned} \tag{2.31}$$

The Lagrangian density can be expanded into several terms representing the interactions and mass terms for the W and Z bosons, respectively, as well as terms for the Higgs kinetic and potential. The mass values of the W boson and Z boson are determined by the EW coupling parameters, and can be written as:

$$m_W = \frac{1}{2}g_W v; \quad m_Z = \frac{1}{2}v\sqrt{g_W^2 + g'^2} = \frac{m_W}{\cos^2\theta_W} \tag{2.32}$$

The Higgs field also provides mass to the fermions, including the leptons and the quarks, through the Yukawa interaction. Taking the electron-neutrino doublet as an example, the Lagrange density of the Yukawa coupling has the form:

$$\mathscr{L}_{\text{Yukawa}} = -y_e(\bar{\psi}_L \phi e_R + \bar{e}_R \phi^\dagger \psi_L) \tag{2.33}$$

where ψ_L is the isospin doublet in Equation 2.22, e_R is the isospin singlet for electrons, and y_e is the coupling parameter. Applying the Higgs field in Equation 2.29, the $\mathscr{L}_{\text{Yukawa}}$ is written as:

$$\begin{aligned}\mathscr{L}_{\text{Yukawa}} &= -\frac{y_e}{\sqrt{2}}(\bar{e}_L(v+H)e_R + \bar{e}_R(v+H)e_L) \\ &= -\frac{y_e v}{\sqrt{2}}(\bar{e}_L e_R + \bar{e}_R e_L) - \frac{y_e H}{\sqrt{2}}(\bar{e}_L e_R + \bar{e}_R e_L) \\ &= -\frac{y_e v}{\sqrt{2}}\bar{e}e - \frac{y_e H}{\sqrt{2}}\bar{e}e\end{aligned} \tag{2.34}$$

The mass of the electron can be derived from the first term, refer to as $m_e = \frac{y_e v}{\sqrt{2}}$, where y_e is the coupling strength. The second term shows the interaction between the electron and the Higgs field. The Lagrange density of the Higgs field is the sum of Equation 2.31 and Equation 2.34.

$$\mathscr{L}_{\text{Higgs}} = \mathscr{L}_\phi + \mathscr{L}_{\text{Yukawa}} \tag{2.35}$$

2.1.6 Cross Section and Decay Width

The cross section and the decay width, predicted by Lagrangian field theory, are common observables for the experimental validation of a theory. In a collision experiment, the expected number of events (N) from a certain physics process can be calculated with $N = L_{int} \cdot \sigma$, where L_{int} is the integrated luminosity depending on the collider performance and running time, and σ is the cross section derived from theoretical calculation.

The cross section of a process can be derived from the transition probability and the particle flux. To compute it, a perturbation theory is adopted, assuming that a process can be expanded to several terms from the leading order to the highest order; each term is visualized by a Feynman diagram that contains external lines, internal lines, and vertices.

In a proton-proton collision process $pp \to$ X, the cross section can be calculated as:

$$\sigma \simeq \sum_{a,b} \int dx_a \int dx_b f_a(x_a, \mu_F^2) f_b(x_b, \mu_F^2) \hat{\sigma}_{ab \to X}(\hat{s}, \mu_R^2), \tag{2.36}$$

where f_a and f_b are parton distribution functions (PDF) inside the two protons making the elementary collision, and $\hat{\sigma}_{ab \to X}$ is the elementary cross section calculated with the matrix element derived from the Feynman diagrams. The effective center-of-mass energy of the parton-parton interaction is defined as $\hat{s} = x_a \cdot x_b \cdot s$, where s represents the center-of-mass energy at the LHC. A factorization factor, μ_F, is introduced to separate the short- and long-distance interactions described by $f_{a/b}$ and $\hat{\sigma}_{ab \to X}$. A renormalization factor, labeled as μ_R, is introduced to avoid the divergences caused by high-order Feynman diagrams and ensure the physical observables from all orders are finite.

In a physics process, the probability of the transition between the initial state and the final state per unit time is defined as

$$d\Gamma = V(2\pi)^4 \delta^4(P_f - P_i) |\bar{\mathscr{M}}_{fi}|^2 \prod_{f,i} \frac{1}{2E_{f,i}V} \prod_{f} \frac{V d^3 p_f}{(2\pi)^3} \tag{2.37}$$

where P_f and P_i are the momentum of the initial and final states, E is the energy, V is the space volume, and $\mathscr{M}_{fi}$ is the matrix element derived from the Feynman diagrams. The absolute square of the matrix element, $|\bar{\mathscr{M}}_{fi}|^2$, is an essential term in the calculation of cross section and decay width.

In a $X \to A + B$ decay process, the probability of the decay can be written as

$$d\Gamma = \frac{1}{(2\pi)^3} \frac{1}{2E_X^2} \delta^4(P_B + P_A - P_X) \frac{d^3 p_B}{2E_B} \frac{d^3 p_A}{2E_A} |\bar{\mathscr{M}}|^2 = \frac{|\vec{P_A}|}{E_X^2} \frac{d\Omega_A}{32\pi^2} |\bar{\mathscr{M}}|^2 \tag{2.38}$$

where $d\Omega_A$ is the differential angle of A in the rest frame of X, and the final term is derived from the middle term by an integral with the delta function. By integrating over the final state particle angle with a known expression of $\mathscr{M}$, one can derive the decay width of the

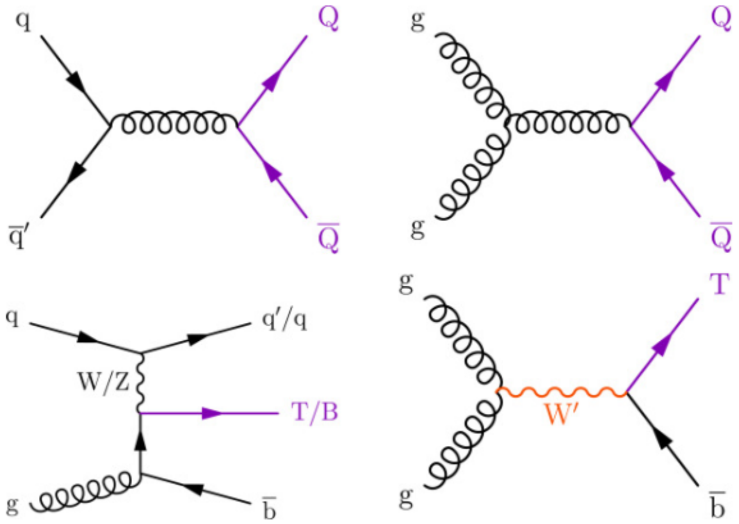

Figure 2.3: Feynman diagrams showing the productions of VLQ. Figure from [10].

particle X, which is labeled as Γ. In case the mother particle has multiple decay modes, its total decay width is the sum of all the partial decay widths from all decay modes. The branching ratio, labeled as $\mathscr{B}$, is the ratio of a partial width to the total width. The sum of all the branching ratios of a particle decay should be equal to 1. The lifetime of a particle can be identified from its total decay width, following the relation $\tau = 1/\Gamma$.

By further integrating over the final state particle angle with a known expression of $\mathscr{M}$, one can derive the decay width of the particle A, which is labeled as Γ. In case the mother particle has multiple decay modes, its total decay width is the sum of all the partial decay widths from all decay modes. Branching ratio, labeled as $\mathscr{B}$, is defined as the ratio of a partial width to the total width. The sum of all the branching ratios of a particle decay should be equal to 1. The lifetime of a particle can be identified from its total decay width, following the relation $\tau = 1/\Gamma$.

2.2 Physics Beyond the Standard Model

2.2.1 Vector-like Quarks

There are many open questions that cannot be answered within the SM framework, such as the hierarchy problem [4, 21–23] and the origin of neutrino masses. As an extension of the SM, vector-like quarks (VLQs) are considered as a possible scenario for a fourth generation of quarks.

In some BSM models, a straightforward extension of the SM is to add a fourth generation of fermions with the same fermionic pattern as the three SM generations [24]. These BSM fermions have been excluded by the joint Higgs cross section measurements by CMS and ATLAS [25]. The VLQs are hypothetical spin 1/2 quarks for which the left- and right-handed chiral components transform in the same way under the SM EW group. Unlike the excluded chiral fourth-generation quarks, the mass of VLQ is obtained from the mixing with SM quarks instead of the Yukawa coupling with the Higgs field, and thus is not constrained by the Higgs cross sections and decay modes. There are typically four types of VLQs considered, named T, B, X, and Y, with electric charges +2/3, −1/3, +5/3, and −4/3, respectively. Assuming the scalar sector only includes $\mathrm{SU}(2)_L$ doublets, there are seven gauge-covariant multiplets of VLQs under the $\mathrm{SU(2)_L} \times \mathrm{U(1)_Y} \times \mathrm{SU(3)}$ symmetry with definite quantum numbers [9]. Two singlets, three doublets, and two triplets of VLQs are listed below:

$$\begin{array}{r} T_{L,R}, \quad B_{L,R} \\ (X,\ T)_{L,R}, \quad (T,\ B)_{L,R}, \quad (B,\ Y)_{L,R} \\ (X,\ T,\ B)_{L,R}, \quad (T,\ B,\ Y)_{L,R} \end{array} \tag{2.39}$$

where T, B, X, and Y are newly introduced fields corresponding to the four types of VLQs. Only the vector-like top quark (T) and the vector-like b quark (B) are included in the singlets, while X and Y are additionally incorporated in the doublets and triplets.

The VLQs can be produced singly or in pairs at the LHC and have been extensively searched for. The single production of VLQ can occur through a t-channel exchange of a W or Z boson via electroweak interaction, as shown in the lower left plot in Figure 2.3. The pair production of VLQs proceeds typically through the strong interaction, as shown in the upper plots in Figure 2.3.

By adding VLQs to the SM, the up-type quark mass eigenstates are extended to u, c, t, T, and the down-type quark mass eigenstates become d, s, b, B. In the up sector, only the top quark has sizable mixing with T according to the experimental constraints [26]. For similar reasons, the b quark dominates the mixing with B in the down sector. Therefore, the t-T and b-B mixing can be parameterized by the following matrices:

$$\begin{aligned} \begin{pmatrix} t_{L,R} \\ T_{L,R} \end{pmatrix} &= U^t_{L,R} \begin{pmatrix} t_{L,R} \\ T_{L,R} \end{pmatrix} = \begin{pmatrix} \cos\theta^t_{L,R} & -\sin\theta^t_{L,R} e^{i\phi_t} \\ \sin\theta^t_{L,R} e^{-i\phi_t} & \cos\theta^t_{L,R} \end{pmatrix} \begin{pmatrix} t_{L,R} \\ T_{L,R} \end{pmatrix} \\ \begin{pmatrix} b_{L,R} \\ B_{L,R} \end{pmatrix} &= U^b_{L,R} \begin{pmatrix} b_{L,R} \\ B_{L,R} \end{pmatrix} = \begin{pmatrix} \cos\theta^b_{L,R} & -\sin\theta^b_{L,R} e^{i\phi_b} \\ \sin\theta^b_{L,R} e^{-i\phi_b} & \cos\theta^b_{L,R} \end{pmatrix} \begin{pmatrix} b_{L,R} \\ B_{L,R} \end{pmatrix} \end{aligned} \tag{2.40}$$

The two unitary matrices are determined by requiring that the mass matrices in the mass eigenstate bases are diagonal. The Lagrangian term describing the third generation and heavy quark mass based on the weak eigenstates is

$$\begin{aligned} \mathscr{L}_{\text{mass}} = &- \begin{pmatrix} \bar{t}_L & \bar{T}_L \end{pmatrix} \begin{pmatrix} y^t_{33} v/\sqrt{2} & y^t_{34} v/\sqrt{2} \\ y^t_{43} v/\sqrt{2} & M^0 \end{pmatrix} \begin{pmatrix} t_R \\ T_R \end{pmatrix} \\ &- \begin{pmatrix} \bar{b}_L & \bar{B}_L \end{pmatrix} \begin{pmatrix} y^b_{33} v/\sqrt{2} & y^b_{34} v/\sqrt{2} \\ y^b_{43} v/\sqrt{2} & M^0 \end{pmatrix} \begin{pmatrix} b_R \\ B_R \end{pmatrix} + H.c. \end{aligned} \tag{2.41}$$

where $y_{ij}^{t/b}$ are Yukawa coupling parameters, v is the Higgs vacuum expectation value introduced in section 2.1.5, and M^0 is a bare mass term, indicating that the VLQ mass does not arise from Yukawa coupling. The two matrices ($M^{t/b}$) in the Lagrangian formula 2.41 can be diagonalized with the mixing matrices in Equation 2.40

$$U_L^{t/b} M^{t/b} (U_R^{t/b})^\dagger = \begin{pmatrix} m_{t/b} & 0 \\ 0 & m_{T/B} \end{pmatrix} \tag{2.42}$$

The mixing angles are derived from the diagonalization of the mass matrix, and their left- and right-handed sectors are correlated. The expressions for left- and right-handed mixing angles, and the relations between them, are listed below for different multiplets.

For singlets and triplets:

$$\tan 2\theta_L^q = \frac{\sqrt{2}|y_{34}^q| v M^0}{(M^0)^2 - |y_{33}^q|^2 v^2/2 - |y_{34}^q|^2 v^2}, \quad \tan\theta_R^q = \frac{m_q}{m_Q} \tan\theta_L^q \tag{2.43}$$

For doublets:

$$\tan 2\theta_R^q = \frac{\sqrt{2}|y_{43}^q| v M^0}{(M^0)^2 - |y_{33}^q|^2 v^2/2 - |y_{43}^q|^2 v^2}, \quad \tan\theta_L^q = \frac{m_q}{m_Q} \tan\theta_R^q \tag{2.44}$$

where q is the SM top or bottom quark in up or down sectors, Q represents the vector-like quark T or B. Since X and Y do not mix with other fermions, they are not shown in these expressions. Since the VLQs have much larger mass than the SM quarks, one can clearly see that the mixing angle is always dominated by one chirality, especially when mixing with the b quark. Additionally, the mixing angles from the up and down sectors in triplets are also correlated. The relations of the mixing angles from the left-handed component, which is the dominant one, are listed below for two triplets.

$$\begin{aligned} \sin 2\theta_L^b &= \sqrt{2}\frac{m_T^2 - m_t^2}{m_B^2 - m_b^2} \sin 2\theta_L^t \quad (X,\ T,\ B) \\ \sin 2\theta_L^b &= \frac{1}{\sqrt{2}}\frac{m_T^2 - m_t^2}{m_B^2 - m_b^2} \sin 2\theta_L^t \quad (T,\ B,\ Y) \end{aligned} \tag{2.45}$$

To conclude, all the multiplets except the (T, B) doublet can be parameterized by a CP-violating phase ($\phi_{t/b}$ in Equation 2.40) and two parameters, including a heavy quark mass and a mixing angle parameter. Since the up- and down-sector components of the (T, B) doublet are uncorrelated, the model requires two mixing angle parameters and two CP-violating phases. These CP-violating phases are negligible in the simplified VLQ models in this section. In case of doublets and triplets, the bare mass term M^0 is shared by multiple heavy quarks and is split based on the mixing with SM fermions. The free parameters and the mass splitting in each multiplet are summarized below, where $s_{L/R}^{t/b}$ is short for $\sin\theta_{L/R}^{t/b}$, and $c_{L/R}^{t/b}$ is short for $\cos\theta_{L/R}^{t/b}$.

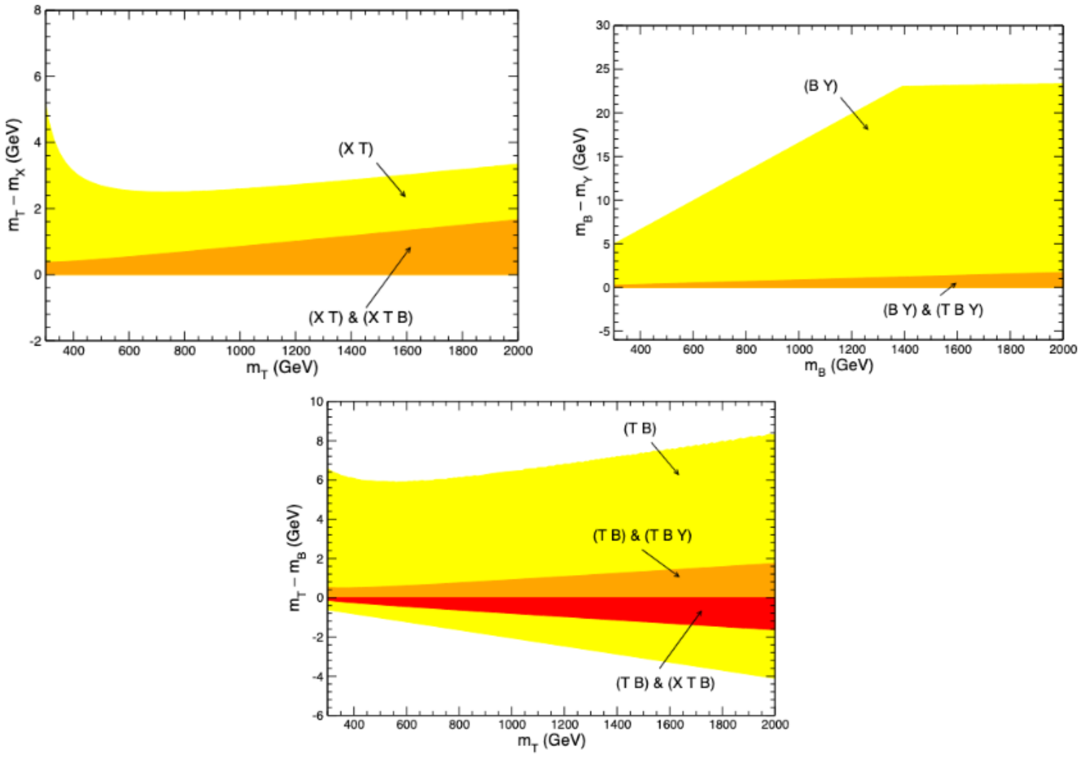

Figure 2.4: Allowed ranges for VLQ mass splittings. Plots are taken from [9].

- (T) singlet: s_L^t, m_T
- (B) singlet: s_L^b, m_B
- (X, T) doublet: s_R^t, m_T; $m_X^2 = m_T^2(c_R^t)^2 + m_t^2(s_R^t)^2$
- (T, B) doublet: s_R^t, s_R^b, m_T; $m_T^2(c_R^t)^2 + m_t^2(s_R^t)^2 = m_B^2(c_R^b)^2 + m_b^2(s_R^b)^2$
- (B, Y) doublet: s_R^b, m_B; $m_Y^2 = m_B^2(c_R^b)^2 + m_b^2(s_R^b)^2$
- (X, T, B) triplet: s_L^t, m_T; $m_X^2 = m_T^2(c_L^t)^2 + m_t^2(s_L^t)^2$, $m_T^2(c_L^t)^2 + m_t^2(s_L^t)^2 = m_B^2(c_L^b)^2 + M_b^2(s_L^b)^2$
- (T, B, Y) triplet: s_L^t, m_T; $m_Y^2 = m_B^2(c_L^b)^2 + m_b^2(s_L^b)^2$, $m_T^2(c_L^t)^2 + m_t^2(s_L^t)^2 = m_B^2(c_L^b)^2 + M_b^2(s_L^b)^2$

Based on these constraints, Figure 2.4 visualizes the allowed ranges for VLQ mass in each multiplet. From these plots, one can see that the mass values of the four types of VLQs are similar to each other, following $m_T \geq m_X$, $m_B \geq m_Y$. Therefore, the VLQs can not decay into each other. The allowed decay modes for different types of VLQs are:

$$\begin{aligned} T \to W^+b,\ T \to Zt,\ T \to Ht \\ B \to W^-t,\ B \to Zb,\ B \to Hb \\ X \to W^+t,\ Y \to W^-b \end{aligned} \tag{2.46}$$

Considering all the possible mixing terms, including the mixing with the SM Higgs, W bosons, Z bosons, and quarks, the partial width for T singlet $T \to tH$ decay becomes

$$\Gamma(T \to tH) = \frac{g^2}{128\pi}\frac{m_T}{m_W^2}\lambda(m_T, m_t, m_H)^{1/2}|Y_{tT}|^2[1 + 6r_t^2 - r_H^2 + r_t^4 - r_t^2 r_H^2] \tag{2.47}$$

where the λ function is defined as $\lambda(x, y, z) = (x^4 + y^4 + z^4 - 2x^2y^2 - 2x^2z^2 - 2y^2z^2)$, r_x is the ratio between m_x and m_T, and Y_{tT} indicates the light-heavy couplings to the Higgs boson. The left- and right-handed terms of Y_{tT} are $Y_{tT}^L = \frac{m_t}{m_T} s_L c_L$ and $Y_{tT}^R = s_L c_L$, where Y_{tT}^L is smaller than Y_{tT}^R.

The partial widths for $T \to tZ$ and $T \to bW^+$ of a T singlet have the following expressions:

$$\begin{aligned} \Gamma(T \to tZ) &= \frac{g^2}{128\pi c_W^2}\frac{m_T}{m_Z^2}\lambda(m_T, m_t, m_Z)^{1/2}(|X_{tT}^L|^2 \times (1 + r_Z^2 - 2r_t^2 - 2r_Z^2 + r_t^4 + r_Z^2 r_T^2)) \\ \Gamma(T \to bW^+) &= \frac{g^2}{64\pi}\frac{m_T}{m_W^2}\lambda(m_T, m_t, m_W)^{1/2}(|V_{Tb}^L|^2 \times (1 + r_W^2 - 2r_b^2 - 2r_W^4 + r_b^4 + r_W^2 r_b^2)) \end{aligned} \tag{2.48}$$

where $X_{tT}^L = s_L c_L$ is the coupling parameter to the Z boson, and $V_{Tb}^L = s_L$ is the coupling parameter to the W boson.

From the expressions of the partial widths, one can find that the branching ratios for the three T decay modes vary with the T mass. Figure 2.5 shows the allowed branching ratios for the decays of T and B in several different multiplets. The red points show the results assuming $m_{T/B} = 2$ TeV, which agrees with the very heavy VLQ mass assumption. Under this assumption, λ values in Equation 2.47 and 2.48 are close to m_T, and the terms with r_i are close to 1. On the left plot, the red point for T singlet shows that $\mathscr{B}(T \to tZ) \approx \mathscr{B}(T \to tH) \approx 0.5\ \mathscr{B}(T \to bW^+) = 25\%$. For a (T, B) doublet with $s_R = 0$, the T quark does not couple to Z and H, leading to a different partial width calculation with the result $\mathscr{B}(T \to bW^+) = 100\%$. The cross points are limits from the experimental measurements, showing good agreements with the theoretical predication.

The thesis focuses on the single T production through the exchange of a W boson in association with a b quark. The total cross section for a T decaying to a specific final state can be written as

$$\sigma(C_W, C, m_T, \Gamma_T) = C_W^2 C^2 \hat{\sigma}(m_T, \Gamma_T) \tag{2.49}$$

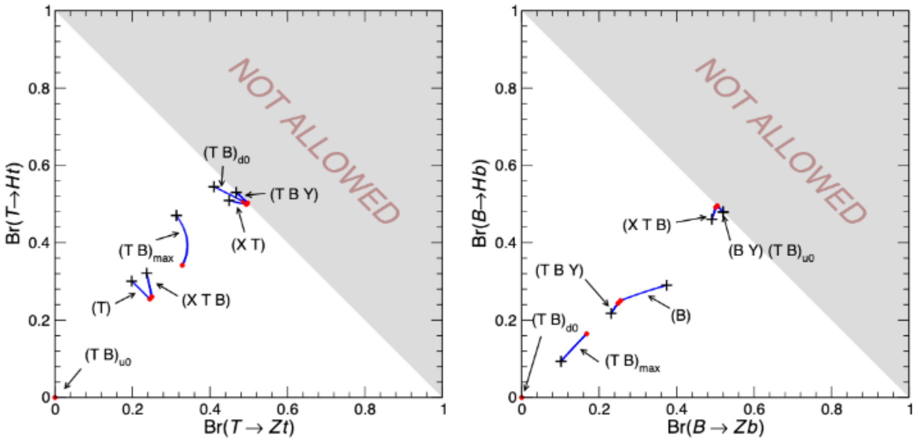

Figure 2.5: Allowed branching ratios for the decays of T and B. Plots are from [9].

where C_W is the production coupling parameter with the W boson, C represents the decay coupling parameter depending on the specific decay mode of T, Γ_T is the total decay width of T, and $\hat{\sigma}$ represents the reduced cross section for T with any decay width. As discussed previously, the decay width of T from the "Tbj" process depends on the coupling strength to the W boson. By assuming a relatively small coupling strength or constraining the ratio of the total T decay width to the T mass (Γ_T/m_T), the T is characterized by a small decay width compared to the experimental resolution. This approach is referred to as the narrow-width assumption (NWA), which typically requires $\Gamma_T/m_T \leq 10-15\%$. Under the NWA, the cross section in Equation 2.49 is simplified, and thus can be written as

$$\sigma(C_W, C, m_T, \Gamma_T) = C_W^2 \hat{\sigma}_{\text{NWA}}(m_T) \mathscr{B}(T \to \text{decay channel}) \tag{2.50}$$

The reduced cross section ($\hat{\sigma}_{\text{NWA}}$) calculation results and uncertainties under the NWA are shown in the right plot of Figure 2.6, where the red band shows the process studied by this thesis. The theoretical cross section of T also depends on the coupling parameters, which can be transferred to the corresponding mixing angles labeled as κ. In CMS searches, the cross section calculation for the T singlet assumes that the mixing angles between the VLQ and the W/Z/H bosons are equal: $\kappa_H = \kappa_W = \kappa_Z = \kappa$. Thus, the coupling factor κ is determined for a fixed VLQ mass and total width, and is directly connected to Γ_T/m_T.

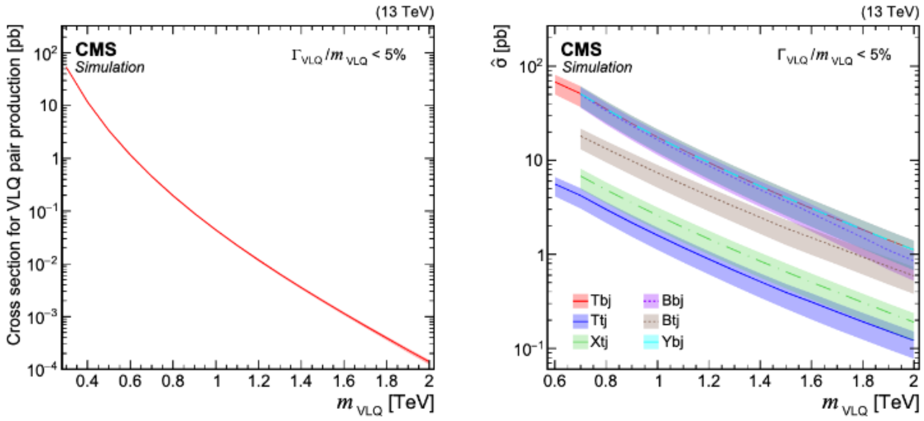

Figure 2.6: Theoretical cross sections of VLQs as a function of VLQ mass. Plots are from [10].

CHAPTER 3

CMS Experiment at the LHC

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is the highest-energy particle collider in the world today. Located at CERN, in France and Switzerland, the machine is installed in a 27-kilometer ring tunnel 100 meters underground, which was originally constructed for the Large Electron–Positron (LEP) machine. The physics motivation for this promising accelerator mainly includes the validation of the standard model and the discovery of new physics that potentially exists at the TeV scale.

As shown in Figure 3.1, the proton beams generated in linear accelerator 4 (LINAC 4) are accelerated in the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS). After having been transferred to the LHC beam pipes, two proton beams ramped to the target energy and ready to collide. The four interaction points are inside the four detectors ALICE, ATLAS, CMS, and LHCb. In this way, the collision process can be recorded by these detectors for various physics studies. The CMS and ATLAS experiments use general-purpose detectors constructed with different technologies, yet they share the same scientific goals, allowing them to cross-check each other. The ALICE experiment focuses on heavy-ion physics using e.g. lead-lead collisions, while the LHCb experiment focuses on flavour physics.

The center-of-mass energy and the luminosity are two key parameters of a collider. During the LHC operation, two proton or ion beams travel in opposite directions with the same energy. At the full intensity of proton-proton collisions, each beam consists of 2808 bunches, and each bunch consists 1.15×10^{11} protons [27]. Each of the two symmetric beams has an energy of E_{beam}; therefore, the center-of-mass energy is $\sqrt{s} = 2E_{beam}$. The instantaneous luminosity depends only on the beam parameters, and is computed as:

$$L = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi \epsilon_n \beta^*} F \tag{3.1}$$

where N_b is the number of particles per bunch, n_b is the number of bunches per beam, f_{rev} is the revolution frequency, γ_r is the relativistic gamma factor, ϵ_n is the normalized transverse beam emittance, β^* is the beta function at the collision point, and F is the geometric luminosity reduction factor due to the crossing angle at the interaction point [28]. The bunches cross every 25 ns, yielding multiple inelastic interactions in the detector. The additional interactions from hard scattering in each bunch crossing are referred to as pileup. A high instantaneous luminosity leads to a high pileup. The integrated luminosity is defined as $L_{int} = \int Ldt$, which is usually given in a unit of inverse femtobarn (fb^{-1}).

For proton-proton collisions, the LHC was designed to reach the highest center-of-mass energy at 14 TeV and a peak luminosity of $L = 7.5 \times 10^{34}$ cm^{-2}s^{-1} [28]. The first LHC operation, Run 1, from 2010 to 2012, led to the discovery of the Higgs boson at a center-of-mass energy of 7-8 TeV. The LHC Run 2 followed in 2015-2018, after the upgrade shutdown, with a center-of-mass energy of 13 TeV. The latest LHC operation from 2022 to 2026, Run 3, reached a center-of-mass energy of 13.6 TeV. The LHC has delivered over 300 fb^{-1} integrated luminosity to CMS in Run 3. After the machine upgrade, the High-Luminosity LHC will start in 2030 and reach the maximum designed center-of-mass energy at 14 TeV. As shown in Figure 3.3, the average pileup in Run 2 is around 40 per event, which rises to around 60 in Run 3 due to the increase in instantaneous luminosity. This thesis uses data collected by the CMS detector during Run 2, corresponding to a total delivered integrated luminosity of 150 fb^{-1}. Figure 3.2 shows the integrated luminosity delivered to CMS from 2011 to 2025 for proton-proton collisions.

3.2 The Compact Muon Solenoid Detector

The Compact Muon Solenoid (CMS) detector is a general-purpose detector designed to study proton-proton collisions at the TeV energy scale. As mentioned in its name, the CMS detector has a compact structure that contains all the detector materials at a height of 15 meters and a length of 21 meters. Thanks to the superconducting solenoid, the magnetic field can reach 3.8 Tesla inside the CMS detector. A dedicated muon detector, composed of a cylindrical barrel section and two planar endcap regions, provides precise muon measurements. Figure 3.4 shows the perspective view of the whole CMS detector.

Based on the geometry of the CMS detector, the CMS coordinate is defined in Figure 3.5 [31]. The left plot shows the coordinate systems including the LHC, and the right plot shows the coordinate systems inside the CMS detector. The origin is the Interaction Point (IP). In the Cartesian coordinate system, the x-axis points at the center of the LHC, and the z-axis points to the Jura along the beamline. Since the detector is cylindrical, CMS analyses often use the polar coordinate system. The momentum in the transverse plane, labeled as p_{T}, is widely used in CMS analyses. Another widely used variable, the pseudorapidity η is defined by $\eta = -\log[\tan(\theta/2)]$, and thus directly related to the polar angle θ.

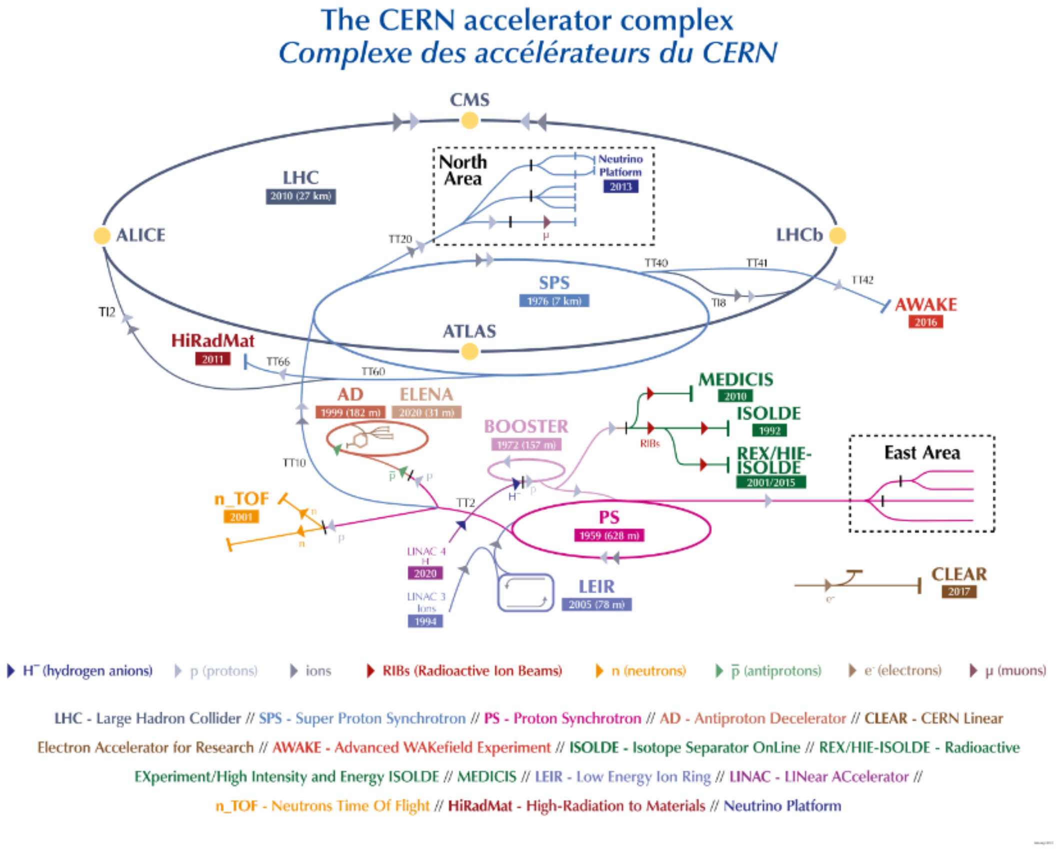
The CERN accelerator complex
Complexe des accélérateurs du CERN
CMS
North Area
Neutrino Platform
2013
LHC
2010 (27 km)
ALICE
LHCb
TT20
TT40
TT41
TI8
TT42
AWAKE
2016
SPS
1976 (7 km)
TI2
ATLAS
HiRadMat
2011
TT66
TT60
MEDICIS
2010
ISOLDE
1992
RIBs
REX/HIE-ISOLDE
2001/2015
East Area
AD
1999 (182 m)
ELENA
2020 (31 m)
BOOSTER
1972 (157 m)
TT10
n_TOF
2001
TT2
PS
1959 (628 m)
LINAC 4
H⁻
2020
LINAC 3
Ions
1994
LEIR
2005 (78 m)
CLEAR
2017
H⁻ (hydrogen anions)
p (protons)
ions
RIBs (Radioactive Ion Beams)
n (neutrons)
p̄ (antiprotons)
e⁻ (electrons)
μ (muons)
LHC - Large Hadron Collider // SPS - Super Proton Synchrotron // PS - Proton Synchrotron // AD - Antiproton Decelerator // CLEAR - CERN Linear Electron Accelerator for Research // AWAKE - Advanced WAKefield Experiment // ISOLDE - Isotope Separator OnLine // REX/HIE-ISOLDE - Radioactive EXperiment/High Intensity and Energy ISOLDE // MEDICIS // LEIR - Low Energy Ion Ring // LINAC - LINear ACcelerator // n_TOF - Neutrons Time Of Flight // HiRadMat - High-Radiation to Materials // Neutrino Platform

Figure 3.1: The CERN accelerator layout, showing the complex chain of the particle accelerators. The LHC is the dark blue ring. [29]

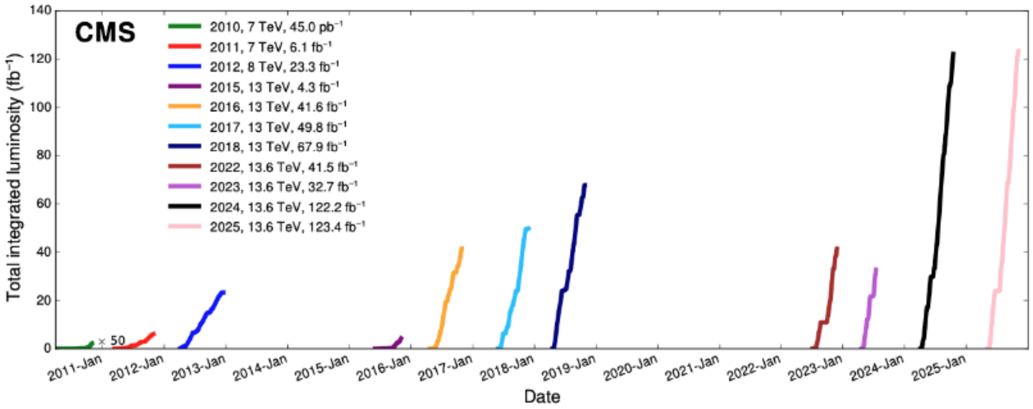

Figure 3.2: Integrated luminosity delivered to CMS from 2011 to 2025 (proton-proton data only) [30].

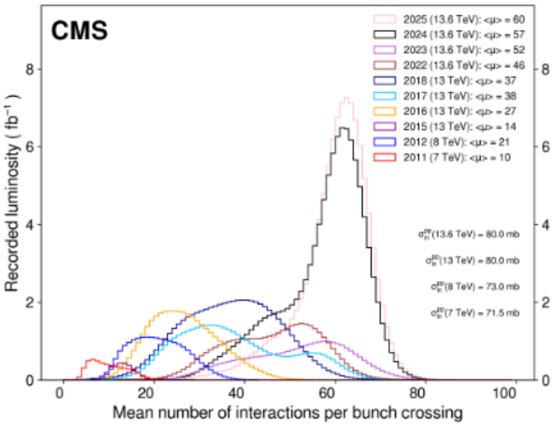

Figure 3.3: Interactions per crossing (pileup) for each year from 2011 to 2025 [30].

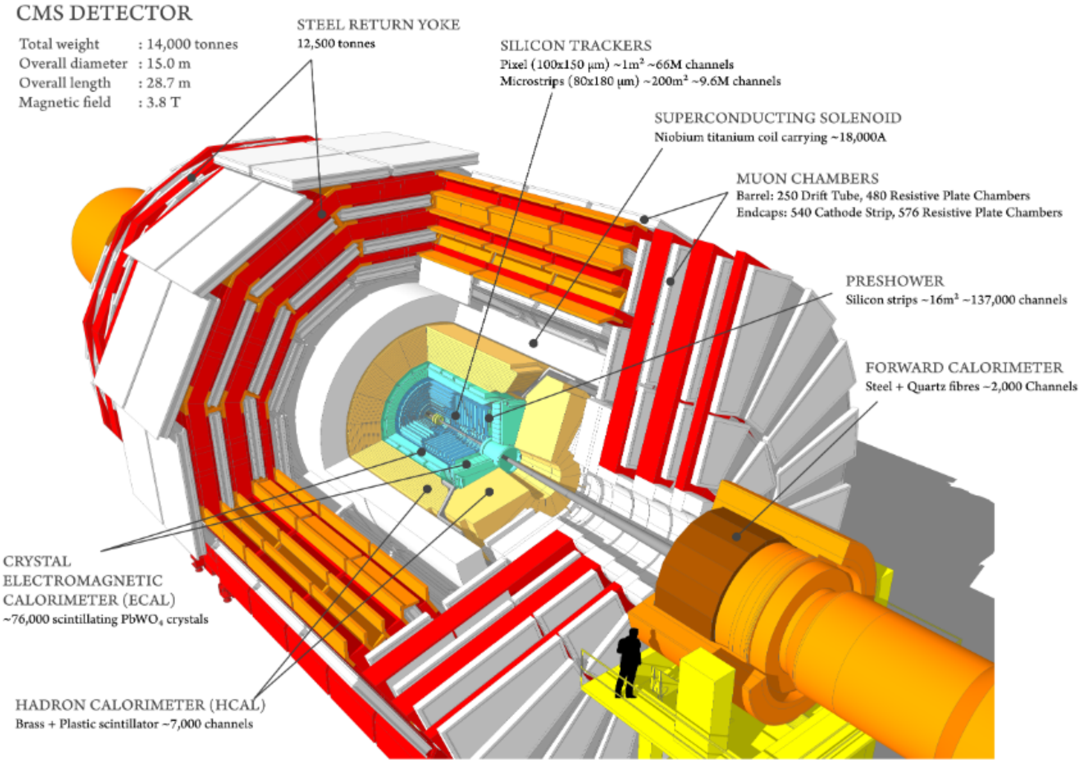
CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~1m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO$_4$ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 3.4: The CMS detector overview [32]

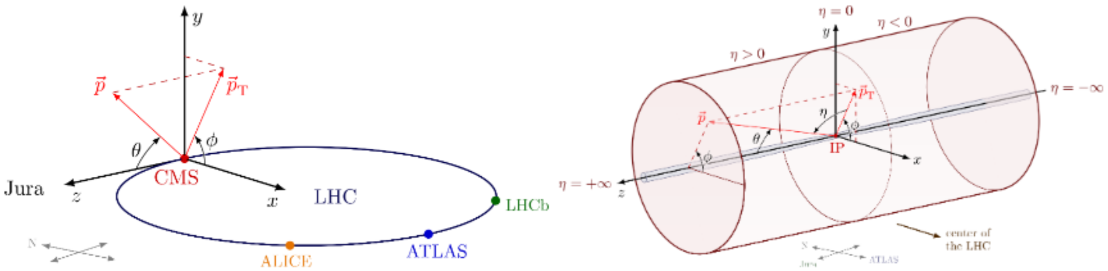

Figure 3.5: The CMS coordinate [33]

3.2.1 The Tracker System

The CMS tracker is an all-silicon detector, that can be divided into the pixel and strip systems. The charged particles interact with the silicon sensors, leaving hits when they pass through the magnetic field. Based on the hit information, track reconstruction, secondary vertex reconstruction, and even further particle identification can be performed. Figure 3.6 shows the CMS inner tracker schematic in cylindrical coordinates. The barrel part of the cylindrical tracker is within the $|\eta| < 1.6$ region, and the endcap part in the $1.6 < |\eta| < 2.5$ region.

The pixel system, located at the core of the detector near the interaction point, plays a crucial role in vertex reconstruction and precise measurement of charged particles. The pixel system went through three upgrade stages. The version for Run 2 data collection since 2017 is referred to as the Phase-1 pixel detector [34]. It consists of 1,856 modules, comprising the barrel pixel detector (BPIX) and the forward disks (FPIX), and covers the pseudorapidity range of $|\eta| < 2.5$.

The strip system surrounds the pixel detector, and it consists of 15148 modules from four subsystems: the Tracker Inner Barrel (TIB), the Tracker Inner Disks (TID), the Tracker Outer Barrel (TOB), and the tracker endcap (TEC). Compared to the pixel system, the strip system covers a bigger area and suffers less from the high radiation environment.

The information collected by the tracker system is used for impact parameter (IP) and secondary vertex measurements, both of which aid the b-tagging. For high p_T tracks at the 100 GeV scale, the transverse momentum resolution is around 1–2 % in the barrel region. The transverse impact parameter measurement for high p_T tracks is accurate to about 10 μm, thanks to the high-resolution pixel hit measurement [31]. The tracker reconstruction efficiency for muons can reach 99% over most of the acceptance.

3.2.2 The Electromagnetic and Hadronic Calorimeters

The CMS electromagnetic calorimeter (ECAL) and the hadronic calorimeter (HCAL) provide direct measurements of the particle energy from the collisions, as well as indi-

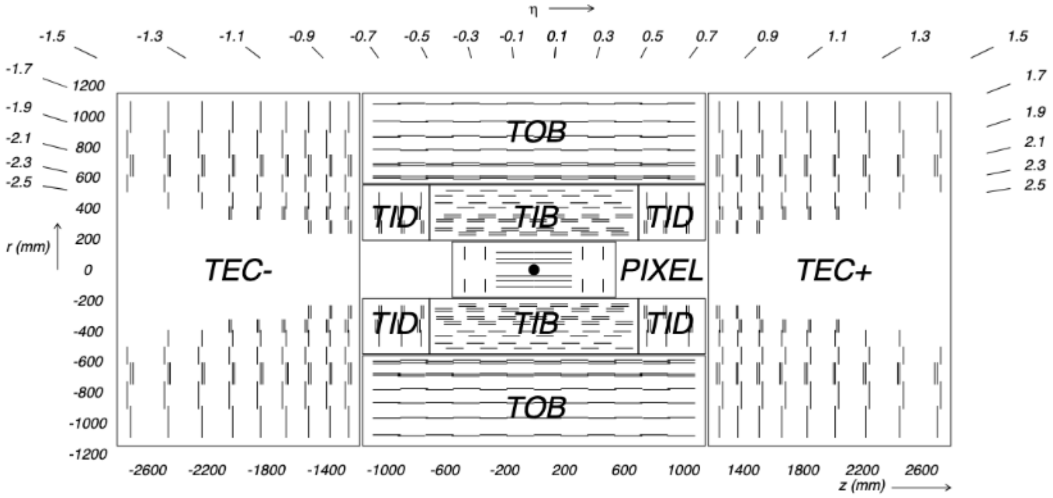

Figure 3.6: The CMS tracker sketch in r-z view [31]

rect measurements of missing energy. The ECAL measures the energy of electrons and photons, and the HCAL measures the energy of charged or neutral hadrons.

The ECAL is the inner layer of the two calorimeters [35, 36]. While the particles are traversing the ECAL, they interact with the scintillator material, and their energy is absorbed and transferred into light. The photodetectors, which are glued on the back of the scintillators, will detect the scintillation light and convert it into an electrical signal. The read-out system will collect the electronic signal and pass it to further analysis. As shown in Figure 3.7, the ECAL consists of a barrel part ($|\eta| < 1.479$) made of 61,200 lead tungstate ($\mathrm{PbWO_4}$) crystals and two endcap parts ($1.479 < |\eta| < 3$) with over 7,000 crystals. The crystals have a high density, a short radiation length, and a small Molière radius, ensuring low noise and high resolution. The preshower detectors are installed in front of the endcaps for the primary purpose of identifying neutral pions.

The energy resolution is a key parameter of the calorimeter performance, and the value for ECAL is defined by the following equation.

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{\mathrm{S}}{\sqrt{E}}\right)^2 + \left(\frac{\mathrm{N}}{E}\right)^2 + \mathrm{C}^2 \tag{3.2}$$

where E is the energy of the measured particle, S is the stochastic term, N is the noise term, and C is the constant term. Those terms are determined from the measurement discussed in [37], showing the results of $\mathrm{S} = 2.8\%,\ \mathrm{N} = 12\%,\ \mathrm{C} = 0.3\%$.

The HCAL [38], located behind the ECAL, measures the energy from the hadronic showers. It is a sampling calorimeter, tiling layers of absorber material and active elements. As shown in Figure 3.8, the hadron calorimeter consists of the HCAL barrel (HB),

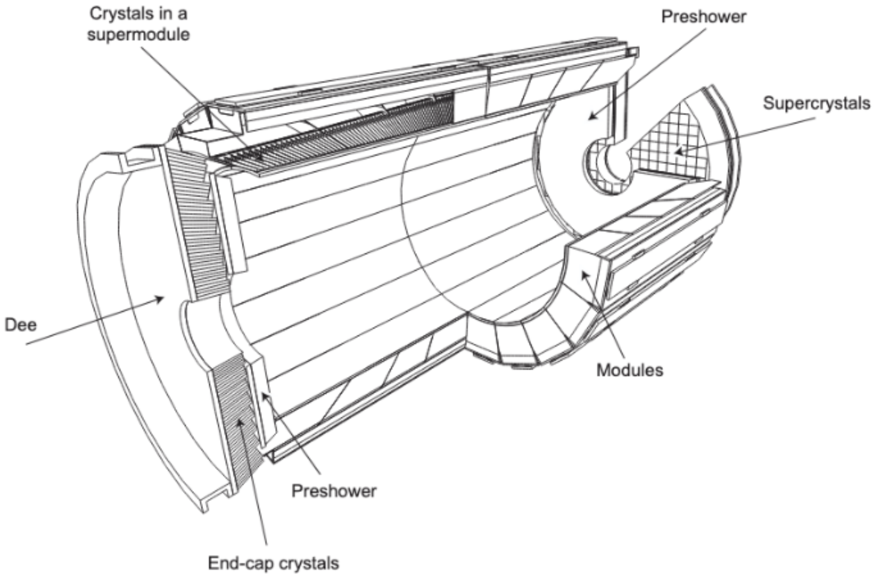

Figure 3.7: The CMS ECAL layout [36]

the HCAL endcap (HE), the HCAL outer (HO), and the HCAL forward (HF) parts; each part covers a specific η range. The HB made of two half-barrels covers the $|\eta| < 1.3$ range, while the two HE modules covering the range $1.3 < |\eta| < 3$. The HB and HE are placed in the high-magnetic field, so their absorb layers are made of copper and stainless steel, which are non-magnetic materials. Their active elements are plastic scintillator tiles. The HO is located behind the solenoid covering the HB η area, and acts as a tail-catcher for the hadron shower. The HF covers the high pseudorapidity range $3 < |\eta| < 5$, is a Cherenkov calorimeter designed for identifying forward jets, measuring missing energy, and monitoring luminosity. Quartz fibers are chosen as active layers for HO to detect Cherenkov light and to cope with the high radiation and high hadron rate environment in that area.

The combined energy resolution σ of ECAL and HCAL is defined by equation 3.3, where E is the energy of the measured particle, a is the stochastic term, and b is the constant term. According to the calibrations in [39], the terms measured with HE or HB are $a = 0.847\ GeV^{1/2}$ and $b = 0.074$. The corresponding values measured with HF are $a = 1.98\ GeV^{1/2}$ and $b = 0.09$.

$$\frac{\sigma}{E} = \frac{a}{\sqrt{E}} \oplus b \tag{3.3}$$

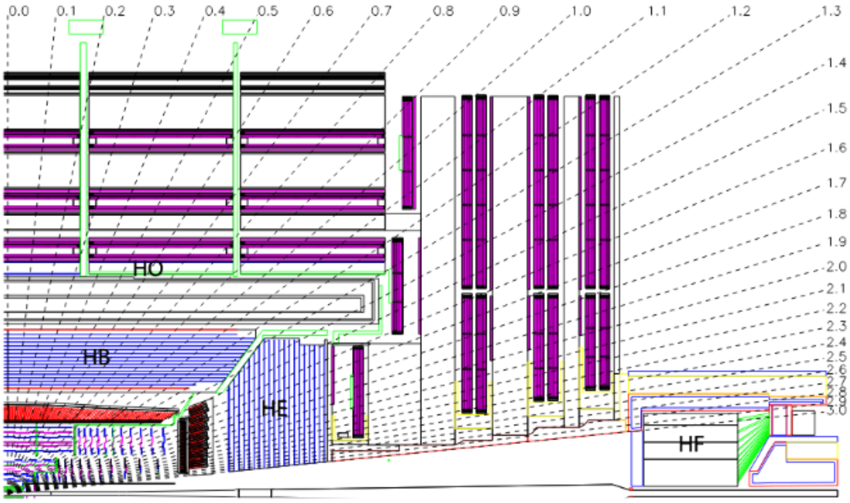

Figure 3.8: The CMS HCAL layout [40]

3.2.3 The Muon Detector System

The muon detector system provides muon information, including muon identification, momentum measurement, and triggering. Muons can travel through the inner detector materials with negligible energy loss, reach the outer muon chamber, and create clear signatures for various physics processes. The muon measurement by the muon detector system is independent from the inner tracker system, thus providing a possibility for cross-check. The muon detector includes the drift tube (DT) system, the cathode strip chambers (CSC), the resistive plate chamber (RPC) system, and the gas electron multiplier (GEM) chambers (Figure 3.9). Like other sub-detectors, the muon system is also cylindrical and can be divided based on pseudorapidity.

The DT system consists of four concentric cylindrical stations, covering the barrel region in $|\eta| < 1.2$, an area with a low signal rate expected. Each station is made of 60 or 70 drift tube chambers, filled with a gas mixture of 85% Ar and 15% CO_2 [31]. The main functions of the DT system are to provide muon position by hit pattern reconstruction and to contribute to the muon components of the level-1 triggers. The numbers and orientation of DT chambers are designed to achieve a good muon reconstruction efficiency at around 97% [41].

The CSC system contains four stations, functioning as multi-wire proportional counters with a finely segmented cathode strip readout [41]. It is located in the endcap regions $1.0 < |\eta| < 2.4$, where the muon rates and the background levels are higher than in the barrel region, and the magnetic field is non-uniform. Besides muon position measurement, its function also includes the muon trigger. The CSC modules are filled with a mixture

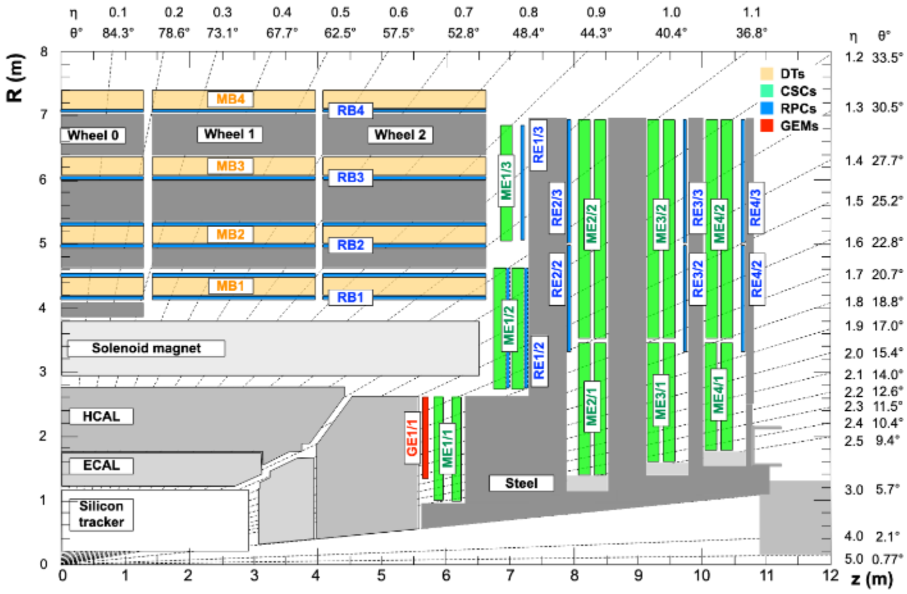

Figure 3.9: The CMS muon detector system [43].

of gases of 40% Ar, 50% CO_2, and 10% CF_4, providing a fast detector response time.

The GEM chamber and the RPCs form a trigger system that can quickly and independently identify the LHC bunch crossings associated with muon tracks. The RPCs are embedded in the barrel and endcap muon detectors, covering $|\eta| < 1.6$. The GEM chamber is placed in front of the first endcap muon station in the $1.6 < |\eta| < 2.2$ region to enhance forward muon triggering and tracking performance [42].

3.2.4 The Trigger System

The LHC bunch crossing rate can reach 40 MHz, but most events from collisions do not contain the target physics process for further analysis. Thus, it is crucial to introduce a trigger system to decide whether to keep or discard the events. The CMS trigger system is designed to reduce the event rate to a few kHz and consists of two levels: the Level 1 (L1) trigger, based on hardware, and the high-level trigger (HLT), based on software. During LHC Run 2, the highest instantaneous luminosity reached around 2×10^{34} cm^{-2}s^{-1}, and the average pileup reached 40, necessitating an upgrade for the trigger system [44].

The L1 trigger system uses information collected by the muon detector and the calorimeters, and makes a quick decision for each event in at most 4 μs after the collision occurs. With the L1 trigger's effort, the event rate reduced to 100 kHz from 40 MHz. In the muon detector system, the trigger electronics decide the existence of the muons based on the track segments or hit patterns in the muon chamber. The calorimeters' trigger determines the existence of jets, electrons, and photons based on their energy deposits.

The global trigger (GL) decides to keep or drop the event based on the feedback from the sub-detectors. Events possibly involving an interesting physics object, such as an electron, a photon, a muon, or a jet, will be passed to the HLT for further processing.

The HLT system uses the complete read-out data and performs analysis-like calculations to select events that contain specific physics processes. More than 400 HLT paths participated in the CMS Run 2 data-taking, each of which included specific requirements of the physics objects. For example, this thesis uses the single lepton triggers with various p_T and isolation requirements. Events that pass the specific HLT are directed to a corresponding primary dataset. The analyzers will select the primary datasets and the matched HLT paths based on the physics process they are searching for.

The HLT reduced the event rate to 1.5 kHz in Run 2. To keep the total event rate within the maximum bandwidth, each HLT path is assigned a threshold. Once an HLT is expected to record a large amount of events that exceeds its bandwidth threshold, a measure named "prescale" could be applied. The prescale method records a fraction of events that pass the HLT requirement, rather than accepting all events, to keep the HLT bandwidth within its assigned maximum. It is usually applied to HLT paths with high expected event rates, such as the low-p_T triggers used in the B-physics study. As for the HLTs used in this thesis, the prescale factors are 1, indicating that all events containing the required high-p_T leptons are recorded in the primary datasets.

3.2.5 Data Acquisition and Processing

The Data Quality Monitoring (DQM) system monitors the data-taking process to ensure that no technical issues affect the data quality. It certifies the good data and puts the list of good run numbers into JavaScript object notation (JSON) files. To ensure high quality of the data used in the analysis, the JSON files in Table 3.1 for all three years are applied.

The raw data collected by the CMS detector go through several processing steps before being used in physics analysis. The output of each step corresponds to a data tier: RAW, RECO, AOD, MiniAOD, and NanoAOD. When data is promoted from a lower tier to a higher tier, it loses information, but becomes more lightweight and readable for physics analyzers. The RECO level datasets contain the reconstructed objects and hits/clusters. AOD is short for Analysis Object Data, which contains the reconstructed physics objects and reduced hit information; therefore, it is the starting point for physics analysis. As reflected in their names, MiniAOD and NanoAOD are the simplified and optimized versions of AOD; thus, they are widely used in the CMS physics analyses. This analysis uses the NanoAOD level datasets, as they include all the necessary information. The processing requires up-to-date configurations to simulate the detector condition. To centralize the procedure, the CMS Software (CMSSW) has been developed. The CMSSW is not only used for data processing, but also in Monte Carlo simulation and physics analysis setups. The analysis uses the NanoAODv9 datasets from MiniAODv2, which are

2016	Cert_271036-284044_13TeV_Legacy2016_Collisions16_JSON.txt
2017	Cert_294927-306462_13TeV_UL2017_Collisions17_GoldenJSON.txt
2018	Cert_314472-325175_13TeV_Legacy2018_Collisions18_JSON.txt

Table 3.1: Golden JSON files used for this analysis.

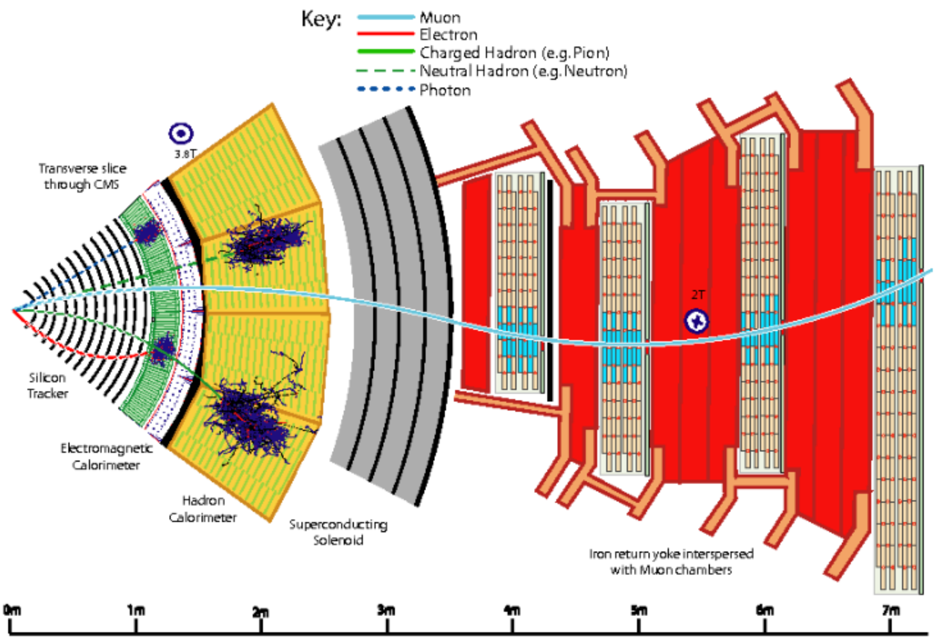
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
3.8T
Transverse slice through CMS
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
2T
Iron return yoke interspersed with Muon chambers
0m
1m
2m
3m
4m
5m
6m
7m

Figure 3.10: Particle interactions in a transverse slice of the CMS detector [45].

the recommended versions for analyses using Run 2 data.

3.3 Physics Objects Reconstruction

The particle-flow (PF) algorithm [45] reconstructs and identifies physics objects by combining information from the CMS sub-detectors, including the tracker, HCAL, ECAL, and muon chamber. Figure 3.10 shows that each particle type interacts distinctively with the CMS detector. As particles travel from the beam interaction area through the subdetectors, they may leave hits in the tracker and energy clusters in the calorimeters. Their detailed behavior according to their kind allows for reconstruction and identification.

3.3.1 Tracks and Vertices

The tracks and vertices are reconstructed from the hits in the tracker system. A combinatorial track finder based on the Kalman Filter [46]reconstructs tracks in three steps. Firstly, an initial "seed" is found with a few hits compatible with a predicted charged-particle trajectory. Next, hits from all tracker layers are included to build a further trajectory candidate with the Kalman Filter. Finally, a fitting step determines charged-particle properties, such as the origin, p_{T}, and direction. The reconstruction efficiency under stringent track quality criteria reaches 99 % for isolated muons at the 10 GeV scale, then slightly drops in the higher p_{T} region.

The primary-vertex (PV) reconstruction measures the location of proton-proton interactions using the reconstructed tracks in a single event [47]. The tracks near the beam spot area are selected for reconstruction. Various requirements, such as the maximum significance of the transverse impact parameter to the beam spot to be less than 5 ($d_0 < 5$), are applied to select the tracks that are comparable with those coming from the IP. Then the selected tracks are clustered using a deterministic annealing (DA) algorithm [48], based on the minimal distances between the tracks and the beam spot center. The vertex positions and associated information, such as the goodness of fit, are obtained with an adaptive vertex fitter. The vertices that are most comparable to the interaction point are selected as PV, while others are considered as pileup vertices.

A secondary vertex (SV) can originate from the decay of an unstable particle or from an interaction with detector material [49], so its reconstruction and measurements are crucial for b-quark tagging. The key parameter to identify a track from SV is the impact parameter, which can be measured in different coordinate systems. The transverse impact parameter (d_0) and the longitudinal impact parameter (d_z) are the distances between the track and the PV in the transverse and longitudinal planes. The impact parameter can be measured in three dimensions (3D) (d_{3D}), and its significance is defined as $s_{\mathrm{3D}} = d_{\mathrm{3D}}/\sigma[d_{\mathrm{3D}}]$, where $\sigma[d_{\mathrm{3D}}]$ is the uncertainty of d_{3D}. An SV is reconstructed from tracks with high impact-parameter significance, and its position is determined by fitting the reconstructed vertex.

3.3.2 Muons

Muons in CMS are reconstructed combining the information from two sub-systems: the inner tracker and the muon chamber. As shown in Figure 3.10, a muon generated near to the Interaction Point (IP) travels through the inner tracker, passes the calorimeters almost losing no energy, and reaches the muon system. There are three types of reconstructed muons, distinguished by the information used for their reconstruction: standalone muons, global muons and tracker muons. A standalone muon is reconstructed with the hits from DT, CSC, and RPC in the muon system, independent from the other CMS sub-systems. A global muon track is based on the standalone muon track, then matched and fitted with the hits from the inner tracker. It increases the muon p_{T} measurement resolution and the reconstruction efficiency of high p_{T} muons, especially those with p_{T} higher than 200 GeV. The tracker muons are reconstructed differently. Starting with the tracks from the inner tracker, the tracks with $p_{\mathrm{T}} > 0.5$ GeV and $p > 2.5$ GeV are extrapolated to the muon system, and the ones with at least one matching segment in the muon system are identified as tracker muons. The tracker muon reconstruction benefits the low-p_{T} muons, especially the ones with $p_{\mathrm{T}} < 10$ GeV. By combining the three types of tracks, about 99 % of muons can be identified as a tracker muon or a global muon, or both.

In the PF algorithm, the muon identification (ID) criteria are based on selections applied to the reconstructed global muons and tracker muons. There are various muon ID criteria, including cut-based ID and multivariate (MVA) ID. Each ID criterion consists of several working points, progressing from the loosest to the tightest one. The analysis in this thesis requires cut-based ID at a tight working point. A tight-ID muon must be both a tracker muon and a global muon, passing the goodness-of-fit requirements, with a track reconstructed from tracker hits in at least six inner tracker layers and including one pixel hit [41]. The tight muon ID aims to select the prompt muons from the PV instead of from a hadron decay; therefore, it requires the muon to have a transverse impact parameter (d_0) smaller than 0.2 cm, and a longitudinal impact parameter (d_z) smaller than 0.5 cm.

Isolation is an important variable that quantifies the separation of muons from other objects in η /ϕ space [50]. This analysis uses the I_μ variable for muon relevant isolation factor, as defined below:

$$I_\mu = \frac{1}{p_{\mathrm{T}}} \sum_{R=0}^{\Delta R^\mu(p_{\mathrm{T}})} (p_{\mathrm{T}}^{\text{charged hadrons}} + max(p_{\mathrm{T}}^{\text{photons}} + p_{\mathrm{T}}^{\text{neutral hadrons}} - \beta p_{\mathrm{T}}^{\text{charged pileup}}, 0)) \quad (3.4)$$

where the ΔR variable is from Equation 3.5, which describes a cone of an angle around the muon direction. The threshold of the cone, ΔR^μ, is a function of the muon p_{T}, instead of a fixed value. This is an optimization for muons from the decay of a high-momentum particle. A small I_μ value close to zero means a high isolation level. The analysis requires the isolation I_μ of each muon to be smaller than 0.05.

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \quad (3.5)$$

$$\Delta R^\mu = \begin{cases} 0.2, \quad p_{\mathrm{T}} < 50 \text{ GeV} \\ 10 \text{ GeV}/p_{\mathrm{T}}, \quad 50 < p_{\mathrm{T}} < 200 \text{ GeV} \\ 0.05, \quad \text{otherwise} \end{cases} \quad (3.6)$$

3.3.3 Electrons

The electron reconstruction utilizes the information from the inner tracker and the ECAL. In Figure 3.10, a prompt electron firstly travels through the inner tracker and leaves hits in the tracker layers, then arrives at the ECAL and leaves clusters of deposit energy. When an electron interacts with the detector material in front of the ECAL, it emits or produces bremsstrahlung photons, producing a shower containing multiple electrons and photons. The energy of an electron can be mostly absorbed by ECAL.

The ECAL-based electron reconstruction begins with the measurement of ECAL clusters [51], which are energy deposits from the particle showers produced by the electrons. The energy of each ECAL cluster with $E_{\mathrm{T}} > 4$ GeV is combined to obtain a primary energy. The cluster with the biggest energy deposit compared to its neighboring cells is defined as the seed cluster. The clusters around the seed cluster within a given area are combined into a supercluster (SC) that captures the energy of the original electron. To measure the electron's kinematic properties, the position of a SC is matched with the inner tracker hits, using the reconstruction algorithm based on the Gaussian sum filter (GSF) [52].

As a complement to the ECAL-based reconstruction method, the tracker-based method benefits the low-p_{T} electron reconstruction. These tracks are reconstructed using a Kalman Filter (KF), a common tool for general track reconstruction as mentioned earlier. All the tracks in one event with $p_{\mathrm{T}} > 2$ GeV are matched with the electron trajectory hypothesis.

The electron reconstruction combines information from the calorimeters and the tracker, including the SCs, the GSF tracks, and the KF tracks. Identifications are defined based on this information. This analysis uses cut-based electron identification at a tight working point, with the requirements listed in Table 3.2. The cut variables in the table are defined as follows:

- The variable $\sigma_{i\eta i\eta}$ is the second moment of the weighted distribution of crystal energies in η, as defined in Equation 3.7. Here, η is the pseudorapidity of the crystal and w is a weight factor from the energy deposit. The calculation uses a 5×5 matrix centered around the most energetic crystal in the SC.

$$\sigma_{i\eta i\eta} = \sqrt{\frac{\sum_i^{5\times5} w_i(\eta_i - \bar{\eta})^2_{5\times5}}{\sum_i^{5\times5} w_i}} \tag{3.7}$$

- The angular variable $|\Delta\eta_{\mathrm{in}}^{\mathrm{seed}}|$ is defined as $|\eta_{\mathrm{seed}} - \eta_{\mathrm{track}}|$. $|\Delta\phi_{\mathrm{in}}|$ is defined as $|\phi_{\mathrm{SC}} - \phi_{\mathrm{track}}|$. "Seed" means the ECAL cluster seed, and "track" means the track η is extrapolated from the innermost track position.

- H/E is the ratio of energy collected by the HCAL to that of the ECAL. A small value indicates that most of the energy is absorbed by ECAL, consistent with the electron hypothesis.

- $I_{\mathrm{combined}}/E_{\mathrm{T}}$ is the relevant isolation factor, where the isolation factor I_{combined} is divided by the electron E_{T}. I_{combined} is defined as $I_{\mathrm{ch}} + max(0,\ I_n + I_\gamma - I_{\mathrm{PU}})$. It

Variable	Barrel	Endcaps
$\sigma_{i\eta i\eta}$	< 0.01	< 0.035
$\lvert\Delta\eta_{\text{in}}^{\text{seed}}\rvert$	< 0.0025	< 0.005
$\lvert\Delta\phi_{\text{in}}\rvert$	< 0.022 rad	< 0.024 rad
H/E	$< 0.026 + 1.15\ \text{GeV}/E_{\text{SC}}$ $+0.032\rho/E_{\text{SC}}$	$< 0.019 + 2.06\ \text{GeV}/E_{\text{SC}}$ $+0.183\rho/E_{\text{SC}}$
$I_{\text{combined}}/E_{\text{T}}$	$< 0.029 + 0.51\ \text{GeV}/E_{\text{T}}$	$< 0.0445 + 0.963\ \text{GeV}/E_{\text{T}}$
$\lvert 1/E - 1/p\rvert$	$< 0.16\ \text{GeV}^{-1}$	$< 0.0197\ \text{GeV}^{-1}$
$n_{\text{missing hits}}$	≤ 1	≤ 1
conversion veto	Pass	Pass

Table 3.2: Cut-based electron ID requirements at the tight working point, from Table 6 in Ref [51]

is a combination of isolation factors of charged hadrons (I_{ch}), photons (I_γ), neutral hadrons (I_{n}) in side a cone of $\Delta R = 0.3$, including a correction for pileup effects(I_{PU}).

- In the variable $|1/E - 1/p|$, E is the SC energy and p is the momentum of the track closest to the vertex. The electron hypothesis predicts a small value for a high-quality reconstruction.

- The conversion-safe electron veto rejects photons misidentified as electrons using pixel hit information.

3.3.4 Jets

The hadronization and fragmentation products of quarks and gluons produced in proton-proton collisions are referred to as jets. A jet usually consists of several particles, such as hadrons from the parton shower, photons created during hadronization, and soft leptons from hadron decay. As shown in Figure 3.10, charged hadrons leave hits in the tracker system, then produce showers, which typically start at ECAL and extend to the HCAL. Neutral hadrons bypass the tracker, potentially shower in the ECAL, and ultimately create showers and deposit all their energy in the HCAL.

Jet reconstruction uses the anti-κ_t jet clustering algorithm [53], taking the information from the tracker, ECAL, and HCAL as inputs. The anti-κ_t jet clustering algorithm [53] is an infrared and collinear safe method, designed from the typical character of a jet, which is a well-separated hard particle surrounded by many soft ones in a cone. A key parameter in this algorithm is the distance parameter between two particles (d_{ij}) defined in Equation 3.8, where ΔR_{ij} is from Equation 3.5.

$$d_{ij} = min(p_{\mathrm{T},i}^{-2}, p_{\mathrm{T},j}^{-2})\frac{\Delta R_{ij}^2}{R^2} \tag{3.8}$$

Among all possible pairings, the algorithm always combines the particle pair with the smallest d_{ij} value first. The factor $\min(p_{\mathrm{T},i}^{-2}, p_{\mathrm{T},j}^{-2})$ in Equation 3.8 emphasizes pairings dominated by higher p_{T} objects, so a hard particle tend to be clustered with the soft particles surrounding it first, while soft particles are likely to be clustered later with another hard particle. The clustering stops when the calculated $d_{ij} < p_{\mathrm{T},i}^2$. The constant variable R sets a boundary on the clustering, representing the radius of the jet cones. In CMS, the R value can be set to 0.4 for small-radius jets, and 0.8 for large-radius jets. This analysis uses the isolated jets reconstructed with $R = 0.4$, referred to as AK4 jets.

The jet identification (ID) criteria are based on the phenomenological studies of jets. Each jet is expected to have approximately 65% of its energy from charged hadrons, 25% from photons or neutral pions, and 10% from neutral hadrons. The jet ID applies selection criteria on the fractions of charged hadron, neutral hadron, charged electromagnetic, and neutral electromagnetic energy [54]. Each selected fraction must be compatible with the jet hypothesis and fall within predefined thresholds.

The pileup effect from low-energy proton-proton collisions brings difficulties for jet reconstruction and measurement. Multiple methods, such as the Charged Hadron Subtraction (CHS) [55] method and the pileup jet ID, are developed to mitigate the effect. The CHS method removes the charged tracks from pileup vertices before jet clustering, thus improving the jet energy resolution at low-p_{T}. The pileup jet ID based on an MVA is developed to remove the pileup jets, which typically lack a hard particle at their core, unlike the hard-scattering jets. Such pileup ID can reject $90 - 95\%$ pileup jets while keeping 99% hard-scattering jets for jets in the $|\eta| < 2.5$ and $p_{\mathrm{T}} < 30$ GeV region [53].

The jet energy corrections at CMS, including jet energy scale (JES) and jet energy resolution (JER), calibrate the jet measurement for both data and Monte-Carlo (MC) simulation. The JES corrects the mean values of the jet four momentum in MC and data, and the JER smears the jet resolution only in MC.

Figure 3.11 shows the workflow of applying JES corrections to data and MC. The pileup corrections are applied to all reconstructed jets to decrease the effect of pileup tracks in the jet cone. Then the simulated particle response correction, as a function of p_{T} and η of the generated jets, is applied to the momentum of the measured jets. The residual corrections as functions of η and p_{T} are only applied to data, and the correction to improve jet flavor tagging is only applied in MC.

The JER correction modifies the resolution of jet momentum distributions in MC. It improves the MC-data agreement of the jet momentum distribution. The JER correction in this analysis follows a hybrid method provided by CMS. The JER correction factor, labeled as c_{JER}, is a scale factor applied to correct the four-momentum of the jets. The

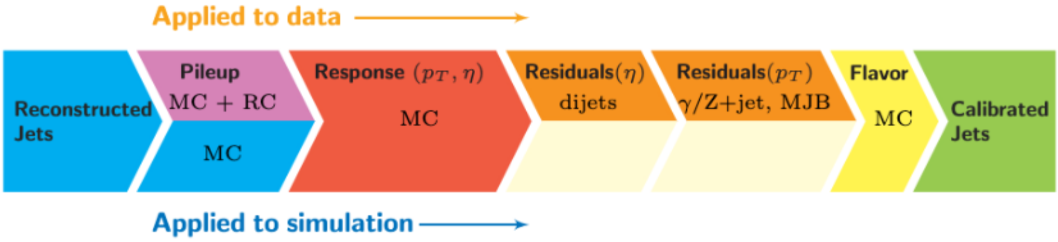

Figure 3.11: Jet energy scale corrections at CMS [56]

method begins with a particle-level matching to the jet. A matching particle-level jet must pass the requirement in Equation 3.9, where p_{ptcl} is the transverse momentum of the particle-level jet for matching, p_{T} is the transverse momentum of the jet to be matched, and σ_{JER} is the relative p_{T} resolution in simulation:

$$\Delta R < R_{\text{cone}}/2, \quad \mid p_{\mathrm{T}} - p_{\mathrm{T}}^{\text{ptcl}} \mid < 3\ \sigma_{\text{JER}}\ p_{\mathrm{T}} \qquad (R_{\text{cone}} = 0.4 \text{ for AK4 jets}) \tag{3.9}$$

If a matching particle-level jet exists, the JER correction can be quantified by equation (3.10), where s_{JER} is the data-to-simulation scale factor found in central produced configuration files:

$$c_{\text{JER}} = 1 + (s_{\text{JER}} - 1)\ \frac{p_{\mathrm{T}} - p_{\mathrm{T}}^{\text{ptcl}}}{p_{\mathrm{T}}} \tag{3.10}$$

If no matching particle-level jet is found, the jet momentum correction factor can be calculated with Equation 3.11, where $N(0, \sigma)$ denotes a random number sampled from a normal distribution with a zero mean and variance σ^2_{JER}:

$$c_{\text{JER}} = 1 + N(0, \sigma_{\text{JER}})\sqrt{max(s^2_{\text{JER}} - 1, 0)} \tag{3.11}$$

3.3.5 B-tagging

The b jet tagging is crucial for establishing the event reconstruction and selection strategy in this analysis. B jets are the hadronization product of b quarks, including b hadrons such as $B^{\pm}$, B^0, B_S, and Λ_b. The b hadrons are heavier than other hadrons, and they have sizable lifetimes of $c\tau = 0.5$ mm. One of their important signatures is the SV reconstructed from their decay products. The p_{T} of the hadron decay products relative to the jet axis, $p_{\mathrm{T}}^{\text{Rel}}$, is an important variable for jet flavor tagging [57]. Most long-lived b hadrons have spin 0, and thus decay isotropically in their rest frame, leading to a large $p_{\mathrm{T}}^{\text{Rel}}$ compared to the light flavor hadrons. A b jet usually contains a large number and various types of tracks, with the presence of leptons from the b hadron decay. The jet flavor classification method used in this analysis is DeepJet [58]. The algorithm utilizes

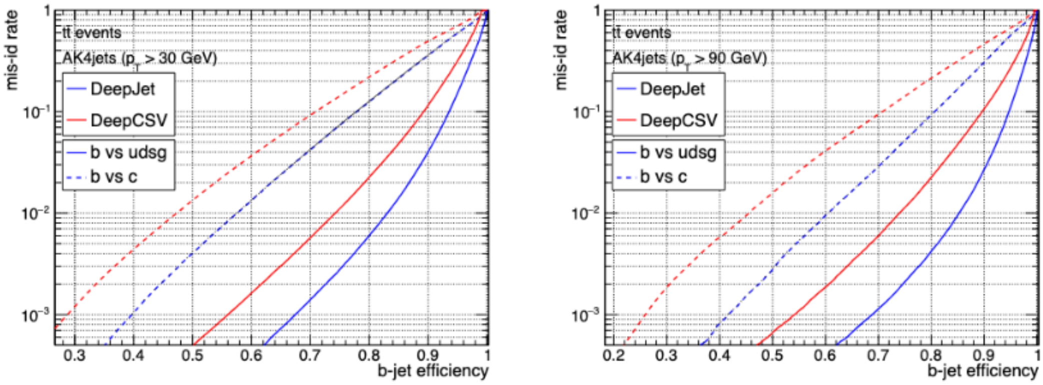

Figure 3.12: DeepJet performance curve: b-jet efficiency vs mis-identification rate [58]

features of b jets and is based on a deep neural network (DNN) trained with approximately 650 input variables, including track multiplicity, jet kinematics, and SV reconstruction information. The training and testing samples are combinations of MC simulations of QCD multi-jet and fully hadronic top-quark pairs.

Figure 3.12 shows the performance of the DeepJet classification algorithm, showing improvement compared to the DeepCSV classifier. This analysis uses the b jet tagging at the medium fixed working point, which has a mis-identification rate of less than 1% and around 80% b-tagging efficiency. To gain a good data-MC agreement, the events in MC are multiplied with a event-based weight $w^{\text{b-tagging}}$, as shown in the following equation.

$$\begin{aligned} w^{\text{b-tagging}} = & \prod_{i=tagged\ jets} \frac{SF_i(p_T,\eta,flavor)\cdot\epsilon_i(p_T,\eta,flavor)}{\epsilon_i(p_T,\eta,flavor)} \\ & \cdot \prod_{j=non\ tagged\ jets} \frac{1-SF_j(p_T,\eta,flavor)\cdot\epsilon_j(p_T,\eta,flavor)}{1-\epsilon_j(p_T,\eta,flavor)} \end{aligned} \tag{3.12}$$

where the scale factor SF_i and the b-tagging efficiency are functions of the tagged jet's p_T, η and flavor. Figures 7.1, 7.2 and 7.3 in the appendix show the b-tagging efficiency maps, individually produced for each data-taking era and jet flavor.

3.3.6 Missing Transverse Momentum

The missing transverse momentum (MET) refers to the total transverse momentum that is not detected in any CMS sub-detectors. It is typically associated to neutrinos that do not interact with the detector materials. The missing transverse momentum is defined as the negative vector sum of the p_T of all the visible particles reconstructed via the PF algorithm, and the missing transverse energy is the magnitude of it:

$$\overrightarrow{p_{\mathrm{T,PF}}^{\mathrm{miss}}} = -\sum_{i=1}^{N_{\mathrm{particles}}} \overrightarrow{p}_{\mathrm{T},i} \;\; ; \tag{3.13}$$

As discussed in the previous section, the JES and JER corrections alter the jet p_{T} values. Therefore, corresponding corrections are required to update the missing transverse momentum, as shown in Equation 3.14.

$$\overrightarrow{p_{\mathrm{T,PF}}^{\mathrm{miss}}} = -\sum_{i=1}^{N_{\mathrm{particles}}} \overrightarrow{p_{\mathrm{T},i}} - \sum_{j=1}^{N_{\mathrm{PF\ jets}}} (\overrightarrow{p_{\mathrm{T,j}}^{\mathrm{corr}}} - \overrightarrow{p_{\mathrm{T},j}}) \tag{3.14}$$

The XY correction is applied to MET ϕ and MET p_{T} in data and MC to mitigate pileup effects and detector modulation. Modulation could arise from detector misalignment, beam spot displacement, and inactive calorimeters. It increases with more pileup interactions. The XY correction modifies the missing p_{T} and missing ϕ values. It is a function of the number of PVs, raw missing p_{T}, and raw missing ϕ. When it applies to data, it also depends on the detector conditions during the data-taking era. The MET ϕ distribution is sinusoidal without the XY correction, and becomes flat after applying it. The correction improves the data-MC agreement in the MET distributions.

CHAPTER 4

Search for singly produced Vector-Like Quarks in the opposite-sign dilepton final state

4.1 Analysis Strategy

This analysis searches for singly produced vector-like top quarks (T) in the decay channel $\mathrm{T} \rightarrow \mathrm{tH}$ (Figure 4.1), focusing on final states with two opposite-sign (OS) leptons, where each lepton could be either an electron or a muon, jets, and missing transverse energy. The search covers the T mass region from 600 GeV to 1200 GeV, where only ”resolved” final states are considered. In a resolved final state, the hadronization products of any two outgoing partons do not merge into a single ”boosted” object. The T signal model is produced under the narrow-width approximation (NWA), requiring a width-to-mass ratio less than 15%. More details about NWA can be found in section 2.2.1. The analysis targets the process $\mathrm{T} \rightarrow \mathrm{tH};\ \mathrm{t} \rightarrow \mathrm{bW} \rightarrow \mathrm{bqq'}$, where the Higgs boson decays to a final state with two opposite-sign leptons, and the top quark decays hadronically. The signature is dominated by $\mathrm{H} \rightarrow \mathrm{WW} \rightarrow \ell\ell\nu\nu$ and $\mathrm{H} \rightarrow \mathrm{WW}; \mathrm{W} \rightarrow (\tau) \rightarrow \ell\nu(\nu)$ decay modes, while $\mathrm{H} \rightarrow \mathrm{ZZ}$ and $\mathrm{H} \rightarrow \tau\tau$ contribute to a small degree. Only the electrons and the muons are considered in the final state, resulting in three different topologies: the dimuon, the dielectron, and the electron-muon final states, accompanied by at least three jets (one of the jets is b-tagged) and missing transverse energy. Table 4.1 lists all the decay modes that contribute to the signal.

The data used in this analysis has an integrated luminosity of 138 fb^{-1}, collected by CMS at $\sqrt{s} = 13$ TeV in the years 2016, 2017 and 2018, referred to as Run 2.

The analysis uses a cut-based approach. The cuts are optimized to suppress the main backgrounds, which are low-mass Drell-Yan production for the $\mu\mu$ and ee channels and

Table 4.1: Higgs boson decay modes contributing to the OS di-lepton signal in the process $\mathrm{pp} \rightarrow \mathrm{Tbq};\ \mathrm{T} \rightarrow \mathrm{t} + \mathrm{H};\ \mathrm{t} \rightarrow \mathrm{b} + \mathrm{W} \rightarrow \mathrm{bqq}'$

H decay mode	Final state		Relative contribution	
			$ee/\mu\mu$	$e\mu$
$\mathrm{H} \rightarrow \mathrm{WW}$		$\rightarrow \ell\ell\nu\nu$	80%	90%
	$\rightarrow \tau\nu_\tau\ell\nu$	$\rightarrow \ell\ell\nu\nu\nu_\tau\nu_\tau$		
	$\rightarrow \tau\nu_\tau\tau\nu_\tau$	$\rightarrow \ell\ell\nu\nu\nu_\tau\nu_\tau\nu_\tau\nu_\tau$		
$\mathrm{H} \rightarrow \tau\tau$		$\rightarrow \ell\ell\nu\nu\nu_\tau\nu_\tau$	5%	10%
$\mathrm{H} \rightarrow \mathrm{ZZ}$		$\rightarrow \ell\ell\nu\nu$	15%	0
	$\rightarrow \tau\tau\nu\nu$	$\rightarrow \ell\ell\nu\nu\nu\nu\nu_\tau\nu_\tau$		
		$\rightarrow \ell\ell\mathrm{qq}'$		
	$\rightarrow \tau\tau\mathrm{qq}'$	$\rightarrow \ell\ell\mathrm{qq}'\nu\nu\nu_\tau\nu_\tau$		

top quark pair ($\mathrm{t\bar{t}}$) production for all channels.

Figure 4.2 shows the dominant production modes of $\mathrm{t\bar{t}}$ at the Leading order (LO) via gluon fusion. The branching ratio of a top quark decaying into a b quark and a W boson is almost 100 %, and the branching ratio of a W boson leptonic decay is 33 %. When two W bosons both decay into leptons and neutrinos, the $\mathrm{t\bar{t}}$ final state includes an opposite-sign lepton pair, jets, and neutrinos, thus becoming a major background in the analysis.

Figure 4.3 shows Z boson production via Drell-Yan (DY) quark-antiquark annihilation at the LO, where the Z boson decays into an opposite-sign (OS) lepton pair. The DY process has a large production cross section, making it a significant background process in the $\mu\mu$ and ee channels.

The discriminating variable in the final signal extraction is the reconstructed mass of the T candidate, labeled as m_{tH}. The reconstructed mass of the fully hadronic decay top quark can be obtained from the tree selected jets. Due to the existence of the two neutrinos in the Higgs boson decay, the Higgs boson cannot be fully reconstructed. This is addressed by a partial reconstruction based on the kinematic properties of the final state.

To maximize analysis sensitivity, the selections are optimized to target a peaking signal shape on a smoothly falling background in the signal region. Considering that the analysis focuses on the T mass points from 600 GeV to 1200 GeV, the signal region is defined as $m_{\mathrm{tH}} \in [400, 1500]$ GeV. Signal models are obtained by fitting the reconstructed T mass distributions from the simulated signal events for each mass point, di-lepton channel, and data-taking era. The background shape in the signal region is determined from the data, with a parameterization motivated by the Monte-Carlo (MC) simulations.

The signal extraction is performed with a binned maximum likelihood fit to the m_{tH} distribution in each category of the data. The signal and background modeling, the

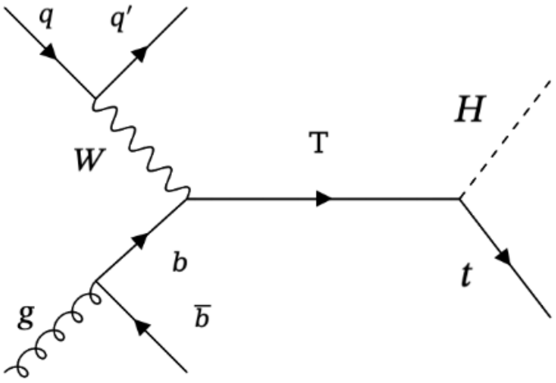

Figure 4.1: Single production of T at LO

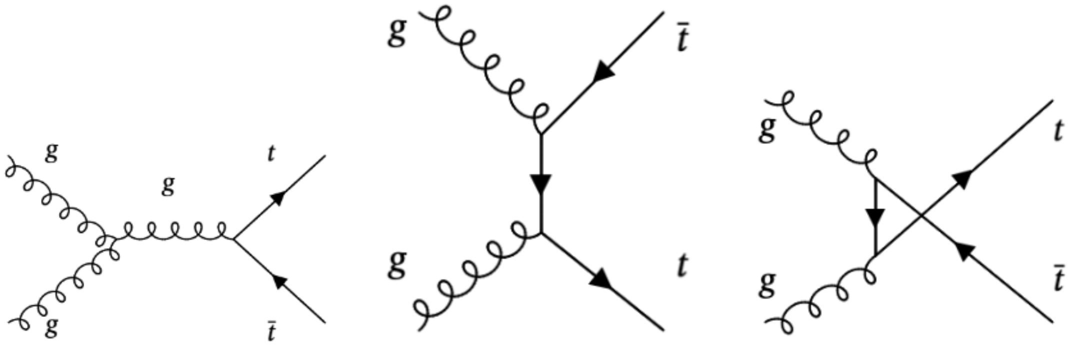

Figure 4.2: $t\bar{t}$ production at LO

systematic uncertainties, and the statistical features are taken into consideration in the final signal extraction. The results can be found in Chapter 5.

4.2 Datasets and Monte Carlo Simulation

4.2.1 Datasets

The analysis in this thesis uses data collected during the LHC Run 2 data-taking period from 2016 to 2018. Different single-lepton datasets with corresponding single-lepton triggers are chosen to maximize the trigger efficiency. For the di-muon channel and the muon-electron channel, the primary datasets named "SingleMuon" are used. For the di-electron channel, the primary datasets named "SingleElectron" or "EGamma" are used. The data eras in CMS during one data-taking year are named with letters in alphabetical order, such as A, B, C, and D. The 2017 B dataset is excluded in the di-electron chan-

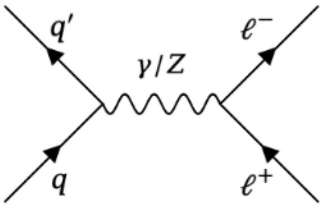

Figure 4.3: DY Z boson production in dilepton final state at LO

nel since there is no proper trigger. The 2016 datasets are divided into "preAPV" and "postAPV" due to the change in silicon strip tracker settings to address issues caused by APV modules. Table 4.2 and 4.3 report the list of datasets used in the analysis, crossing all years and channels.

4.2.2 Monte Carlo Simulation

Monte Carlo (MC) simulation plays a crucial role in physics analysis. In this analysis, the selection strategy is optimized with the signal and background MC. The signal modeling is driven by the signal MC, and the initial background shape is extracted from the background MC. The MC samples are generated using two generators: MadGraph [59] and Powheg. The hadronization, parton showering, and particle decay simulation are processed with Pythia 8 [60]. The detector simulation is provided with GEANT4 [61] [62].

The signal MC pp → Tbq is initially generated at the leading order (LO) with the MadGraph generator (V5.2.6.5). The T is produced via the electroweak (EW) interaction associated with a b quark as shown in the Feynman diagram (Figure 4.1). In the generator, the T decays into a Higgs boson and a top quark, under seven mass hypotheses ranging from $m_{\mathrm{T}} = 600$ GeV to $m_{\mathrm{T}} = 1200$ GeV. The generator-level T width in the signal MC is set to 10 GeV; thus, $\Gamma/m_{\mathrm{T}} \approx 1$ %, which is consistent with the NWA.

The hadronization, parton showering, and the Higgs boson decay are performed by Pythia8.24. Higgs boson decay mode specification and generator-level filters are inserted in the configuration. Higgs bosons from the generator are forced to decay to W boson pairs, Z boson pairs, or τ pairs, since they are the only decay modes significantly contributing to the opposite-sign final state. H → WW is the dominant Higgs boson decay mode in the analysis, while H → ZZ and H → $\tau\tau$ contribute to a smaller degree. The generator-level filter requests at least two leptons (electrons or muons) in one event; each lepton must pass the $p_{\mathrm{T}} > 5$ GeV and $|\eta| < 2.5$ requirements. The motivation of the lepton number requirement is to increase the MC production efficiency. In a typical signal event, there are exactly two leptons in the final state from the Higgs boson decay. Although sometimes additional leptons decaying from b jets are found at the generator level, they usually are soft leptons, with p_{T} below the analysis threshold and beyond the generator-level filter threshold. The existence of those leptons will however change the total number of leptons counted at the generator level. Around 50 % signal events have an additional lepton from the b hadron decay in t → bW. To include those events, it is necessary to ask for at least two leptons at the generator level instead of asking for exactly two leptons. Table 4.4 lists the signal MC samples for all the data-taking eras.

Background MC samples are utilized for selection optimization and initial background shape parameterization. The $\mathrm{t\bar{t}}$ MC samples are generated at next-to-leading order (NLO) using POWHEG, and the Drell-Yan MC samples are generated by MADGRAPH at the leading order. To ensure sufficient MC statistics, the Drell-Yan modeling combines thirteen Drell-Yan MC samples spanning various H_T and lepton pair mass regions, where H_T is the scalar sum of the p_T values of all jets in one event. Since the next-to-leading order for Drell-Yan processes also contributes to the background, a correction factor, calculated with generator-level information, is adopted to scale the event rate, as discussed later. Apart from the main background MC samples, the MC samples for sub-leading background processes are used for background study, namely single top, ttW, ttH, and ttZ. All the background MC samples used in the analysis are listed in Table 4.4, associated with their expected cross sections.

4.3 Events Selection and Signal Reconstruction

4.3.1 Trigger

To achieve high selection efficiencies for target signal decays, the trigger strategy is to combine two lepton triggers in each channel, including a low p_T single-lepton trigger with an isolation requirement and a high p_T single-lepton trigger without an isolation requirement. Table 4.5 lists the trigger combinations used in the $\mu\mu$ and $e\mu$ categories, and Table 4.6 lists the ones used in the ee category.

The p_T threshold for the low-p_T muon trigger is 24 GeV or 27 GeV, depending on the data-taking year. The p_T threshold for the high-p_T muon is 50 GeV in all years. The electron triggers are analogous to the muon triggers. The low p_T single electron triggers with the isolation requirement have p_T thresholds of 27 GeV or 32 GeV. The high p_T triggers without the isolation requirement have p_T thresholds of 115 GeV.

In general, the low p_T triggers increase the signal acceptance of online selection by including signal events with low p_T leptons. The high p_T triggers include high p_T objects, which can fail the isolation requirement applied in the low p_T trigger. As a result, the single lepton triggers and the resulting offline p_T cuts do not entail a significant loss of signal efficiency.

The offline trigger efficiency for this analysis is estimated centrally in CMS via the tag-and-probe technique. The tag-and-probe method is based on a two-object system consisting a ”tag” and a ”probe”. The tag and the probe are produced in pair through

2016 HIPM Dataset	Run–range
/SingleElectron/Run2016B*	273150–275376
/SingleElectron/Run2016C*	275656–276283
/SingleElectron/Run2016D*	276315–276811
/SingleElectron/Run2016E*	276831–277420
/SingleElectron/Run2016F*	277932–278807
Integrated luminosity (SL trigger)	19.67 fb^{-1}
2016 post Dataset	
/SingleElectron/Run2016F*	278769–278769
/SingleElectron/Run2016G*	278820–280385
/SingleElectron/Run2016H*	281613–284044
Integrated luminosity (SL trigger)	16.98 fb^{-1}
2017 Dataset	
/SingleElectron/Run2017C*	299368–302029
/SingleElectron/Run2017D*	302030–302663
/SingleElectron/Run2017E*	303818–304797
/SingleElectron/Run2017F*	305040–306460
Integrated luminosity (SL trigger)	36.75 fb^{-1}
2018 Dataset	
/EGamma/Run2018A*	315257–316995
/EGamma/Run2018B*	317081–318310
/EGamma/Run2018C*	319337–320065
/EGamma/Run2018D*	320500–325175
Integrated luminosity (SL trigger)	59.83 fb^{-1}

Table 4.2: `/EGamma` and `/SingleElectron` primary datasets used for the dielectron channel.

2016 HIPM Dataset	Run–range
/SingleMuon/Run2016B*	273150–275376
/SingleMuon/Run2016C*	275656–276283
/SingleMuon/Run2016D*	276135–276811
/SingleMuon/Run2016E*	276831–277420
/SingleMuon/Run2016F*	277932–278807
Integrated luminosity (SL trigger)	19.67 fb^{-1}
2016 post Dataset	
/SingleMuon/Run2016F*	278769–278808
/SingleMuon/Run2016G*	278820–280385
/SingleMuon/Run2016H*	281613–284044
Integrated luminosity (SL trigger)	16.98 fb^{-1}
2017 Dataset	
/SingleMuon/Run2017B*	297047–299329
/SingleMuon/Run2017C*	299368–302029
/SingleMuon/Run2017D*	302031–302663
/SingleMuon/Run2017E*	303824–304797
/SingleMuon/Run2017F*	305040–306462
Integrated luminosity (SL trigger)	41.48 fb^{-1}
2018 Dataset	
/SingleMuon/Run2018A*	315257–315789
/SingleMuon/Run2018B*	317080–319310
/SingleMuon/Run2018C*	319337–320065
/SingleMuon/Run2018D*	320500–325175
Integrated luminosity (SL trigger)	59.83 fb^{-1}

Table 4.3: `/SingleMuon` primary datasets used for the dimuon and the electron-muon channels.

Signal	
TprimeBToTH_M-[1]_AtLeast2L_LH_TuneCP5_13TeV-madgraph-pythia8	
Background	Cross section (pb^{-1})
TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8	88.29
DYJetsToLL_M-4to50_HT-70to100[2]	314.8
DYJetsToLL_M-4to50_HT-100to200[2]	190.2
DYJetsToLL_M-4to50_HT-200to400[2]	42.27
DYJetsToLL_M-4to50_HT-400to600[2]	4.05
DYJetsToLL_M-4to50_HT-600toinf[2]	1.216
DYJetsToLL_M-50_HT-70to100[2]	140.0
DYJetsToLL_M-50_HT-100to200[2]	139.2
DYJetsToLL_M-50_HT-200to400[2]	38.4
DYJetsToLL_M-50_HT-400to600[2]	5.174
DYJetsToLL_M-50_HT-600to800[2]	1.258
DYJetsToLL_M-50_HT-800to1200[2]	0.56
DYJetsToLL_M-50_HT-1200to2500[2]	0.13
DYJetsToLL_M-50_HT-2500toinf[2]	0.003
ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8	39.65
ttWJets[2]	0.21
ttHToNonbb_M125[2]	0.59
ttZJets[2]	0.86

[1] 600, 700, 800, 900, 1000, 1100, 1200

[2] _TuneCP5_13TeV-madgraphMLM-pythia8

Table 4.4: MC samples names used in the analysis

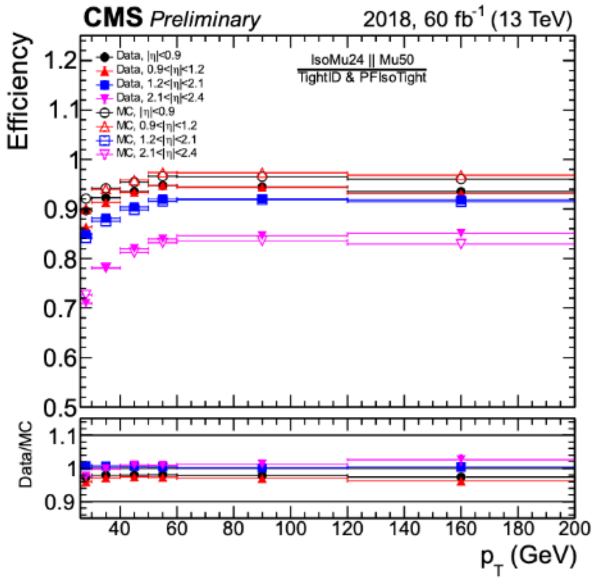

Figure 4.4: Single muon trigger efficiency plot from CMS [63].

Year	single μ trigger
2018	HLT_IsoMu24 or HLT_Mu50
2017	HLT_IsoMu27 or HLT_Mu50
2016	HLT_IsoMu24 or HLT_Mu50

Table 4.5: Triggers for $\mu\mu$ and $e\mu$ channels of each year. "HLT" is short for high-level trigger, the numbers "Mu" are p_{T} thresholds for muons, and "Iso" is short for isolation.

the same process, so they share similar physics properties. Strict selection criteria apply to the tag to ensure purity. The probe objects are used to measure the trigger efficiency. Efficiency is measured as the ratio of probes passing the trigger to the total number of probes. Here, the tag-and-probe system uses muon pairs or electron pairs from Z boson decay produced in the DY process. Figure 4.4 shows an example of the trigger efficiency as a function of muon p_{T} and η. The trigger efficiency is calculated as ε = (IsoMu24 || Mu50)/(Tight ID & PFIsoTight). The denominator indicates that the calculation applies to muons passing both the tight ID and tight isolation requirements.

Year	Triggers
2018	HLT_Ele115_CaloIdVT_GsfTrkIdT or HLT_Ele32_WPTight_Gsf
2017	HLT_Ele115_CaloIdVT_GsfTrkIdT or HLT_Ele32_WPTight_Gsf_L1DoubleEG
2016	HLT_Ele115_CaloIdVT_GsfTrkIdT or HLT_Ele27_WPTight_Gsf

Table 4.6: Triggers for the ee channel of each year. "HLT" is short for high-level trigger, the numbers after "Ele" are p_{T} thresholds for electrons, and "WPTight" refers to the isolation requirement at the tight work point. "CaloIdVT_GsfTrkIdT" indicates that the trigger uses electron ID, which favors high efficiency over purity, based on deposited energy information in ECAL and HCAL and on the matching of ECAL clusters to tracker tracks.

4.3.2 Basic Selection

As the first part of the offline selection, basic selections provide initial suppression of the background and select the physics objects for signal reconstruction. These selections must be tighter than the online selections in triggers. The detector acceptance, due to geometry, should also be taken into consideration. Based on these concerns, the basic selections are developed as follows.

- In each event, there must be exactly two OS leptons and at least three jets. At least one of the jets must be medium b-tagged.

- Muon candidates are required to have $p_{\mathrm{T}} > 30$ GeV, $|\eta| < 2.4$, tight ID, tight isolation, and pass the selection on the significance of the 3D impact parameter ($s_{\mathrm{3D}} < 3$) to ensure production comparable with the primary vertex.

- Electron candidates are required to have $p_{\mathrm{T}} > 35$ GeV, $|\eta| < 2.5$, tight ID, and pass an impact parameter selection applied on the transverse and longitudinal planes: $d_0 < 0.05$ cm and $d_z < 0.1$ cm in the barrel region; $d_0 < 0.1$ cm and $d_z < 0.2$ cm in the endcap. Electrons in the transition region between the barrel and endcap of ECAL ($1.444 < |\eta| < 1.566$) are vetoed to avoid areas with low-quality reconstruction.

- The AK4 jet candidates are required to have $p_{\mathrm{T}} > 30$ GeV, $|\eta| < 2.5$, and pass tight ID. The jets with $p_T < 50$ GeV need to pass the loose pile-up ID. A veto applies to the jets lying in the overlap region, defined as a cone of $\Delta R(j, \mu/e) = \sqrt{\Delta\eta(j, \mu/e)^2 + \Delta\phi(j, \mu/e)^2} < 0.4$ around a selected lepton.

- To remove contributions from di-lepton mass resonances in the low-mass region, the invariant mass of the lepton pair is required to be $m_{ll} > 12$ GeV.

- To further reject background events, the scalar sum of all jets p_{T} in one event is required to satisfy $H_{\mathrm{T}} > 80$ GeV.

- The HEM issue in 2018 is caused by several broken HCAL modules in $-3.2 < \eta < -1.3$ and $-1.57 < \phi < -0.87$ regions. It affects a small fraction of the data collected in 2018. The jets and electrons in the affected η-ϕ region during the affected data-taking period are vetoed.

Figures 4.5, 4.6, and 4.7 show distributions of a few basic kinematic variables from signal and background MC after the basic selection. MC corrections discussed in the following section 4.3.3 are applied to these distributions. All the distributions from background processes in Table 4.4 are included: $\mathrm{t\bar{t}}$, Drell-Yan, $\mathrm{t\bar{t}V}$, and single top productions. The signal distributions are from three representative mass points (600 GeV, 900 GeV, and 1200 GeV) after the basic selection, additionally requiring the leptons to stem from $\mathrm{H} \rightarrow \mathrm{WW}$ decay. In each plot, the total number of events (n) from each signal or background process is scaled according to the integral luminosities (L_{int}) and the theoretical cross-section (σ), defined as $n = L_{int} \cdot \sigma \cdot n_{\text{after selections}}/n_{\text{no selection}}$.

From the distributions of the numbers of jets and b jets, one can infer that the signal process has more jets or b jets than the background processes. In the p_{T} distributions of the leptons and jets, objects from the signal have higher p_{T} than those from the background, because the signal decay products are initially from the heavy particle T. Besides the basic kinematic distributions, two variables, namely the invariant mass of the lepton pair and the ΔR between the two leptons, show characteristic features resulting from the Higgs boson decay [64]. As the Higgs boson has spin 0 and its daughter W bosons have spin 1, the charged leptons from their decays appear preferentially in the same hemisphere [65]. Therefore, the $\Delta R(l,l)$ values from the signal are smaller than those from the background. Due to the same reason, the m_{ll} variable from the signal populates mostly in the low mass region below $m_{\mathrm{H}}/2$. The $\Delta R(l,l)$ variable is sensitive to the Higgs boson p_{T}, resulting in difference between T mass points: In general, the signal distributions from high T mass points have smaller $\Delta R(l,l)$ values than those from low T mass points. Meanwhile, as m_{ll} is Lorentz invariant and depends only on the W decay process, its distribution remains unchanged among the different T mass points.

4.3.3 MC Corrections

To mitigate the MC mismodeling effect and improve data-MC agreement, a series of corrections is applied to all the MC samples used in this analysis. Some corrections modify the values of the kinematic variables, while others are applied as event-based weights.

As introduced in the physics object section 3.3, the JES correction adjusts the mean value of the jet p_{T} distribution, and the JER correction modifies its resolution. As a

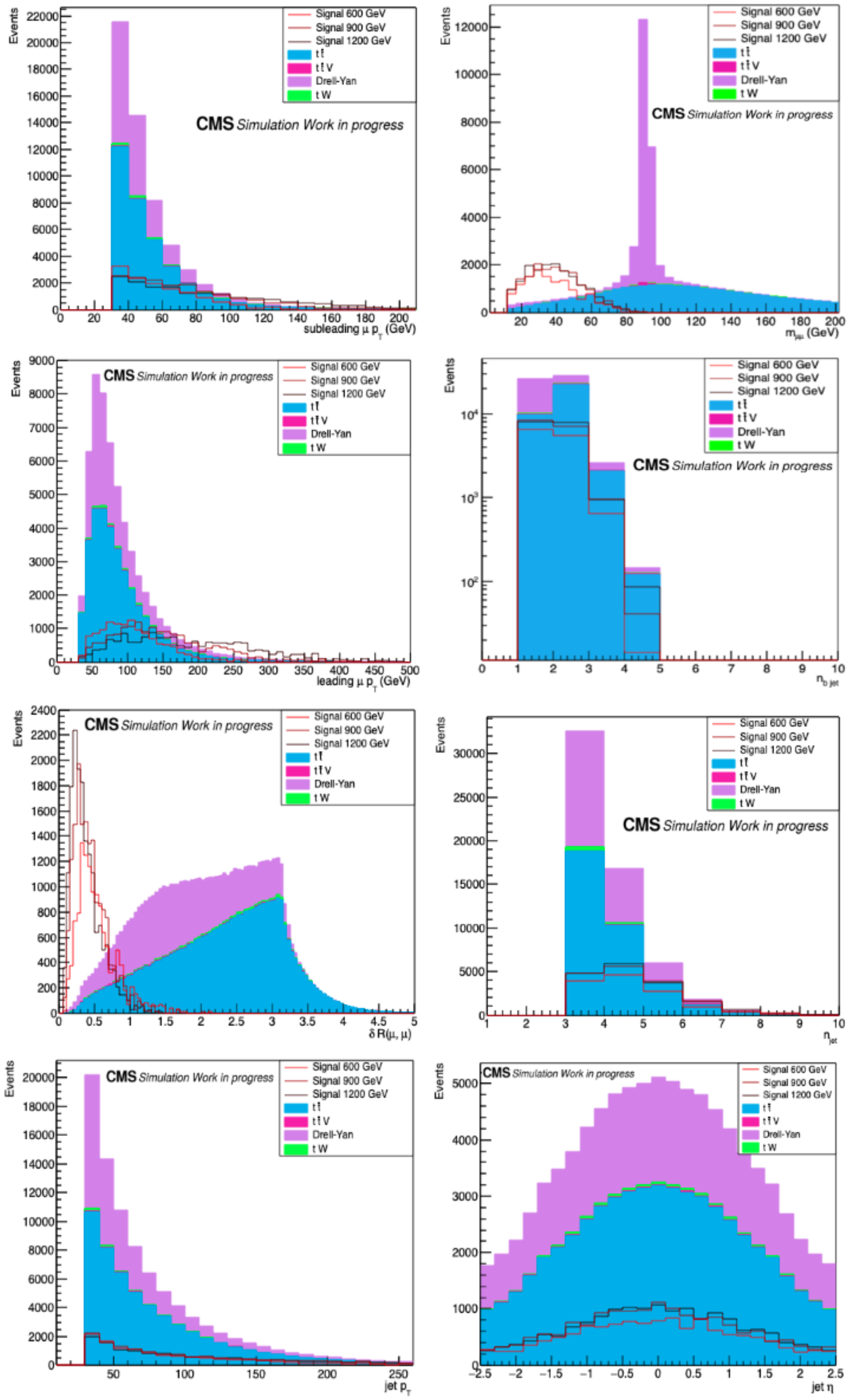

Figure 4.5: Distributions of kinematic variables in the $\mu\mu$ channel after the basic selection.

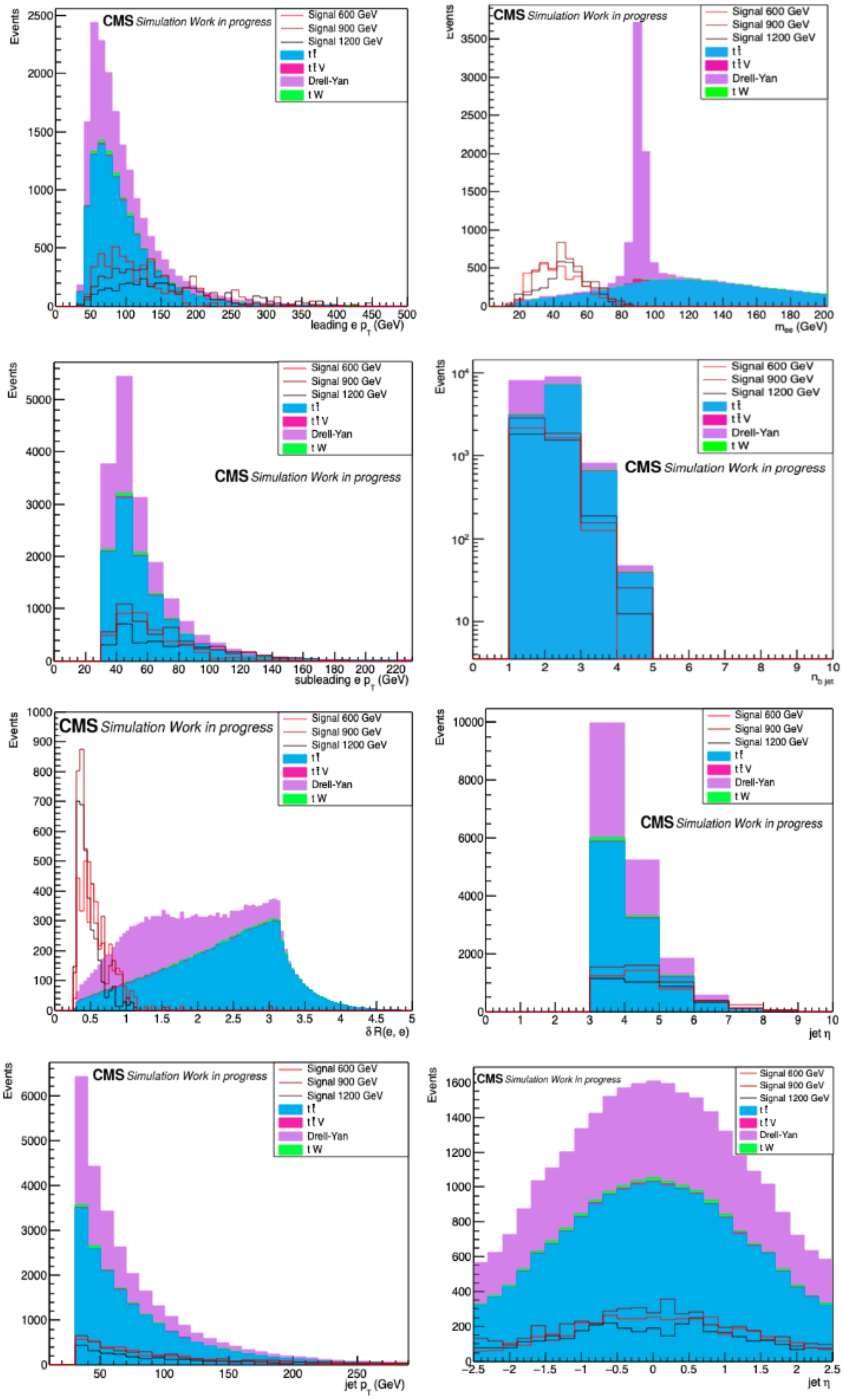

Figure 4.6: Distributions of kinematic variables in ee channel after the basic selection .

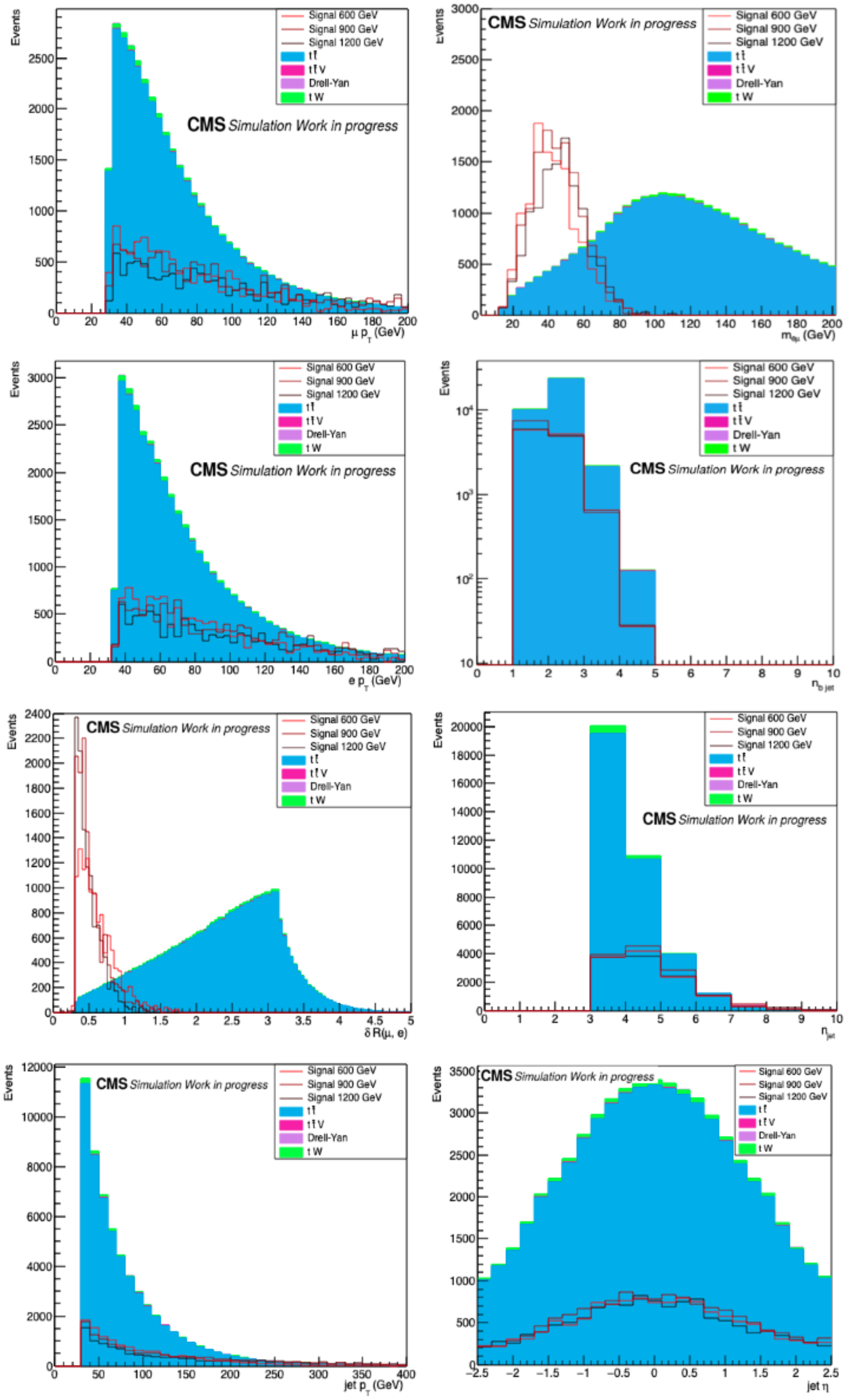

Figure 4.7: Distributions of kinematic variables in $e\mu$ channel after the basic selection.

consequence, the missing transverse momentum $p_{\mathrm{T}}^{\mathrm{miss}}$ is modified to adopt the jet p_{T} corrections. The XY corrections modify MET ϕ and MET $p_{\mathrm{T}}^{\mathrm{miss}}$ to reduce pile-up effect and modulation.

A total event-based weight, which is the product of scale factors reflecting different physics motivations, is computed for each MC event following the expression in Equation 4.1.

$$\begin{aligned} w^{\mathrm{MC}} = \mathrm{SF}^{\text{pile-up}}(n_{\mathrm{Interaction}}) \cdot \mathrm{SF}^{\mathrm{GEN}} \\ \cdot w^{\text{b-tagging}}(p_{\mathrm{T}}^{\text{tagged jet}}, \eta^{\text{tagged jet}}, flav^{\text{tagged jet}}) \\ \cdot \mathrm{SF}^{\mu\ \mathrm{ID}}(p_{\mathrm{T}}^{\mu}, \eta^{\mu}) \cdot \mathrm{SF}^{\mu\ \mathrm{reco}}(p_{\mathrm{T}}^{\mu}, \eta^{\mu}) \cdot \mathrm{SF}^{\mu\ \mathrm{Iso}}(p_{\mathrm{T}}^{\mu}, \eta^{\mu}) \\ \cdot \mathrm{SF}^{e\ \mathrm{ID}}(p_{\mathrm{T}}^{e}, \eta^{e}) \cdot \mathrm{SF}^{e\ \mathrm{reco}}(p_{\mathrm{T}}^{e}, \eta^{e}) \\ \cdot \mathrm{SF}^{\ell\ \mathrm{Trigger}}(p_{\mathrm{T}}^{\ell\ \mathrm{Trig}}, \eta^{\ell\ \mathrm{Trig}}) \cdot \mathrm{SF}^{\mathrm{L1\ prefiring}} \\ \cdot w^{\mathrm{top}\ p_{\mathrm{T}}\ \mathrm{reweighting}}(p_{\mathrm{T}}^{t}, p_{\mathrm{T}}^{\bar{t}}) \end{aligned} \tag{4.1}$$

The calculation of w^{MC} includes all scale factors centrally calculated in CMS, each of which is listed below:

- The pile-up scale factor, $\mathrm{SF}^{\text{pile-up}}$, mitigates the mismodeling of the pile-up distribution in MC simulations. It depends on the number of interactions generated in one MC event.

- A b-tagging weight $w^{\text{b-tagging}}$, as firstly introduced in the physics object section 3.3.5, is computed for each event considering all tagged jets. It uses information from the scale factor maps produced by CMS and the b-tagging efficiency maps produced by analyzers. The full expression of $w^{\text{b-tagging}}$ appears in Equation 3.12.

- The scale factors for muons and electrons include identification, isolation, and reconstruction, reduce the MC simulation bias in each term. They depend on the working points, the data-taking era, and the p_{T} and η of all leptons in a single event.

- The offline trigger scale factors ($\mathrm{SF}^{\ell\ \mathrm{Trigger}}$) are obtained from the trigger scale factor maps. These 2D maps, given as functions of p_{T} and η, result from dividing the trigger efficiency maps from data by those from MC. The p_{T} and η values used to determine the scale factor are from the lepton that actually triggered the event.

- The L1 prefiring scale factor, $\mathrm{SF}^{\mathrm{L1\ prefiring}}$, corrects the gradual timing shift of the ECAL and the muon systems during the 2016 and 2017 data taking. The time shift is the side effect of the L1 rules, which forbid two consecutive bunch crossings from firing. This causes some triggered events to be mistakenly assigned to the previous bunch crossing.

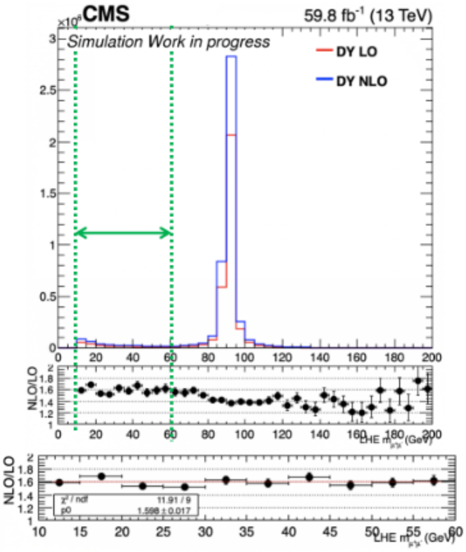

Figure 4.8: Estimate NLO-to-LO correction factor from DY MC

- The top quark p_T reweighting only applies to $\mathrm{t\bar{t}}$ MC samples. The weight is defined as $w^{\text{top } p_\mathrm{T} \text{ reweighting}} = \sqrt{\mathrm{SF}^t(p_\mathrm{T}^t) \cdot \mathrm{SF}^t(p_\mathrm{T}^{\bar{t}})}$, where $\mathrm{SF}^t(p_\mathrm{T}) = e^{0.0615 - 0.0005 \cdot p_\mathrm{T}}$ with p_T values from parton level top quark and anti-top quark. This formula describes the ratio of the top p_T measured in data with respect to the simulation from Powheg and Pythia. The overall effect of top p_T reweighting is a softening of the top p_T distributions in $\mathrm{t\bar{t}}$ MC.

In addition to w^{MC} mentioned above, a self-developed correction factor is applied to the DY MC samples. Although the LO DY samples provide sufficient statistics for the analysis, their cross sections are much less accurate than those from NLO. To address this issue, a correction is introduced to approximate the NLO-level description. This correction is estimated by comparing the generator-level $m_{\ell\ell}$ distributions from the LO and NLO DY MC samples. The upper panel of Figure 4.8 shows the differential cross sections as a function of the generator-level di-muon mass, where NLO appears in blue and LO in red. The lower panels show the ratio of the NLO and the LO distributions. The green arrows point out the region relevant for this analysis, 12 GeV $< m_{\ell\ell} <$ 60 GeV. In the ratio panels, the NLO-to-LO ratio within this region is well described by a constant factor of 1.6 ± 0.02. Since this correction factor can be suboptimal for events outside the range 12 GeV $< m_{\ell\ell} <$ 60 GeV, the following data-MC comparisons will be restricted to it. It is especially important for the $\mu\mu$ and ee channels, where DY dominates the background. As expected, no major difference is observed between data-taking eras or lepton flavors. Therefore, one correction factor of 1.6 ± 0.02 will be applied for all the $\mu\mu$, $e\mu$ and ee channels of all years.

4.3.4 Signal Reconstruction

The reconstructed mass of the T candidate is the main discriminating variable in this analysis. In the target signal process, the top quark decays fully hadronically. Its invariant mass can be inferred from the four momenta of the three selected jets. The Higgs boson in the target signal process decays leptonically, resulting in final states that includes neutrinos. The existence of neutrinos brings challenges to the Higgs boson mass reconstruction.

A χ^2 sorting algorithm is applied to select the kinematically optimal combination of three jet candidates for the top quark reconstruction. This algorithm selects jets by matching their reconstructed mass values to the known W boson and the top quark mass values. It proceeds in two steps. The first step loops over all possible jet pairings to identify the optimal combination for the W boson reconstruction. The combination having the minimal value of $\chi^2_{\mathrm{W}} = (m_{jj} - m_{\mathrm{W}})^2$ is selected, where m_{W} is the W boson mass from PDG [66]. The second step loops over all the remaining jet candidates, and selects the one passing the medium b-tagging working point and giving the minimal value of $\chi^2_{\mathrm{t}} = (m_{bjj} - m_{\mathrm{t}})^2$, where m_{t} is the top quark mass from PDG.

The Higgs boson reconstruction method is developed for the Higgs boson decay modes that share a similar topology with $\mathrm{H} \rightarrow \mathrm{WW} \rightarrow \ell\ell\nu\nu$. The neutrinos in the final state are considered as a single invisible part and later combined with the lepton pair to obtain the reconstructed Higgs boson mass. The transverse momentum of the invisible part is assumed to be equal to the missing p_{T}, referred to as $\overrightarrow{p_T^{inv}} = \overrightarrow{p_{\mathrm{T}}^{\mathrm{miss}}}$.

Because the neutrinos are decay products with high energy, they can be considered approximately collinear with the lepton pair stemming from the same Higgs boson decay. Thus, the azimuthal angle of the invisible part can be approximated as equal to that of the lepton pair: $\theta_{inv} = \theta_{\ell\ell}$. With this approximation, the longitudinal momentum of the invisible part can be calculated as $p_{\mathrm{z}}^{inv} = \frac{p_{\mathrm{T}}^{\mathrm{miss}}}{\tan\theta_{inv}}$. To validate this, Figure 4.9 shows the distributions of $\theta_{\ell\ell} - \theta_{inv}$ for various T mass points of signal in the $\mu\mu$, $e\mu$ and ee channels. The distributions approximately peak around zero. The angular variables are sensitive to the Higgs boson momentum, so distributions from high mass T decays show smaller variances than those from low mass T decays. No clear differences are observed across the different lepton flavors.

The last kinematic variable is the invariant mass of the invisible part. It is approximated by the mean value of its distribution obtained from the simulated signal events, which is found to be $\langle m_{inv} \rangle = 30$ GeV. As shown in Figure 4.10, the m_{inv} distribution is found not to vary with T mass points or channels. Thus, the same mean m_{inv} value applies to the reconstruction of all Higgs boson candidates in this analysis. The reconstructed Higgs boson mass is inferred from the four momenta of the OS leptons and the invisible part.

Figure 4.11 shows the distribution of the Higgs boson mass reconstructed using leptons from the $\mathrm{H} \rightarrow \mathrm{WW}$ decay after the basic selection with the signal MC. The distribution of

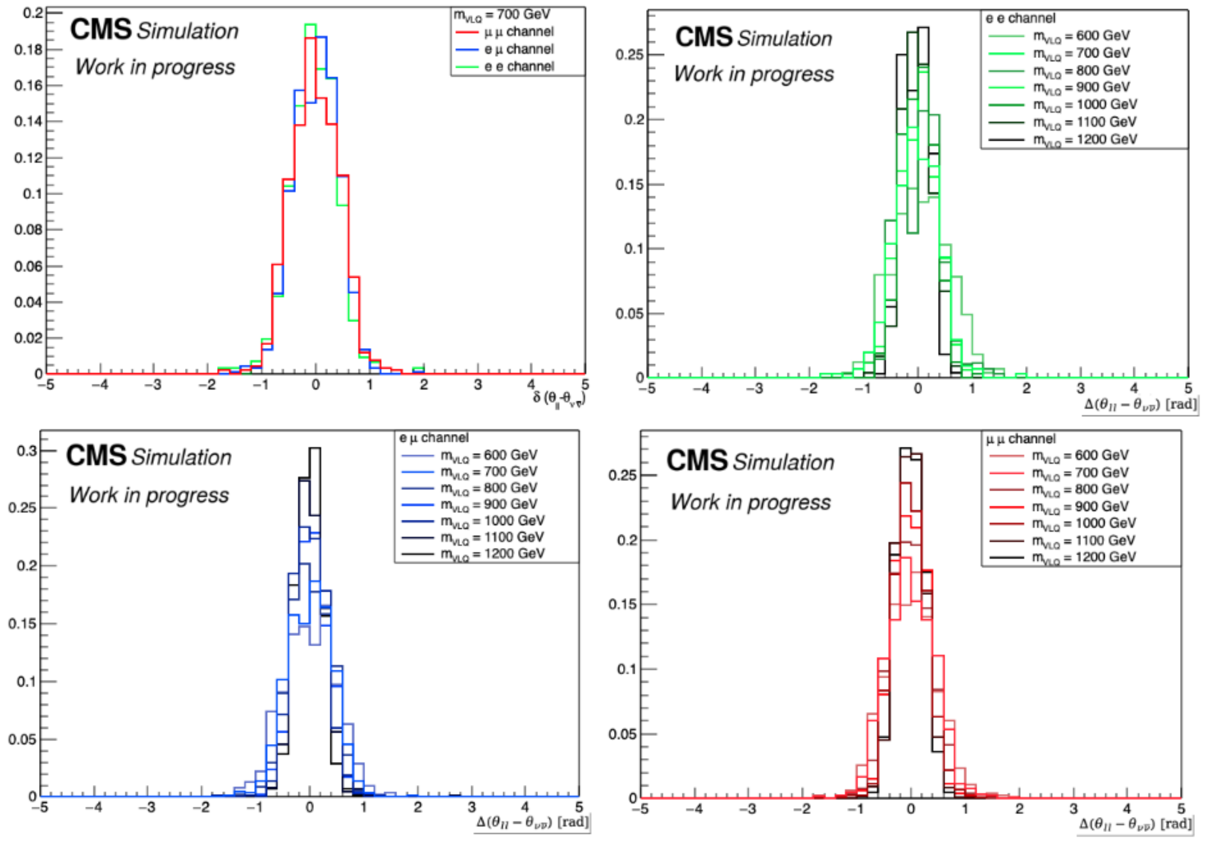

Figure 4.9: Distributions of $\theta_{\ell\ell} - \theta_{\nu\bar{\nu}}$ in the $\mu\mu$, $e\mu$ and ee channels for various T mass points. The unit on the abscissa is radians (rad).

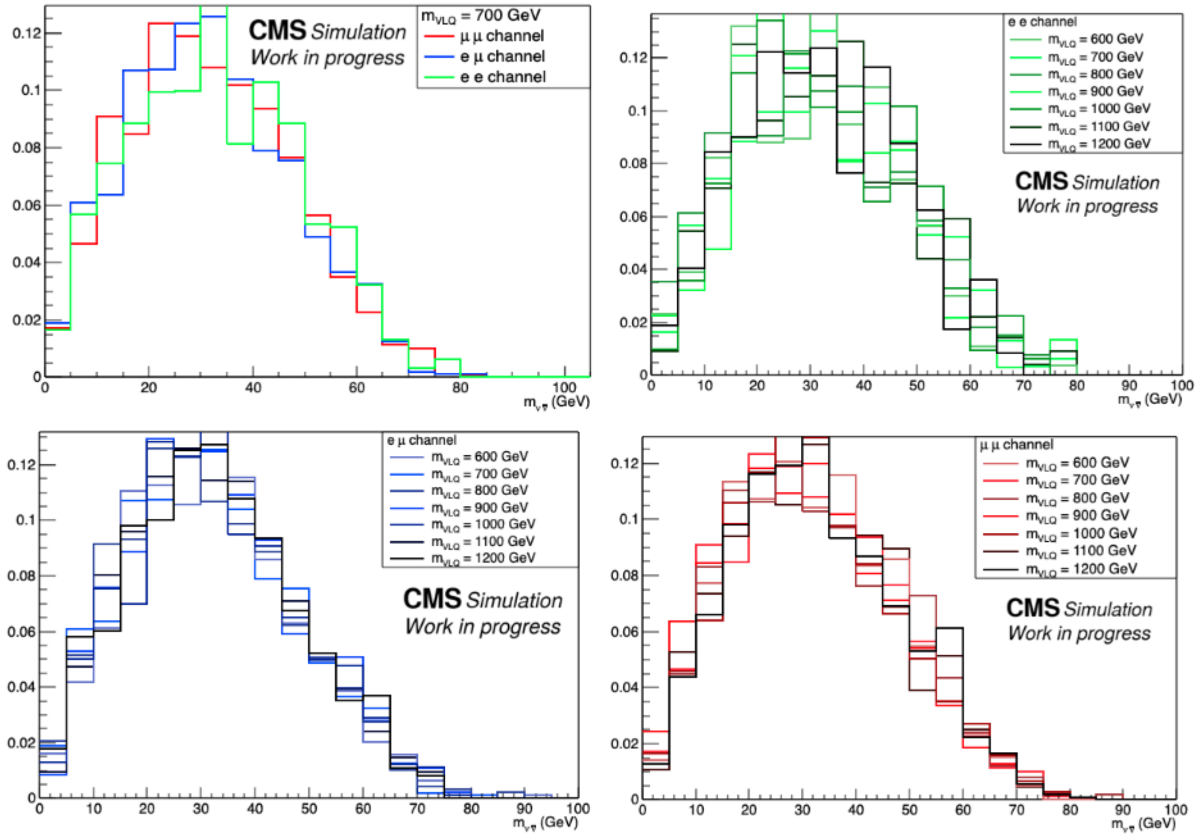

Figure 4.10: $m_{\nu\bar{\nu}}$ distributions in the $\mu\mu$, $e\mu$ and ee channels and T mass points

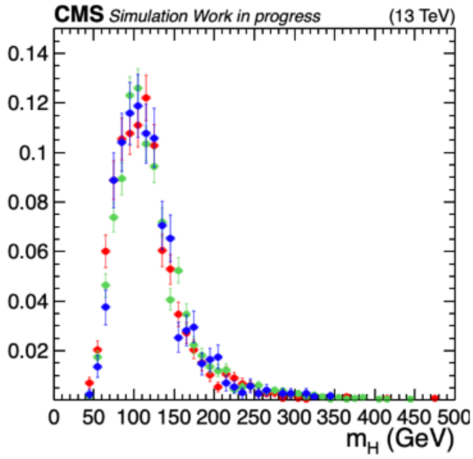

Figure 4.11: Reconstructed Higgs boson mass from signal MC

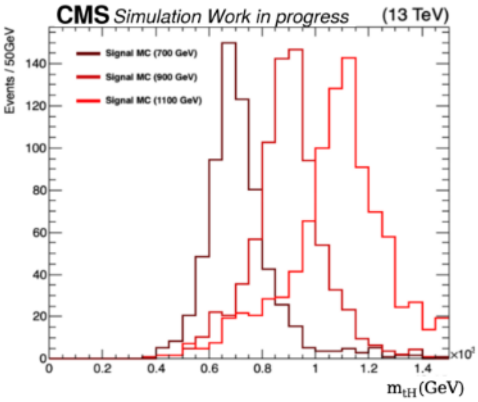

Figure 4.12: Reconstructed T mass from signal MC

the reconstructed Higgs boson mass peaks near the nominal Higgs boson mass at 125 GeV. Figure 4.12 shows the reconstructed T mass distribution from signal MC, using the three example T mass points. All the signal distributions peak around the nominal T mass values, showing good resolution. These reconstructed T mass histograms will be used later in signal modeling.

4.3.5 Main Selection

The main selection criteria detailed in this section are designed based on the differences in kinematics between the signal and background processes, and optimized using the Punzi significance [67], which is defined as $\epsilon/\sqrt{1+B}$, where ϵ is the signal selection efficiency, and B is the numbers of background events passing the selection. The key point of the main selection optimization is to achieve good sensitivity by discriminating the signal from the background. Meanwhile, it is also crucial to ensure that the analysis is orthogonal to other similar analyses in view of potential future combinations. Before introducing the main selection criteria, a study of sub-processes comprising the signal is given below.

Sub-processes in the signal

The study utilizes the generator-level information of the leptons passing the basic selection and the $m_{\ell\ell} < 60$ GeV selection (see later for this selection cut). The signal MC generates the pp → Tbq process, including all the decay modes of the top quark and the Higgs boson. The branching ratios of each decay mode of the top quark and the Higgs boson follow the SM predictions. All the sub-processes contributing to the signal region are listed below. Table 4.7 shows the percentage for each of them in the $\mu\mu$, $e\mu$ and ee channels. Figure 4.13 visualizes the components in the signal MC. The sum of the known sub-processes percentages is 100%.

- $\mathrm{T} \rightarrow \mathrm{tH};\ \mathrm{t} \rightarrow \mathrm{bW} \rightarrow \mathrm{bqq};\ \mathrm{H} \rightarrow \mathrm{WW} \rightarrow \ell\ell\nu\nu$. This process dominates the signal modeling in all channels: around 70% in the $\mu\mu$ and ee channels; 80% in $e\mu$ channel. The selection strategy is primarily optimized for this process, and the T mass reconstruction method benefits it the most.

- $\mathrm{T} \rightarrow \mathrm{tH};\ \mathrm{t} \rightarrow \mathrm{bW} \rightarrow \mathrm{bqq};\ \mathrm{H} \rightarrow \mathrm{WW};\ \mathrm{W}(\rightarrow \tau\nu_\tau) \rightarrow \ell\nu\nu_\tau\nu_\tau$. In this process, a W boson could decay into a τ lepton, then later decays into a muon or electron and two neutrinos. The process exhibits very similar final states and kinematic behaviors compared to the dominant signal process.

- $\mathrm{T} \rightarrow \mathrm{tH};\ \mathrm{t} \rightarrow \mathrm{bW} \rightarrow \mathrm{bqq};\ \mathrm{H} \rightarrow \tau\tau \rightarrow \ell\ell\nu\nu\nu_\tau\nu_\tau$. This process contributes to all channels. Thanks to the topology of its final state, the T mass reconstruction method is valid for it.

- $\mathrm{T} \rightarrow \mathrm{tH};\ \mathrm{t} \rightarrow \mathrm{bW} \rightarrow \mathrm{bqq};\ \mathrm{H} \rightarrow \mathrm{ZZ} \rightarrow \ell\ell\nu\nu/\mathrm{qq}$. Since the lepton pair is from the Z boson decay, this process only contributes to the $\mu\mu$ and ee channels. Its topology is slightly different from other sub-processes. As a result, the mean value of the reconstructed T mass from this process is slightly different from that in the $\mathrm{H} \rightarrow \mathrm{WW}/\tau\tau$ processes.

- $\mathrm{T} \rightarrow \mathrm{tH};\ \mathrm{t} \rightarrow \mathrm{bW} \rightarrow \mathrm{b}\ell\nu;\ \mathrm{H} \rightarrow \mathrm{WW} \rightarrow \ell\nu\mathrm{qq}$. In this process, one lepton comes from the top quark decay. The final state has a different topology compared to the main signal process, causing a bias on the mass reconstruction. Its contribution is less than 1% after the $m_{\ell\ell}$ selection, and is fully removed after all main selections are applied. This analysis is thus orthogonal to other analyses targeting this channel.

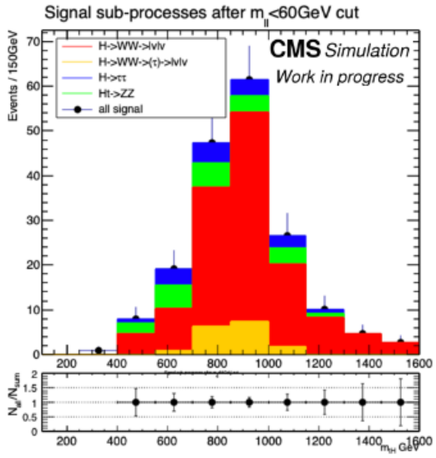

Figure 4.13: Combinations of sub-processes in the signal MC, shown for $m_{\mathrm{T}} = 800$ GeV.

Process in signal MC	Inclusive signal MC[1]		
T → tH	$\mu\mu$	ee	eμ
t → bW → bqq; H → WW → $\ell\ell\nu\nu$	70%	72%	80%
t → bW → bqq; H → WW; W(→ $\tau\nu_\tau$) → $\ell\nu\nu_\tau\nu_\tau$	9%	13%	9%
t → bW → bqq; H → ZZ → $\ell\ell\nu\nu$/qq	13%	8%	0%
t → bW → bqq; H → $\tau\tau$ → $\ell\ell\nu\nu\nu_\tau\nu_\tau$	6%	6%	8%
t → bW → b$\ell\nu$; H → WW → $\ell\nu$qq	< 1%		

Table 4.7: Signal processes and their corresponding percentages in the signal MC samples.

Main selection

The main selection in this analysis consists of four individual selections, which are applied sequentially to the analysis after passing the basic selection. Figure 4.14, 4.15, and 4.16 shows the distributions of four variables used in the main selection. Each subplot shows the distribution after all previous cuts have been applied.

- The first selection requires $m_{\ell\ell} < 60$ GeV, where $m_{\ell\ell}$ is the invariant mass of the OS lepton pair. Due to the property of the H → WW process, the $m_{\ell\ell}$ distribution from the signal dominates the lower region comparing to the distribution from $\mathrm{t\bar{t}}$ process. Moreover, the cut reduces the Drell-Yan process whose distribution peaks near the Z boson mass in the $\mu\mu$ and ee channels. Both main background processes are effectively reduced after this selection. Since $m_{\ell\ell}$ distribution remains unchanged cross the T mass points, the selection works well in the entire search range.

- Secondly, the selection $\Delta R_{\min}(b^t, \ell) > 2$ is applied, where b^t is the b-tagged jet used for the top quark reconstruction. In the signal process, the lepton and the b-tagged jet stem from different mother particles. On the other hand, in the $\mathrm{t\bar{t}}$ process, the lepton and the b-tagged jet emerge from the same top quark decay. This requirement will thus reduce the $\mathrm{t\bar{t}}$ background. Figure 4.17 shows that the variable $\Delta R_{\min}(b^t, \ell)$ is not correlated with the reconstructed T mass m_{tH}. Thus, requiring $\Delta R_{\min}(b^t, \ell) > 2$ does not sculpt the background shapes, making it a good selection criterion.

- The third selection is $\Delta R(b^t, \mathrm{W}^t) < 2.5$, where W^t is reconstructed from the two jets forming the W boson candidate. In signal, the hadronically decaying top quark originates from a heavy mother particle, so its decay products are naturally collimated. There is no hadronically decaying top quark in any of the main background processes. This selection thus suppresses both $\mathrm{t\bar{t}}$ and Drell-Yan background processes.

- The last selection is $S_{\mathrm{T}} > f_4(m_{t\mathrm{H}})$, where $f_4(m_{t\mathrm{H}})$ is shown in Equation 4.2. S_{T} is defined as the scalar sum of the Higgs boson p_{T} and the top quark p_{T}. The T particle is a heavy resonance, so its decay products tend to have much higher p_{T} than those from the background processes. However, the reconstructed T mass is correlated with the p_{T} of T decay products, and thus with S_{T}, as visible in Figure 4.18. Setting a fixed threshold for S_{T} would sculpt the falling background shape and create a background bump that brings difficulty in signal extraction and decreases analysis sensitivity. To mitigate this effect, the S_{T} threshold is chosen as a function of m_{tH} instead of a fixed value.

$$f_4(m_{\mathrm{tH}}) = 72.9 + 0.18 m_{\mathrm{tH}} + 6.07 \cdot 10^{-4} m_{\mathrm{tH}}^2 - 3.47 \cdot 10^{-7} m_{\mathrm{tH}}^3 \tag{4.2}$$

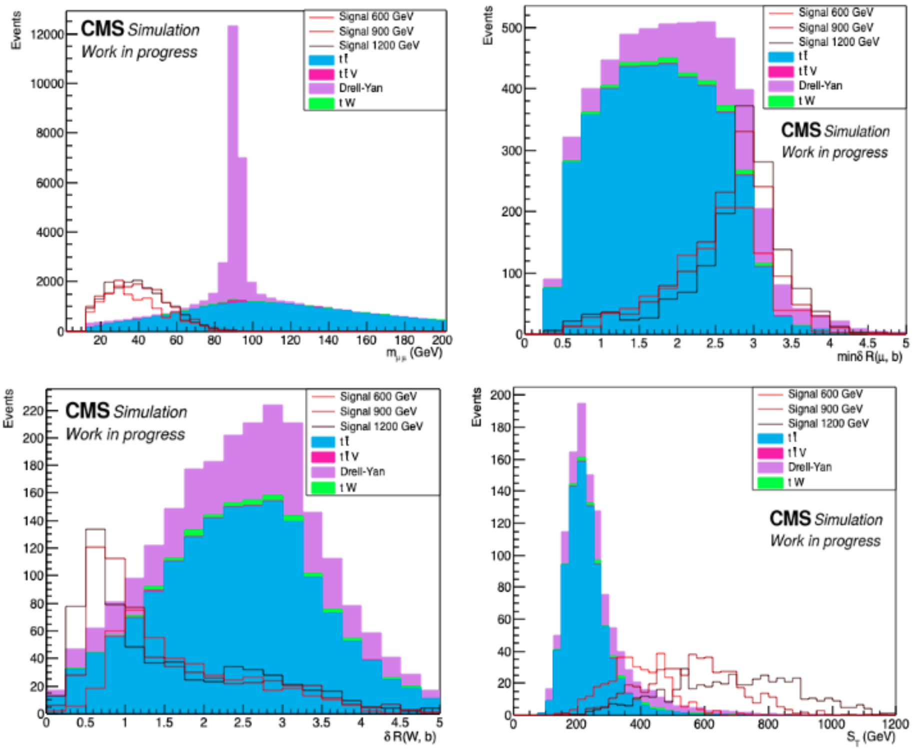

Figure 4.14: The variables used for main selections in $\mu\mu$ channel. The plots show the distributions of $m_{\ell\ell}$, $\delta R_{\min}(b^{\mathrm{t}}, \ell)$, $\Delta R(b^{\mathrm{t}}, W^{\mathrm{t}})$ and S_{T} before applying requirements on them.

> This sliding S_{T} selection maintains a constant fraction of selected events spanning the whole signal region. This constant fraction is optimal as determined by Punzi significance, using the signal ($m_{\mathrm{T}} = 700$ GeV) and the background MC. The S_{T} function is parameterized by fitting the S_{T} points from background MC (Figure 4.19). Thus, the background selection efficiency becomes independent of m_{tH}. Figure 4.20 shows the m_{tH} background shapes before and after the S_{T} selection. The background shape after the selection remains the same as the one before, showing a smooth failing shape in the signal region. Most importantly, the background is significantly suppressed by the S_{T} selection.

To show the influence of each selection criterion for each channel, the cut-flow tables are presented here (Tables 4.8, 4.9, and 4.10). Each cut-flow table begins with the event yield after the basic selection and ends with the number of events retained after the full selection. The percentage retained of each background component is also computed to further motivate the selection criteria. In the signal columns, several extra lines show the effect of the trigger selection on the signal efficiency. The background is derived from

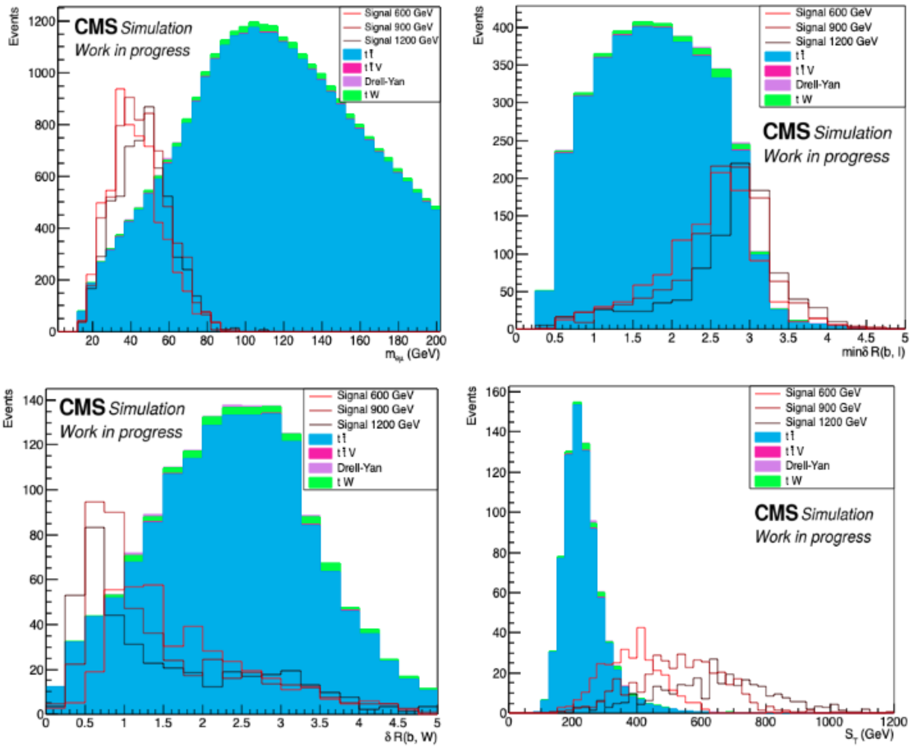

Figure 4.15: The variables used for main selections in $e\mu$ channel. The plots show the distributions of $m_{\ell\ell}$, $\delta R_{\min}(b^{\mathrm{t}}, \ell)$, $\Delta R(b^{\mathrm{t}}, W^{\mathrm{t}})$ and S_{T} before applying requirements on them.

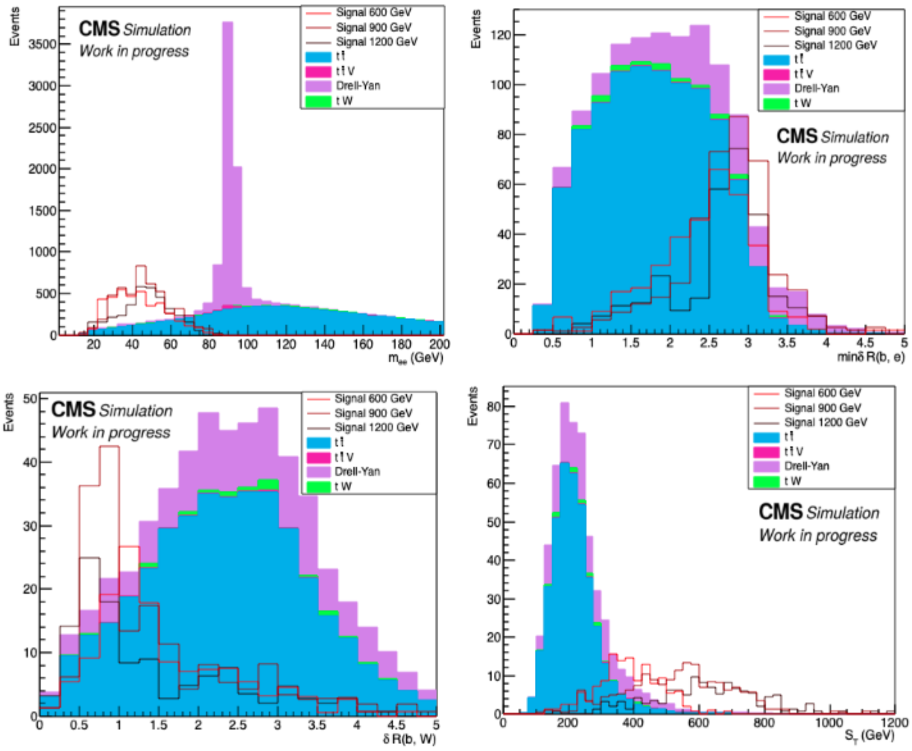

Figure 4.16: The variables used for main selections in ee channel. The plots show the distributions of $m_{\ell\ell}$, $\delta R_{\min}(b^{\mathrm{t}}, \ell)$, $\Delta R(b^{\mathrm{t}}, W^{\mathrm{t}})$ and S_{T} before applying requirements on them.

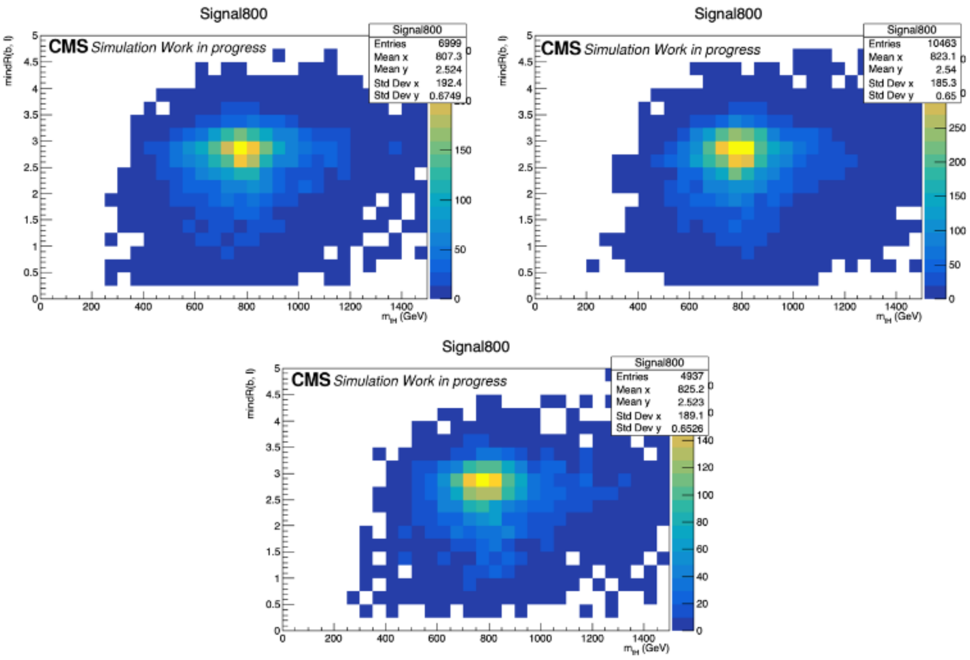

Figure 4.17: Two-dimensional distributions: $\Delta R_{\text{min}}(b^t, \ell)$ and $m_{t\text{H}}$, showing the $\mu\mu$ (upper left), $e\mu$ (upper right) and ee (bottom) channels.

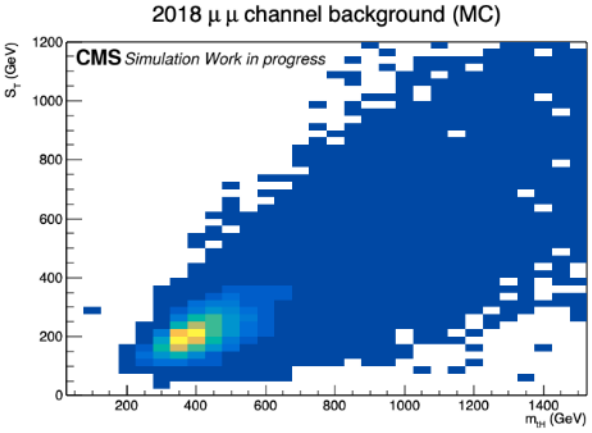

Figure 4.18: Two-dimensional distribution used to determine S_{T} cut function: S_{T} and $m_{t\text{H}}$ in 2018 $\mu\mu$ channel from the merged background MC.

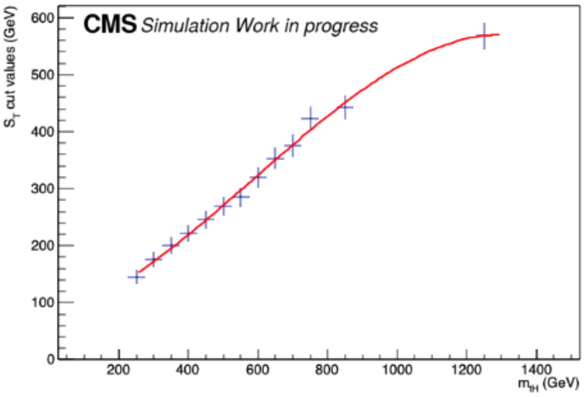

Figure 4.19: The fit used to obtain the $f_{\mathrm{cut4}}(m_{t\mathrm{H}})$ function in S_{T} cut.

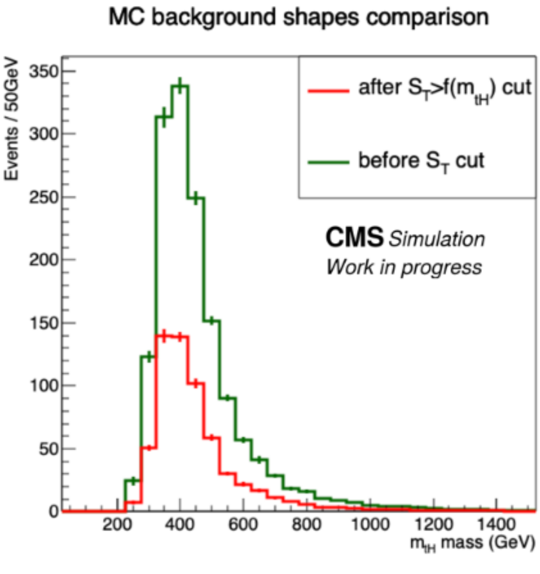

Figure 4.20: m_{tH} distributions of 2018 $\mu\mu$ channel from background MC before and after the sliding S_{T} cut.

cuts	all background[1]	DY	$t\bar{t}$	GEN Signal[2] (900 GeV)
no cut	-	-	-	120
trigger	-	-	-	98%
basic cuts	45288	18890	26398	94%
$m_{\ell\ell} < 60$ GeV	10.1%	4.3%	14.3%	84%
$\delta R_{\min}(b^{\mathrm{t}}, \ell) > 2$	4.5%	3.0%	5.6%	70%
$\delta R(b^{\mathrm{t}}, W^{\mathrm{t}}) < 2.5$	2.3%	1.3%	2.9%	57%
S_{T} cut	1.1%	0.83%	1.3%	44%
full selection	499	157	342	53

[1] Only $t\bar{t}$ and DY are taken into account

[2] Requesting two muons from Higgs and T decays. The signal event rates are scaled according to 1 pb of cross-section and the integrated luminosity of the 2018 dataset.

Table 4.8: Cut-flow table for the 2018 $\mu\mu$ channel

the MC samples scaled to the 2018 integrated luminosity. Additionally, the signal events in these tables are from the four known sub-processes in T → tH; t → bW → bqq, as introduced in section 4.3.5. The signal event rates are normalized to 1 pb of cross-section and the 2018 integrated luminosity. The selection efficiency does not significantly change among T mass points or data-collecting eras.

Figure 4.21 shows the distributions of the discriminating variable m_{tH} in the three channels after all selections. The background is obtained from MC samples scaled with the theoretical cross-sections and the 2018 luminosity. The signal distributions corresponds to three representative T mass points. Each signal is scaled with a cross-section of 5 pb instead of 1 pb to improve the presentation. Thanks to the well-designed selection criteria and mass reconstruction method, the signal shapes peak over the smoothly falling background in each channel in the signal region.

4.3.6 Data-MC Comparisons

The MC-based background modeling is used for selection optimization and initial parameterization of the background. To validate this approach, MC-data comparisons are performed in the $\mu\mu$, $e\mu$ and ee channels of Run 2. For the $e\mu$ channel, the comparison is done after the basic selection. For the $\mu\mu$ and ee channel, due to the NLO-to-LO DY weighting, it is done after the basic selection and the $m_{\ell\ell} < 60$ GeV selection. In these plots, the signal is negligible compared to the enormous background without applying all selections.

cuts	all background[1]	$t\bar{t}$	GEN Signal[2] (900 GeV)
no cut	-	-	195
trigger	-	-	90%
basic cuts	33459	33417	84%
$m_{\ell\ell} < 60$ GeV	12.6%	12.5%	75%
$\delta R_{\mathrm{min}}(b^{\mathrm{t}}, \ell) > 2$	5.1%	5.0%	65%
$\delta R(b^{\mathrm{t}}, W^{\mathrm{t}}) < 2.5$	2.7%	2.7%	55%
S_{T} cut	1.3%	1.3%	44%
full selection	437	432	86

[1] Only $t\bar{t}$ is taken into account

[2] Requesting an electron and a muon from Higgs and T decays. The signal event rates are scaled according to 1 pb of cross-section and the integrated luminosity of the 2018 dataset.

Table 4.9: Cut-flow table for the 2018 $e\mu$ channel

cuts	all background[1]	DY	$t\bar{t}$	GEN Signal[2] (900 GeV)
no cut	-	-	-	125
trigger	-	-	-	89%
basic cuts	20383	8533	11850	60%
$m_{\ell\ell}$ <60 GeV	7.9%	3.6%	11.1%	53%
$\delta R_{\mathrm{min}}(b^{\mathrm{t}}, \ell) > 2$	3.6%	2.6%	4.4%	46%
$\delta R(b^{\mathrm{t}}, W^{\mathrm{t}}) < 2.5$	2.0%	1.2%	2.5%	38%
S_{T} cut	1.1%	0.85%	1.2%	33%
full selection	221	73	148	41

[1] Only $t\bar{t}$ and DY are taken into account

[2] Requesting two electrons from Higgs and T decays. The signal event rates are scaled according to 1 pb of cross-section and the integrated luminosity of the 2018 dataset.

Table 4.10: Cut-flow table for the 2018 ee channel

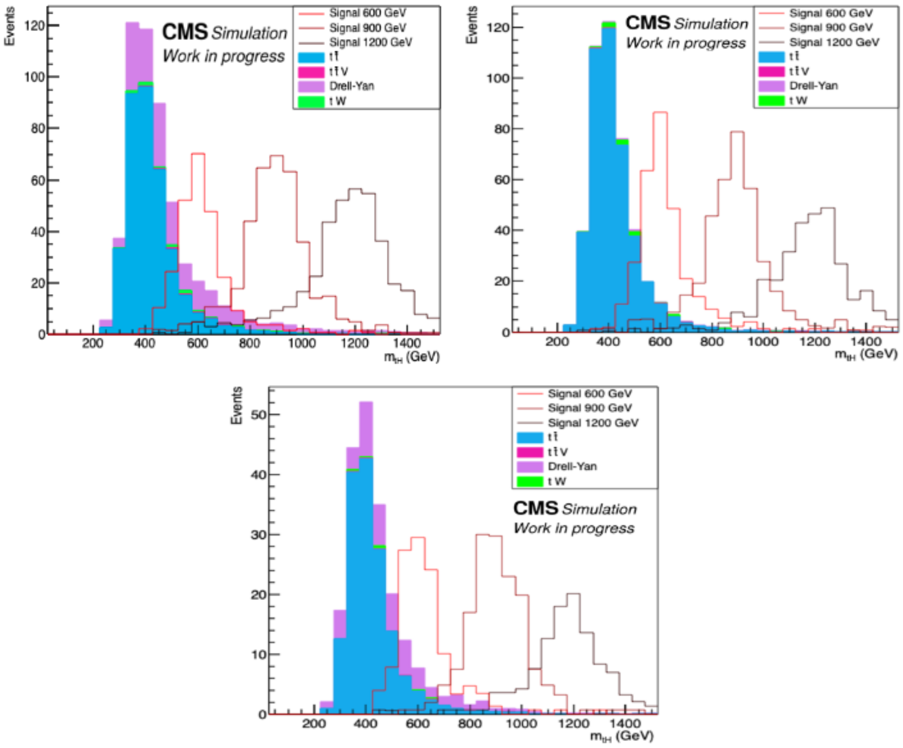

Figure 4.21: $m_{t\mathrm{H}}$ distributions of the the $\mu\mu$ (upper left), $e\mu$ (upper right) and ee (bottom) channels after all selections.

Figure 4.22, 4.23, and 4.24 show the data-MC distributions of $m_{\ell\ell}$, $\Delta R(\ell,\ell)$, $\Delta R_{\min}(b^t,\ell)$, $\Delta R(b,\mathrm{W})$, S_T, and missing p_{T}. The upper panels display the stacked MC distributions and the data points, while the lower panels show the ratios of data to total MC background. The error bar of each data point shows the statistical uncertainties, defined as the square root of the data events in each bin. The gray shadings represent the combination of systematic (see the next section for more details) and statistical uncertainties. Different background processes play different roles in these distributions: $\mathrm{t\bar{t}}$ and DY dominate the background, single top makes a minor contribution, and $\mathrm{t\bar{t}V}$ and VV are negligible compared to the other processes. All the corrections mentioned in Section 4.3.3 are applied to the MC samples. In general, the distributions show good data-MC agreements within uncertainties.

4.4 Signal and Background Modeling

In the signal region, the observed data consist of both background and potential signal. To extract the signal from data, it is essential to build accurate models for both the signal and the background. Signal models aim to describe the signature of the search target T. Background models are built to predict the background distributions in the signal region. In this analysis, the signal models are obtained by fitting the m_{tH} distribution from signal MC. The background models are derived from data, with a parameterization motivated by MC. Before building the models, all corrections are first applied to mitigate the biases from detector effects and theoretical calculations.

4.4.1 Signal Modeling

The signal models are obtained by fitting the $m_{t\mathrm{H}}$ distributions from signal MC after the full selection.

The key features of a signal shape are from the theoretical prediction of the T particle. The T mass hypothesis corresponds to the mean value of a peaking signal shape, and the NWA results in a small width compared to the experimental resolution.

The experimental resolution from detector effects significantly modify the signal shapes. By using fully reconstructed events simulated considering the CMS detector conditions, the detector effects are largely accounted for. Applying all the corrections mentioned in Section 4.3.3 further improves the simulation of the signal shapes.

Considering all the effects in the analysis procedure, as shown in Section 4.12, the signal distribution results in a smeared and shifted peaking shape compared to its original shape from the theoretical prediction.

To precisely model the signal behavior, individual fits are performed on signal distributions from all T mass hypotheses, data-taking eras, and channels ($\mu\mu$, $e\mu$, and ee). The probability density function (pdf) is chosen to be a double-sided Crystal Ball (DSCB)

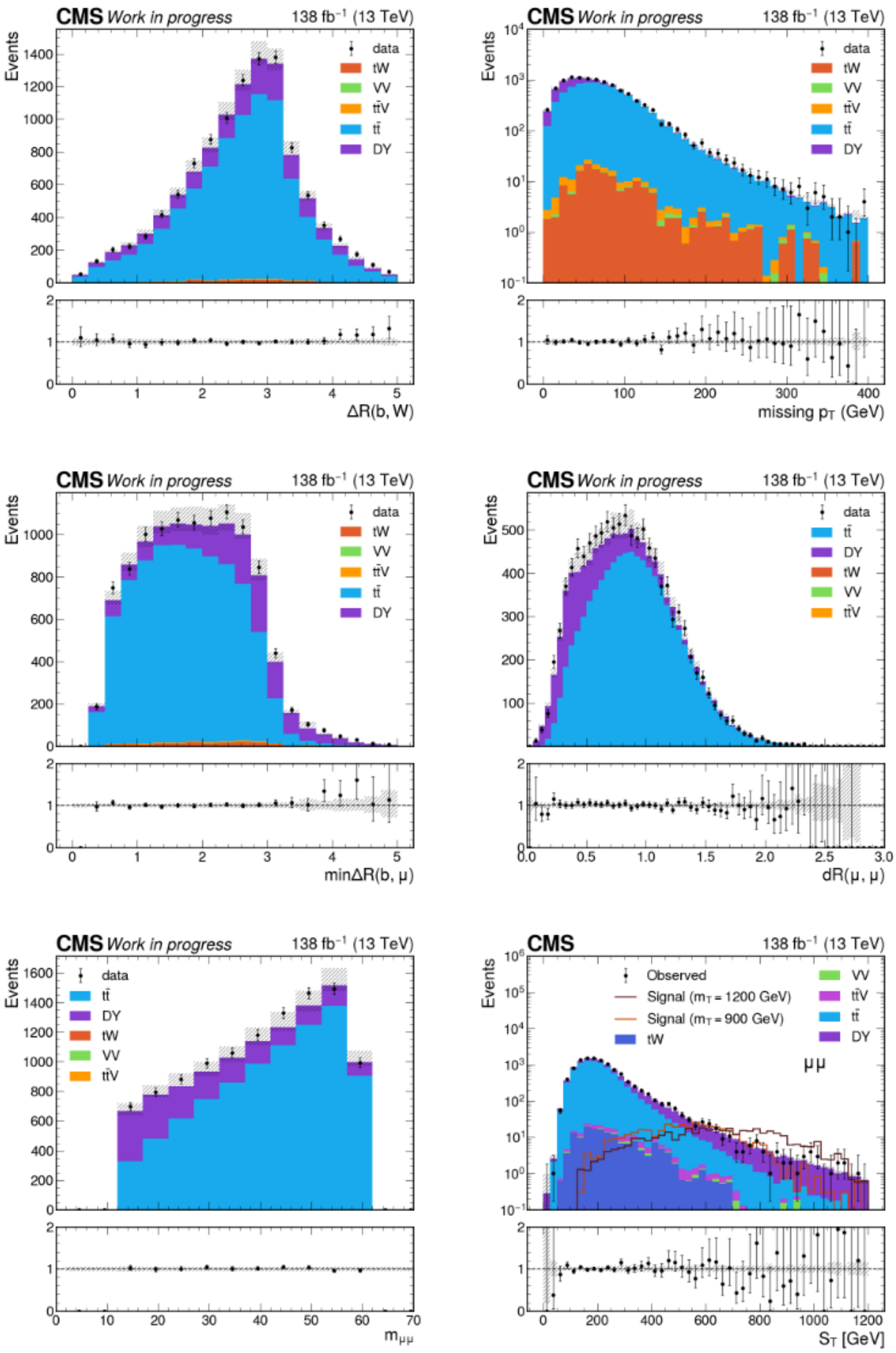

Figure 4.22: Data-MC comparison for the $\mu\mu$ channel in full Run 2.

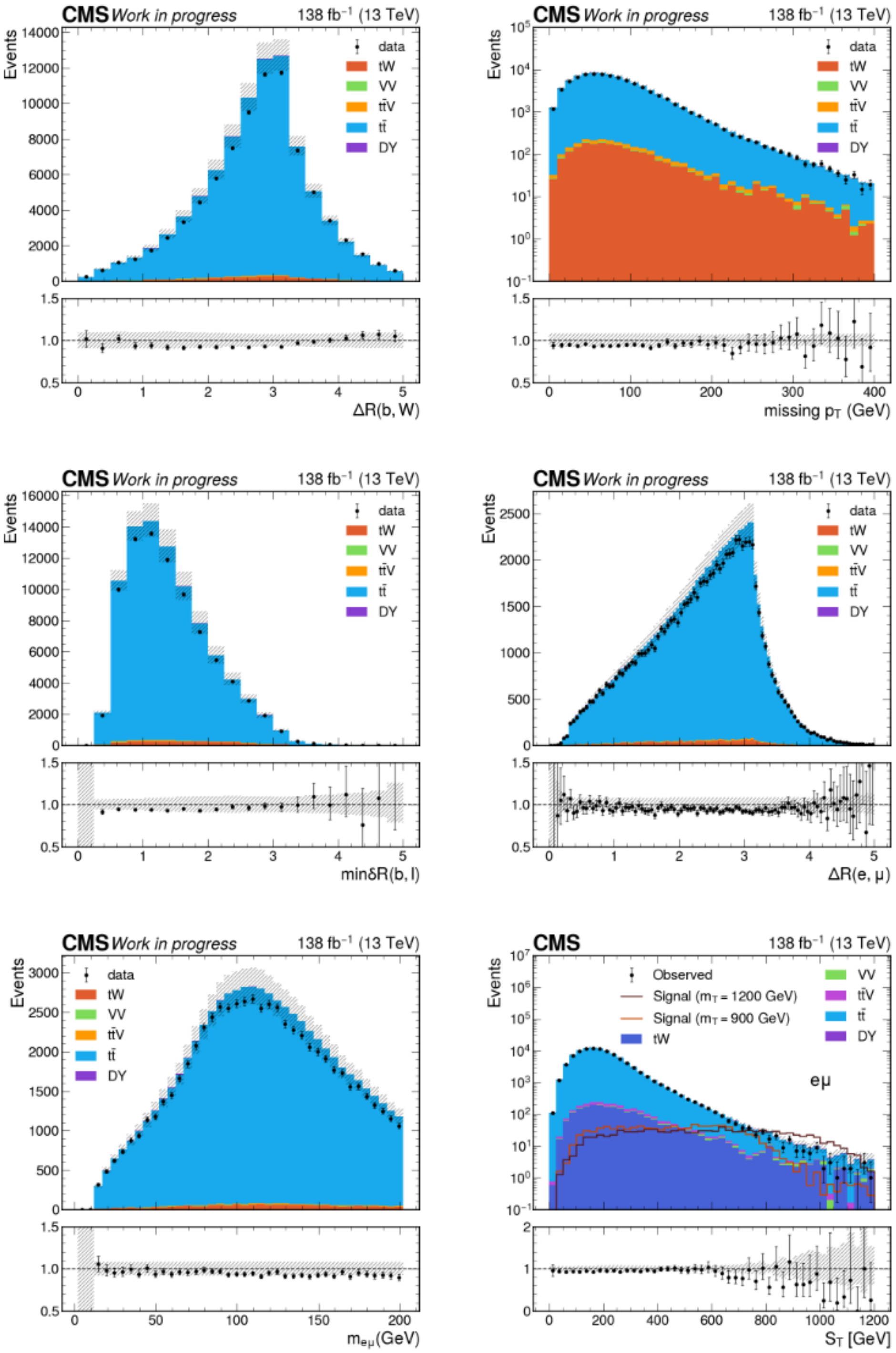

Figure 4.23: Data-MC comparison for the $e\mu$ channel in full Run 2.

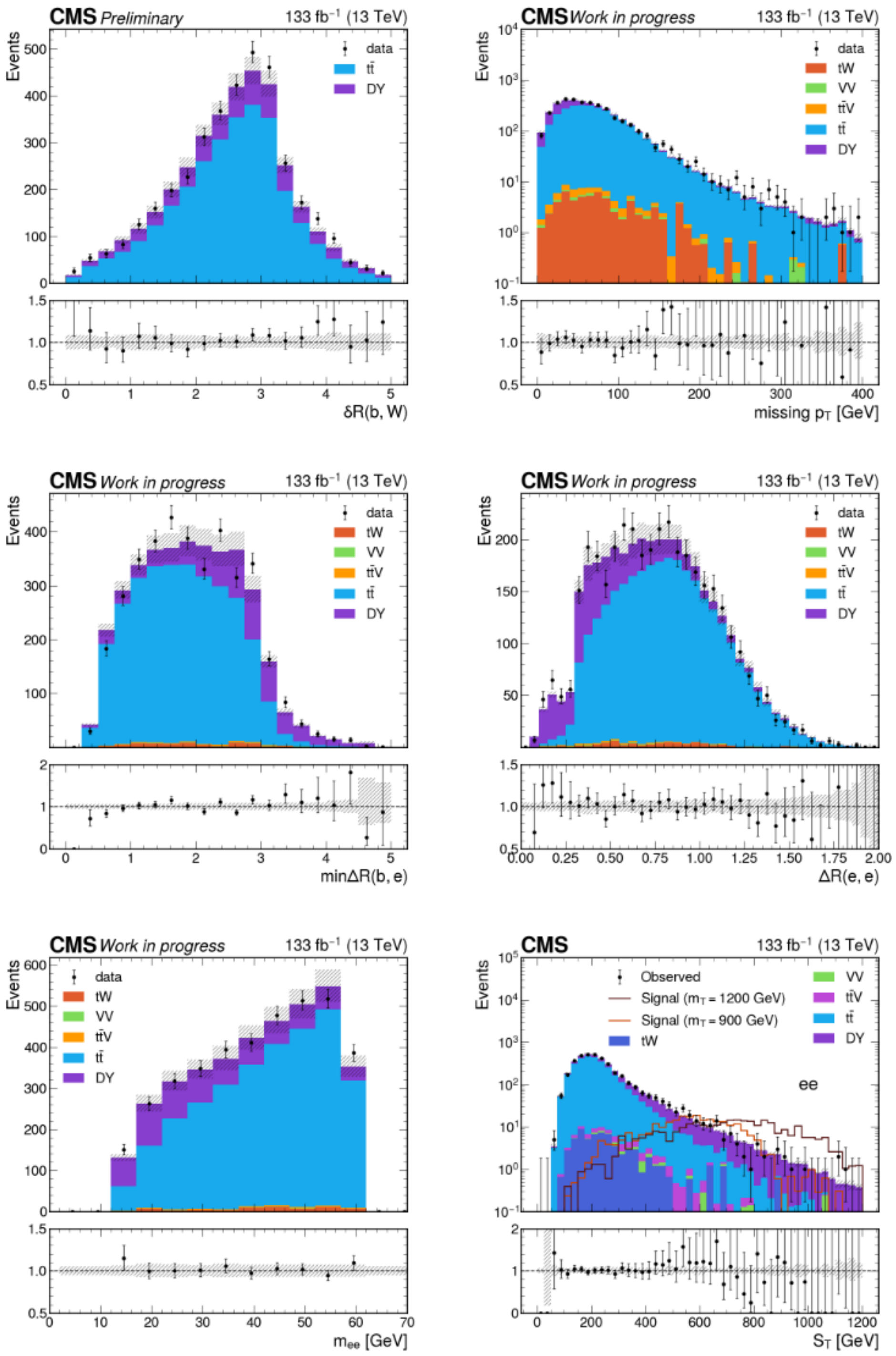

Figure 4.24: Data-MC comparison for the *ee* channel in full Run 2.

function 4.3:

$$f(m_{\mathrm{tH}};m_0,\sigma_0,\alpha_{\mathrm{L}},n_{\mathrm{L}},\alpha_{\mathrm{R}},n_{\mathrm{R}})=\begin{cases}A_{\mathrm{L}}(B_{\mathrm{L}}-\frac{m_{\mathrm{tH}}-m_0}{\sigma_0})^{-n_{\mathrm{L}}}, & \frac{m_{\mathrm{tH}}-m_0}{\sigma_0}<-\alpha_{\mathrm{L}}\\ \exp\Big\{\Big(-\frac{1}{2}[\frac{m_{\mathrm{tH}}-m_0}{\sigma_0}]^2\Big)\Big\}, & -\alpha_{\mathrm{L}}\leq\frac{m_{tH}-m_0}{\sigma_0}\leq\alpha_{\mathrm{R}}\\ A_{\mathrm{R}}(B_{\mathrm{R}}-\frac{m_{\mathrm{tH}}-m_0}{\sigma_0})^{-n_{\mathrm{R}}}, & \frac{m_{\mathrm{tH}}-m_0}{\sigma_0}>\alpha_{\mathrm{R}}\end{cases}\tag{4.3}$$

$$A_{\mathrm{i}}=(\frac{n_{\mathrm{i}}}{|\alpha_{\mathrm{i}}|})^{n_{\mathrm{i}}}\exp\Bigg\{\Bigg(-\frac{{\alpha_{\mathrm{i}}}^2}{2}\Bigg)\Bigg\},\quad B_{\mathrm{i}}=\frac{n_{\mathrm{i}}}{|\alpha_{\mathrm{i}}|}-|\alpha_{\mathrm{i}}|\tag{4.4}$$

This function is composed of a Gaussian core, a left-sided power-law tail, and a right-sided power-law tail. There are six free parameters in one DSCB function: m_0 is the mean of the Gaussian core, σ_0 is the standard deviation of the Gaussian core, α_{R} and α_{L} represents the positions where to switch from the Gaussian to the power-law tails on the right and left sides, n_L and n_R are the exponents shaping the power-law tails.

Figure 4.25 shows the results after fitting the $m_{t\mathrm{H}}$ distributions with the DSCB function, taking T mass points of 600, 800, and 1000 GeV in the 2018 $\mu\mu$ channel as examples. Each signal distribution is obtained after all analysis selections and corrections, with the total number of events normalized to one. The fitting results show that the DSCB function describes the signal well.

4.4.2 Background Modeling

To estimate the SM background in the signal region, in principle two possible approaches could be considered: MC-based modeling and data-driven modeling. The MC-based method uses the normalized SM background MC distributions in the signal region. This method heavily relies on the accuracy of the MC simulations, so they must be verified in a validation region. The data-driven method uses pure data from a control region to model the background. In a control region, the background shape is expected to be proportional to the background in the signal region. Using a data-driven background avoids potential mismodeling from MC simulation. Many analyses use a hybrid method for background modeling, where some background components are modeled with MC and the rest are derived from the control region data.

In this analysis, the background is extracted from data. Unlike a pure data-driven method, background modeling is directly extracted from the signal region data, rather than a control region. As shown in Figure 4.21, the background has a smoothly falling shape in the mass region relevant for this analysis (400 GeV $< m_{tH} <$ 1500 GeV). The initial background shape is determined by MC simulation in the signal region. The MC study indicates that the background can be parameterized by a monotonically decreasing function with exponential-like properties. In the signal extraction fit, the normalization factor and the parameters in the background PDF are freely floated to gain the final background shapes.

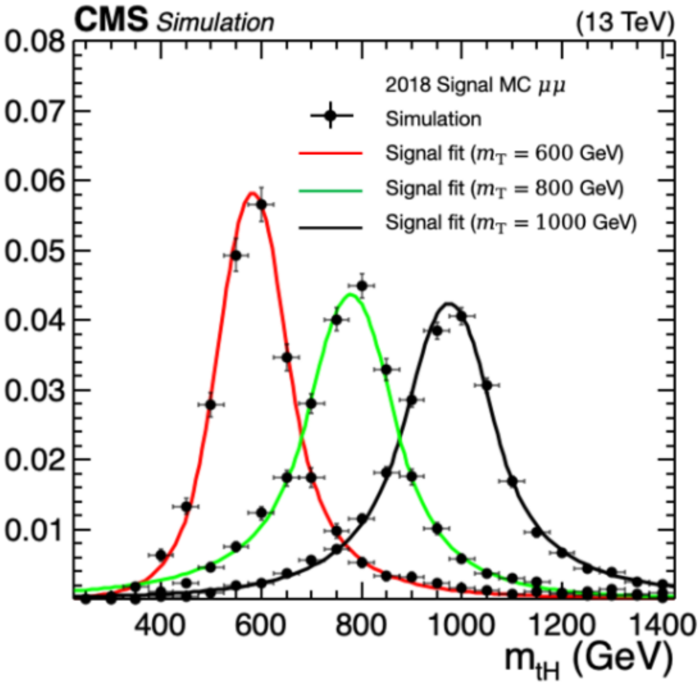

Figure 4.25: Signal models obatined from the $m_{t\mathrm{H}}$ distributions, taking the T mass points 600, 800 and 1000 GeV in the 2018 $\mu\mu$ channel as examples.

Two pdfs (Equation 4.5) are included in the background model. The first one, labeled as f_0, is the exponential of a second-order polynomial in the logarithm of m_{tH}. The second one, labeled as f_1, is the exponential of a second-order polynomial of m_{tH}, and is chosen as an alternative function mainly for extracting the systematic uncertainty from the background pdf choice. In both functions, m_{tH} is associated with a constant m_c. It has been chosen to be 650 GeV solely to reduce the correlation between the two variable parameters (α_1 and α_2).

$$\begin{aligned} f_0(m_{\mathrm{tH}};\alpha_1,\alpha_2) &= \exp\left\{(\alpha_1(\log\frac{m_{\mathrm{tH}}}{m_c}) + \alpha_2(\log\frac{m_{\mathrm{tH}}}{m_c})^2)\right\} \\ f_1(m_{\mathrm{tH}};\alpha_1,\alpha_2) &= \exp\left\{(\alpha_1(m_{\mathrm{tH}} - m_c) + \alpha_2(m_{\mathrm{tH}} - m_c)^2)\right\} \end{aligned} \tag{4.5}$$

Figure 4.26 shows the results of fitting the background MC $m_{t\mathrm{H}}$ distributions with the two pdfs, showing $\mu\mu$, $e\mu$ and ee in all data-taking eras. In the upper panels, the black dots represent the MC simulation background, the red curves denote f_0, and the blue curves denote f_1. The lower panels show the differences between the fitted results and the MC background. Overall, the two functions fit the background well. These fitting results, as shown in Table 4.11, are the initial parameterization of the background models. During the signal extraction, the two PDF candidates are implemented using the discrete profiling method later introduced in Section 4.5.2. All the parameters, including the pdf index and the normalization factor, are floated and derived from the signal region data.

4.5 Statistical Models

4.5.1 Observation Models and Likelihoods

The observation model, $M(r,\ \vec{\theta})$, defines the probability for any set of observations given specific values of the input parameters. All the knowledge about signal and background modeling from Section 4.4 is implemented. The input parameters in the observation model can be roughly divided into two types: the parameter of interest and the nuisance parameters. The former type is the quantity measured by an analysis, and the latter type includes the parameters that affect the model expectation but are not of direct interest in the measurement. In this analysis, the parameter of interest is the signal strength of T production, labeled as r. The signal strength can be derived from the number of expected events (n^{exp}) under the signal-plus-background hypothesis, which is written as $n^{\mathrm{exp}} = r \cdot n^{\mathrm{exp}}_{\mathrm{signal}} + n^{\mathrm{exp}}_{\mathrm{bkg}}$. A set of nuisance parameters, labeled as $\vec{\theta}$, consists of the background modeling parameters, the background normalization factors, and auxiliary parameters representing systematic uncertainties. The likelihood,

$$\mathscr{L}_M(r,\ \vec{\theta}) = p_M(\mathrm{data};\ r,\ \vec{\theta}) \tag{4.6}$$

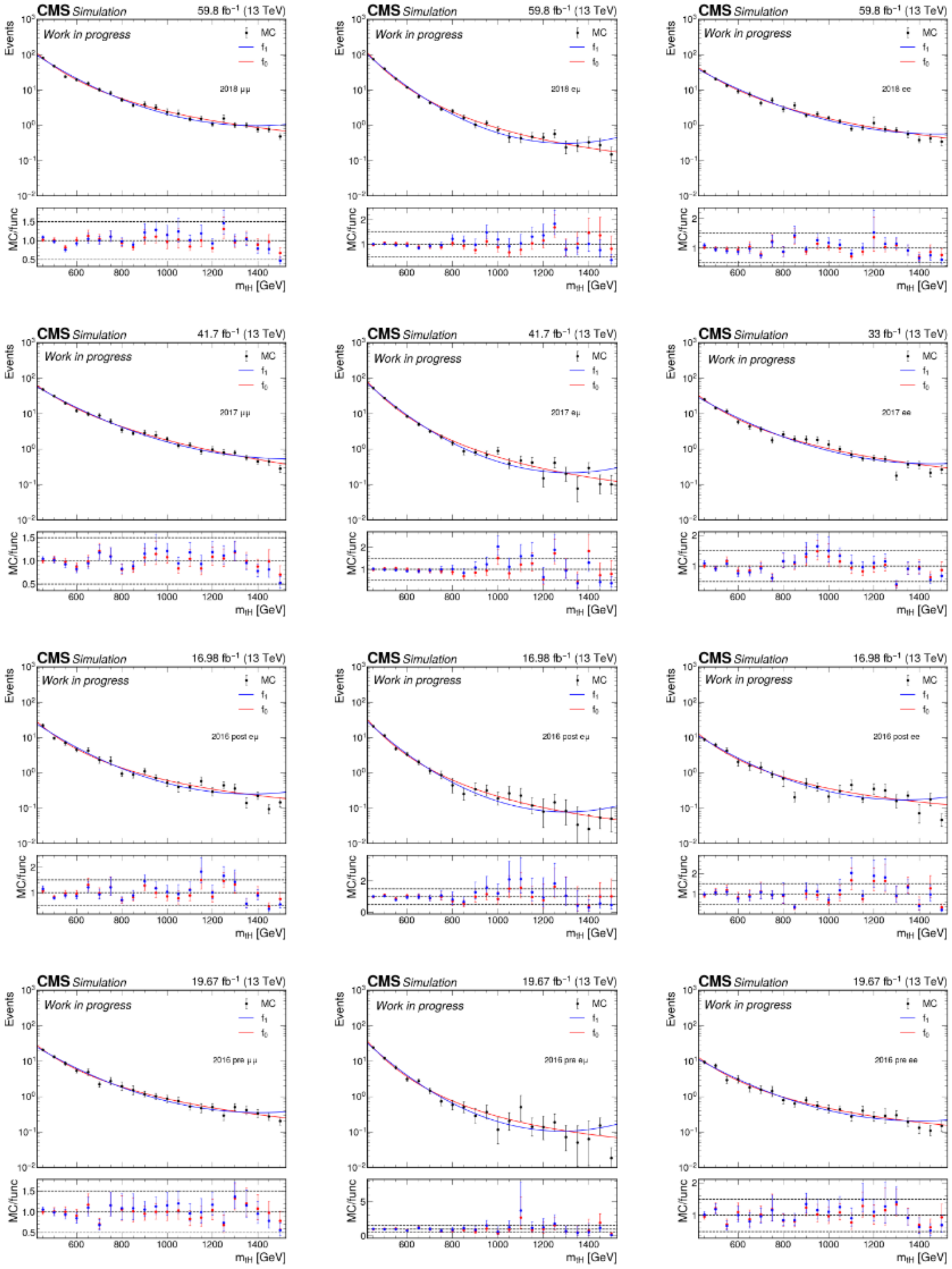

Figure 4.26: Fit background $m_{t\mathrm{H}}$ distributions with two PDFs.

		f_0		f_1	
		α_1	α_2	α_1	α_2
2018	$\mu\mu$	1.102	4.437	0.490	-0.778
	$e\mu$	1.620	5.769	0.755	-1.060
	ee	0.687	3.894	0.375	-0.663
2017	$\mu\mu$	2.678	5.352	0.376	-0.708
	$e\mu$	1.558	5.748	0.735	-1.05
	ee	0.852	4.052	0.402	-0.693
2016 preAPV	$\mu\mu$	1.421	5.115	0.434	-0.708
	$e\mu$	1.929	5.745	0.778	-1.063
	ee	0.995	3.883	0.418	-0.673
2016 postAPV	$\mu\mu$	1.310	4.486	0.511	-0.789
	$e\mu$	1.757	5.875	0.770	-1.080
	ee	1.302	4.173	0.508	-0.744

Table 4.11: MC background fitting results.

is the probability of observing the data given the observation model M. In this analysis, the m_{tH} distributions from data are presented in binned histograms. Background and signal models are parameterized by PDF functions. The full observation model consists of twelve channels, combining three lepton flavor categories ($\mu\mu$, $e\mu$ and ee) and four data-taking eras (2018, 2017, 2016 post APV and 2016 pre APV). Equation 4.7 shows the expression of the parametric likelihood on binned data,

$$\mathscr{L} = \prod_c \prod_b \mathrm{Poisson}(n_{cb}^{\mathrm{obs}};\ n_{cb}^{\mathrm{exp}}(r,\ \vec{\theta})) \prod_e p_e(y_e;\ \theta_e) \tag{4.7}$$

where c indexes the channel, b indexes the data histogram bin, and e indexes the nuisance parameters. The first term shows the primary likelihood, which is a product of the likelihoods in all bins and channels. Each likelihood is the Poisson probability of observing n_{cb}^{obs} events with an n_{cb}^{exp} expectation. The second term shows the auxiliary likelihood, which is a production of auxiliary terms arising from nuisance parameters representing systematic uncertainties. The form of p_e in each auxiliary term depends on the type of systematic uncertainty, which is sorted by how it affects the observation model. For each p_e term, y_e represents the measured parameter, and θ_e is the measurement from the previous observation. Systematic uncertainties affecting the signal model can be divided into rate and shape uncertainties. The shape systematic uncertainties change the shapes of signal models, while the rate systematic uncertainties vary the expected yields of signal.

Specifically, the rate uncertainties are modeled using a lognormal distribution as the auxiliary term, as shown in the following equation,

$$p_e(y_e;\ \theta_e) \propto \frac{1}{\theta_e \kappa} \exp\left(-\frac{(\ln \theta_e - y_e)^2}{2\kappa^2}\right) \tag{4.8}$$

In a lognormal distribution, the logarithm of the variable follows a normal distribution with the width κ. Effects from each rate-uncertainty source are quantified using "up" and "down" ratios. These ratios compare the yields for a $+1\kappa$ or -1κ deviation to the previously measured nominal yield.

The auxiliary terms of shape uncertainties use a Gaussian distribution, and can be written as:

$$p_e(y_e;\ \theta_e) \propto \exp\left(\frac{-(\theta_e - y_e)^2}{2\sigma^2}\right) \tag{4.9}$$

In the Gaussian constrained terms, the mean value (y_e) is set to 0, and the width (σ) is set to 1. Shape uncertainties affect the fitting parameters in the signal model, such as m_0 in Equation 4.4. For each parameter affected by a shape uncertainty, the up and down variations are obtained by fitting the signal distributions with $\pm 1\sigma$ variations. The ratios between the up or down parameters and the nominal parameter are implemented in the auxiliary terms.

4.5.2 Systematic Uncertainties

Systematic uncertainties arise from experimental and theoretical effects, and are implemented in the statistical model by the auxiliary terms. To study the impact of each systematic uncertainty on signal modeling, histograms are filled using m_{tH} signal MC distributions with one 'up' and 'down' standard deviations after all analysis cuts. The variations are done by floating the scale factor corresponding to the specific systematic uncertainty, while keeping the others at their nominal values. The effects due to systematic variations are quantified by fitting the distributions for all T mass hypotheses, data-taking eras, and channels ($\mu\mu$, $e\mu$, and ee), respectively. All systematic uncertainties related to signal modeling in the analysis are discussed below.

The JES uncertainty has both rate and shape effects on signal modeling. To estimate its influence, the four momentum values of all jets in all events are scaled up or down by one standard deviation. Figure 4.27 shows that the JES variations shift the mean values of the m_{tH} peaks, resulting in shape and rate effects due to changes in jet four momentum. In each plot, two standard deviations are applied instead of one to improve the presentation. The blue points show the distribution with two standard deviations down, the red points show the distribution with two standard deviations up, and the black points show the nominal m_{tH} distribution. To estimate the shape effect, each distribution is individually fitted with a DSCB function 4.3. Two fractions are calculated as $(m_0^{\text{up}} - m_0^{\text{nominal}})/m_0^{\text{nominal}}$

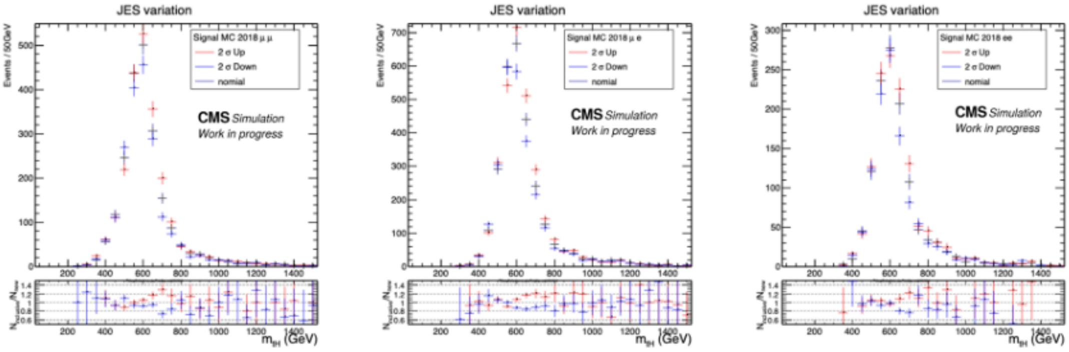

Figure 4.27: Systematic uncertainty study of the JES correction with signal MC after all analysis cuts, taking 2018 $\mu\mu$, $e\mu$, and ee channels under $m_{\mathrm{T}} = 600$ GeV hypothesis as examples.

and $(m_0^{\mathrm{nominal}} - m_0^{\mathrm{down}})/m_0^{\mathrm{nominal}}$, where m_0^{up}, m_0^{down}, and m_0^{nominal} are from the three fitting results. The larger fraction of the two is implemented in the statistical model. Modifying the jet four-momentum also changes the fraction of events passing the selection, thereby explaining the rate effect. The rate effects are quantified by comparing event numbers (n^{up}, n^{down}, and n^{nominal}) in the up, down, and nominal distributions. The rate effect from up variation is calculated as $(n^{\mathrm{up}} - n^{\mathrm{nominal}})/n^{\mathrm{nominal}}$, and from down variation it is $(n^{\mathrm{down}} - n^{\mathrm{nominal}})/n^{\mathrm{nominal}}$. Calculations for both shape and rate effects are implemented in the statistical model using one nuisance parameter per data-taking era.

The JER uncertainty has rate and shape effects on the signal models. As mentioned in Section 3.3.4, the JER correction modifies the resolution of the jet momentum measurement, thus affecting the width of the m_{tH} distributions in signal. The up and down variations are handled by adjusting the s_{JER} values in the JER correction calculation (Equation 3.10, 3.11) for each jet. As in the JES systematic study, the systematic effects from JER uncertainty is obtained by fitting the distributions in Figure 4.28. The JER uncertainty changes the σ_0 values while fitting the signal MC distributions to the DSCB function. The shape effect of JER is quantified by a fraction calculated with $\sigma_0^{\mathrm{up/down}}$ and $\sigma_0^{\mathrm{nominal}}$. The rate effect is obtained by comparing the number of events with and without the JER variations. Both fractions are introduced into the statistical model via a nuisance parameter for each data-taking era.

Ten separate uncertainties related to b-tagging are applied to the analysis, combining jet flavors (b/c jets and light jets), data-taking eras (2018, 2017, 2016 post, and 2016 pre), and correlations between eras (correlated and uncorrelated across all eras). To study each of the ten uncertainties, up and down variations are applied by adjusting the b-tagging scale factor in the b-tagging weight calculation (Equation 3.12). Figure 4.29 shows the effects caused by the systematic variations. The shape effects due to b-tagging systematic uncertainties are negligible compared to JES and JER, hence only rate uncertainties are

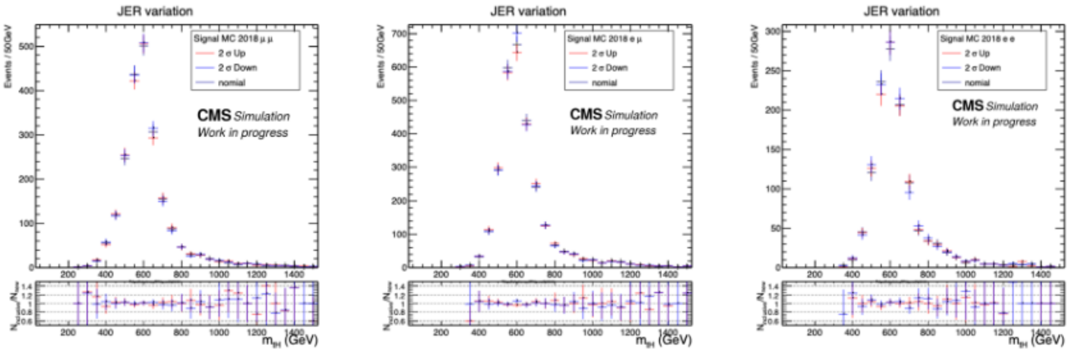

Figure 4.28: Systematic uncertainty study of the JER correction with signal MC after all analysis cuts, taking 2018 $\mu\mu$, $e\mu$, and ee channels under $m_{\mathrm{T}} = 600$ GeV hypothesis as examples.

applied.

The estimation of the integrated luminosity leads to a cross section uncertainty of 0.6–2%. It can be further broken down into five uncertainties: three uncorrelated uncertainties for the three data-taking years (2016, 2017, and 2018), one correlated uncertainty for all three years, and one correlated uncertainty for 2017 and 2018. They all have rate effects because they influence the expected signal events.

Lepton identification, isolation, and single lepton trigger uncertainties have rate effects. Figure 4.30 shows the up and down variation distributions for the electron ID systematic uncertainty study, and Figure 4.31 shows the distributions to study the muon ID and isolation systematic effects. The shape effects from lepton-related systematic uncertainties are negligible. Only rate effects are implemented in the statistical models.

Pile-up reweighting has a rate effect on signal modeling. The effects are quantified by varying the pile-up scale factors separately for all the data-taking eras.

Level 1 ECAL and muon prefiring are addressed using an event-based weight that corrects a gradual timing shift in the ECAL and muon systems. It has a rate effect on signal modeling.

QCD-scale uncertainties from the renormalization and factorization scales in the matrix element generators have a rate effect, while the shape effect is found negligible for this analysis. To quantify the effect, the renormalization factor and the factorization factor are varied by factors of 0.5, 1 and 2, yielding nine combinations and nine weights. The standard deviation of the sum of the scaled weights is taken as the uncertainty.

The parton distribution function uncertainty has a rate effect. The systematic effect is estimated using 101 sets of event-by-event weights corresponding to the 101 different parton distribution functions, including one function used by signal models and 100 alternative ones. The uncertainty is quantified by the standard deviation of all the weights from different parton distribution functions.

Discrete Profiling Method

To account for systematic uncertainties arising from the choice of background pdf, the discrete profiling method [68] is applied. In this method, multiple pdf candidates are included in the background model, each is labeled with an index. One of them will be selected by the computing tool to determine the signal strength, while others are used to estimate the systematic uncertainty. The index of the active pdf in the fit is referred to as a discrete nuisance parameter. It is profiled by looping through all possible pdf index values and identifying the pdf that provides the best fit. To quantify the goodness of the background fit, Λ is defined in Equation 4.10, where for the i^{th} bin in the binned data, n_i^{obs} is the observed and n_i^{exp} is the expected number of events from the background PDF:

$$\Lambda = 2\sum_i n_i^{\text{exp}} - n_i^{\text{obs}} + n_i^{\text{obs}} \ln\left(n_i^{\text{obs}}/n_i^{\text{exp}}\right) \tag{4.10}$$

In the profile scan, for each background PDF choice, the variable Λ is scanned over the signal strength (r). For example, Figure 4.32 displays the Λ scan curves, with each curve representing a background PDF as indicated in the legend. The red curve, which gives the global minimum Λ value, is used as the chosen fit to measure the signal strength. The blue curve is then profiled as the alternative fit to derive the systematic uncertainty from the choice of background PDF. The solid black line, referred to as the "envelope", represents the lower boundary of the red and blue curves in the plot. By reading the 68.3% and 95.4% intervals on r from the envelope curve, the systematic effect brought by the alternative background PDF function is taken into consideration.

4.5.3 The maximum likelihood method

The signal strength of T production is measured using a maximum-likelihood method. Given the observed data, it finds the most probable parameterization of the statistical model, including the parameter of interest (r) and the nuisance parameters $(\vec{\theta})$, that maximize the likelihood. The procedure is usually done by numerically solving the likelihood equations 4.11, where the logarithm is applied due to technical reasons.

$$-\frac{\partial \ln \mathscr{L}}{\partial x} = 0 \quad x \in (r,\ \vec{\theta}) \tag{4.11}$$

4.5.4 Statistical tests

Based on likelihood models and the observed data, statistical tests provide rules for accepting or rejecting a hypothesis, usually a null hypothesis. They are used to estimate parameters, determine the significance of new-physics discoveries, and set limits to exclude certain parameter phase spaces. A "test statistic", as a function of the observed data, is required for the statistical test. In this analysis, the negative logarithm of the likelihood ratio is chosen as the test statistic for limit setting and the goodness of fit tests, as shown in Equation 4.15.

$$t \propto -\log\left(\frac{\mathscr{L}_M}{\mathscr{L}_{M'}}\right) \tag{4.12}$$

The result of a statistical test is evaluated using a p-value, which is the probability of observing data at least as extreme as the actual observation. For a certain test statistic distribution D_M, the p-value can be obtained by the integration in Equation 4.13.

$$p = \int_{t_{min}}^{t_{obs}} D_M dt \tag{4.13}$$

If the p-value is smaller than 0.05, the hypothesis is rejected.

To ensure the statistical model fits the data well, a goodness-of-fit (GOF) test is performed. The GOF test uses the log likelihood ratio function in Equation 4.14 as the test statistic, and is based on the binned data.

$$t_{\text{saturated}} = -2\ln\left(\frac{\mathscr{L}(r,\vec{\theta})}{\mathscr{L}_{\text{saturated}}}\right) \tag{4.14}$$

The likelihood in the nominator is from the fit to the observed data or the pseudodata under the null hypothesis. The likelihood in the denominator comes from the saturated model, which is an alternative hypothesis that matches the prediction and the observation in each bin. A saturated model typically needs as many parameters as there are data bins.

The result of a GOF test is a p-value calculated based on the toy model distribution and the observed data. The toy distribution of $t_{\text{saturated}}$ is obtained by generating pseudodata from the likelihood model, and a $t_{\text{saturated}}$ value is calculated with the observed data under the null hypothesis. The fit is considered a good fit under the null hypothesis when the p-value is bigger than 0.05. The GOF test results for the analysis are presented in the next chapter.

4.5.5 Upper limits setting

In the absence of a signal in the data, upper limits are set to exclude some parameter phase spaces. The limits are computed using the modified frequentist confidence level (CL) criterion [69] [70] with the asymptotic approximation [71]. The test statistic for limit setting is a modified likelihood ratio:

$$t_r = \begin{cases} -2\ln\left(\frac{\mathscr{L}(r)}{\mathscr{L}(r=0)}\right) & \hat{r} < 0 \\ -2\ln\left(\frac{\mathscr{L}(r)}{\mathscr{L}(\hat{r})}\right) & 0 < \hat{r} < r \\ 0 \quad \hat{r} > r \end{cases} \tag{4.15}$$

where $\hat{r}$ is the signal strength from the maximum likelihood fit. If $0 < \hat{r} < r$, the denominator in the logarithm is the maximum likelihood. If $\hat{r} > r$, the test statistic is set to 0, ensuring phase spaces outside the chosen confidence interval are not excluded. For $\hat{r} < 0$, the r value in the denominator is set to 0 so that no negative limits are set. In the search for new physics processes, upper limits are estimated using Equation 4.16, where CL is typically 95%; p_{s+b} and p_b are p-values computed with the test statistic in Equation 4.15. The calculation of p_{s+b} is under the signal-plus-background hypothesis, while the p_b calculation assumes the null-hypothesis. As a conclusion, the phase spaces with the signal strength higher than the upper limit are excluded, since they have smaller p-values than the threshold corresponding to the given confidence level.

$$\frac{p_{s+b}}{1 - p_b} < (1 - \mathrm{CL}) \tag{4.16}$$

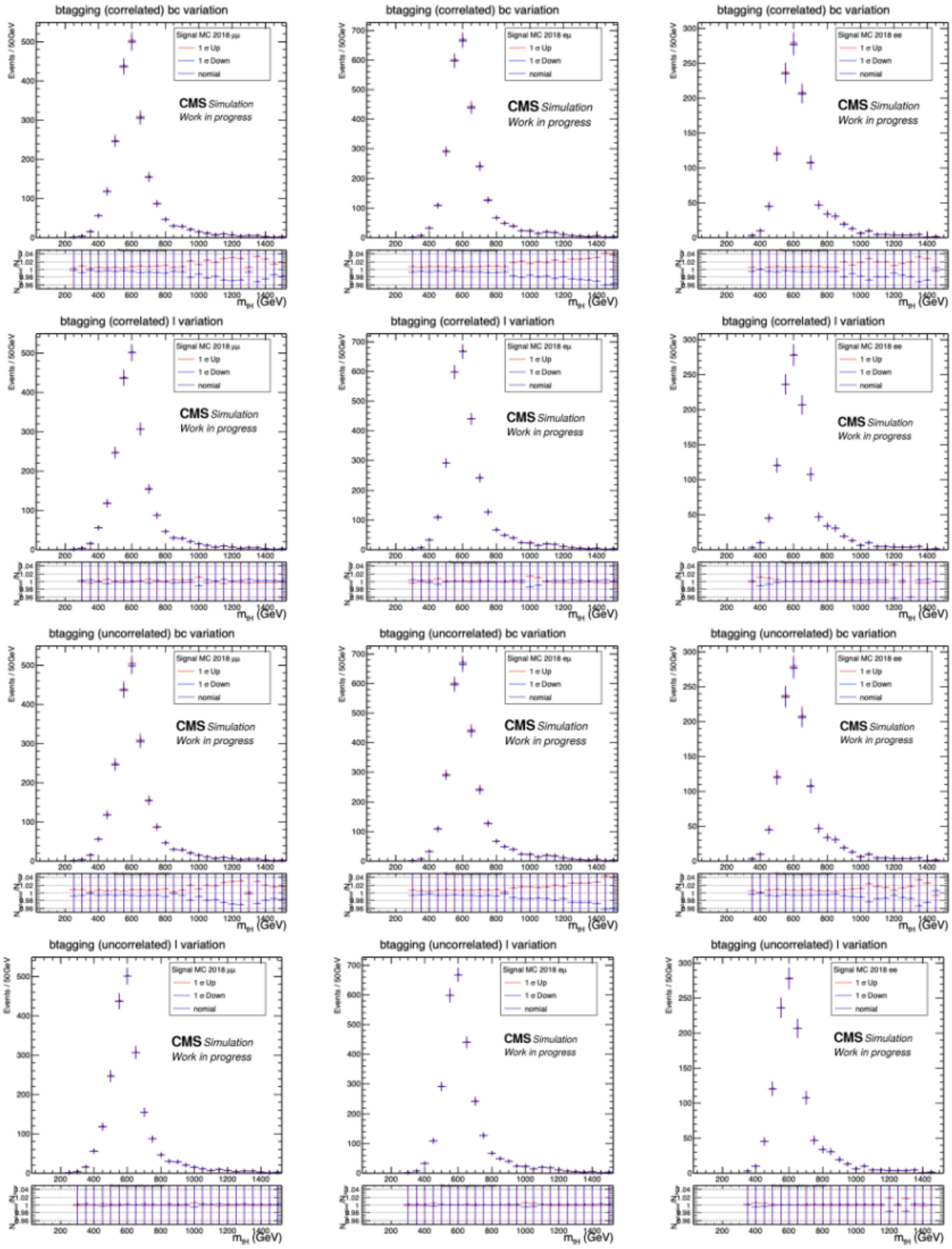

Figure 4.29: Systematic uncertainty study of b tagging corrections with signal MC after all analysis cuts

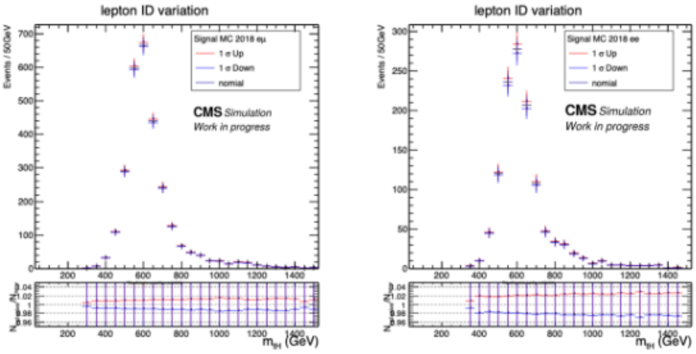

Figure 4.30: Systematic uncertainty study of the electron ID correction with signal MC after all analysis cuts, taking 2018 $e\mu$ and ee channels under $m_{\mathrm{T}} = 600$ GeV hypothesis as examples.

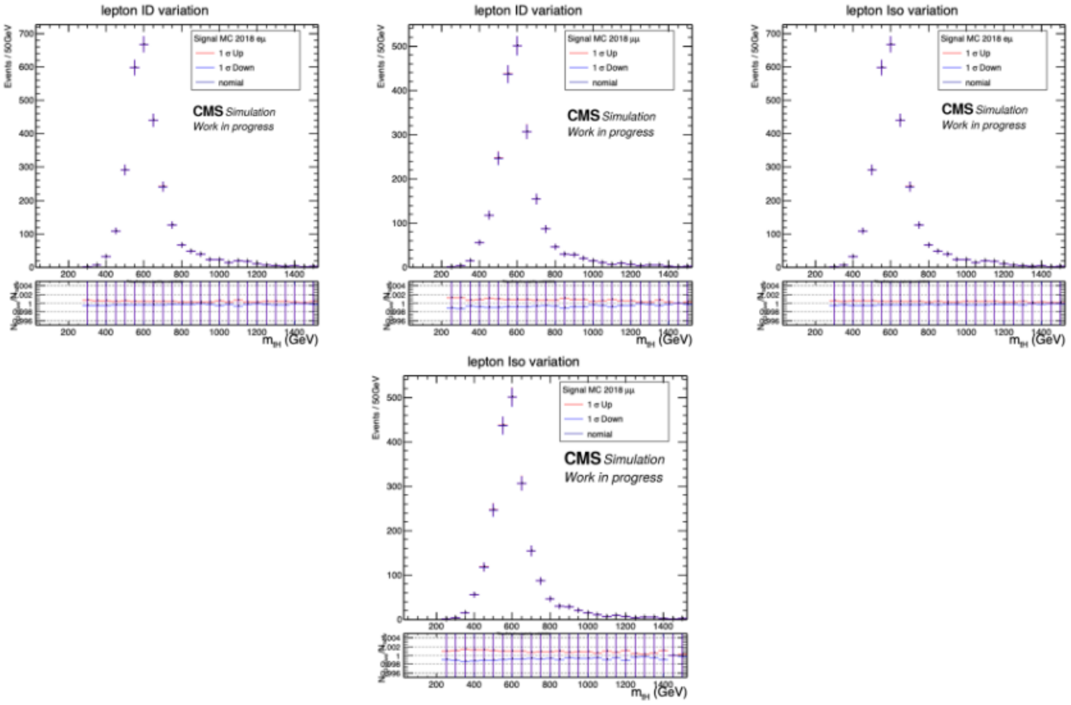

Figure 4.31: Systematic uncertainty study of the muon ID and isolation corrections with signal MC after all analysis cuts, taking 2018 $e\mu$ and $\mu\mu$ channels under $m_{\mathrm{T}} = 600$ GeV hypothesis as examples.

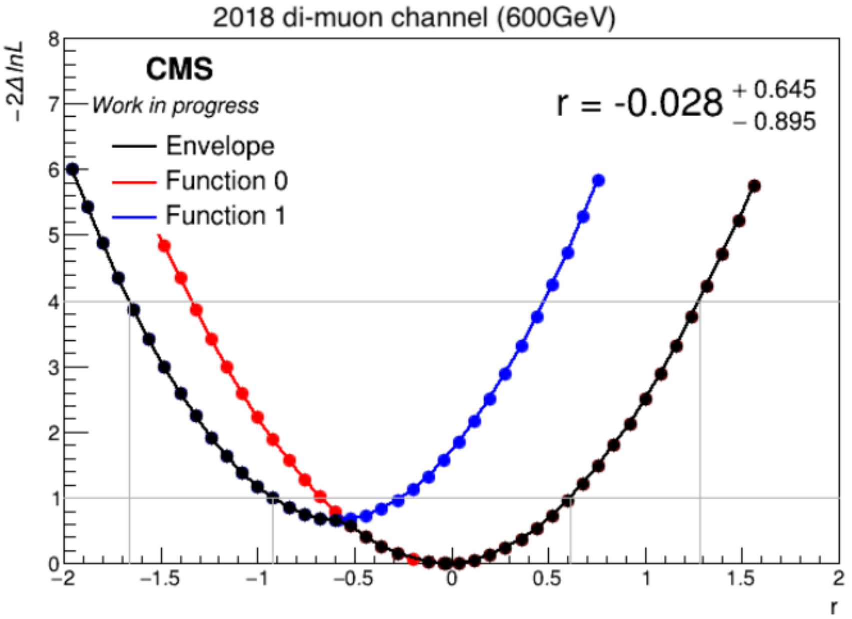

Figure 4.32: Profile scans for the two PDF fits and the envelope obtained from the fits. On the y-axis, $-2\Delta\ln\mathscr{L}$ is twice the negative of the logarithm of the likelihood ratio function, and is equal to Λ defined by Equation 4.10. Note that the plot shows the fits to the MC background; thus, it is not used to extract systematic uncertainty. This is only an exercise to illustrate the discrete profiling method.

CHAPTER 5

Results

This chapter presents the results of searching for T production in the T → tH decay mode in the final states containing two opposite-sign leptons. As discussed earlier, the analysis combines three lepton-flavor categories and four data-taking eras, yielding 12 channels. For each channel, signal models are built under seven T-mass hypotheses. The results are determined by a joint likelihood-based fit combining all channels using a single common signal strength parameter. The CMS COMBINE tool [72], a software framework built on RooFit [73] and RooStats [74], is used to build likelihood-based statistical models, perform statistical tests, and determine results.

The following information is input to COMBINE to build the statistical model. For signal modeling, the T signal distributions across all channels are parameterized using DSCB functions (Equation 4.3). The fitting results describing the signal shapes are implemented in the statistical model as fixed parameters. The expected signal events for each channel are normalized to a cross-section $\sigma(\mathrm{pp} \to \mathrm{T})\mathscr{B}(\mathrm{T} \to \mathrm{tH})$ of 1 pb. Thus, the signal strength in the measurement results is converted to the cross section $\sigma(\mathrm{pp} \to \mathrm{T})\mathscr{B}(\mathrm{T} \to \mathrm{tH})$. For background modeling, all the parameters in the background functions (Equation 4.5), the background normalization parameters, and the index of the active function are implemented in the statistical models as floating parameters. All the systematic uncertainties are included in the statistical model through nuisance parameters. The observed data are presented as binned histograms of m_{tH} across all channels.

In CMS, the signal region data is accessible only with permission to unblind. Before unblinding, the analysis strategy should be shown to be optimal and validated with non-signal-region data and Asimov data. The Asimov data is toy data generated from the statistical model, used for statistical validation rather than the signal region data. An Asimov dataset contains simulated signal and background events, with the expected signal strength set in the configuration. After unblinding, the analysis strategy is frozen, and the signal region data is accessible. Tests are performed both before and after unblinding to validate the analysis methods and the statistical models.

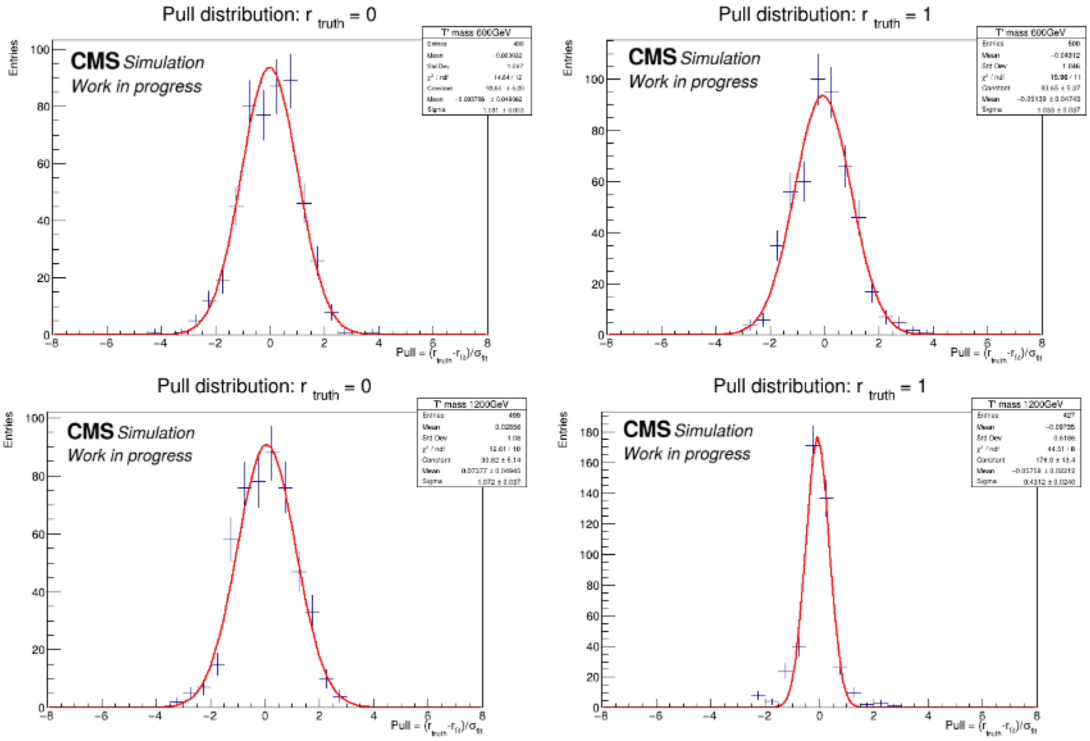

Figure 5.1: Bias test at m_{VLQ} =600 and 1200 GeV combining all channels in Run 2 with $r_{truth} = 1$ *or* 0 assumptions.

5.1 Statistical model validation

5.1.1 Bias tests

The bias tests are done with Asimov datasets to verify the statistical model and the fitting procedure before unblinding. Firstly, an Asimov dataset is generated assuming the signal strength is a known value (r_{truth}), typically 0 or 1. Secondly, a maximum-likelihood fit is performed on the Asimov dataset to obtain the signal strength (r_{fit}) and its uncertainty (σ_{fit}). A pull value is thus calculated from $Pull = (r_{truth} - r_{fit})/\sigma_{fit}$. By repeating the procedure hundreds of times, a distribution of the pull values is generated. If there is no bias in the statistical model or the fitting procedure, the pull distribution should be Gaussian with mean 0 and standard deviation 1. The potential bias is defined as the mean value when fitting the pull distribution to a Gaussian function. The statistical model is considered unbiased, as the potential bias is less than 0.14, leading to a change in total uncertainty of less than 1%. Figure 5.1 shows the bias test results with r_{truth} equal to 0 or 1, under the T mass hypotheses of 600 GeV or 1200 GeV. Each distribution is fitted with a Gaussian function, and the potential bias results agree with the no-bias expectation.

5.1.2 Goodness of fit tests

As discussed in Section 4.5.4, the GOF tests are performed to determine whether the null hypothesis should be accepted based on the observed data. In this analysis, the GOF tests use the signal-region data and are performed after unblinding. Figure 5.2 shows the GOF tests using the binned m_{tH} data distributions; each plot shows a $\mu\mu$, $e\mu$, or ee channel combining full Run 2. In each plot, the stacked distribution is from hundreds of toys generated under the null hypothesis, and the blue arrow is from fitting the observed data. The p-values indicate that the data are well fitted by the statistical model under the null hypothesis.

5.1.3 Impact plots

To study the effects of nuisance parameters on the parameter of interest, impact plots are generated using Asimov data before unblinding and observed data after unblinding. The impact is calculated by measuring the signal strength shift (Δr) when varying a nuisance parameter by 1 standard deviation, up ($+1\sigma$) and down (-1σ). The remaining nuisance parameters are fixed at their nominal values while studying the impact of a single nuisance parameter. In each plot, each column shows the impact information of a nuisance parameter, including the fitting results with uncertainties, the pull values, and the Δr values. The pull values are defined by $(\hat{\theta} - \theta_I)/\sigma_I$ or $(\hat{\theta} - \theta_I)/\sqrt{{\sigma_I}^2 - \sigma^2}$, where $\hat{\theta}$ is the best-fit value, θ_I is the input nominal value, σ_I is the input uncertainty, and σ is the uncertainty from fit. The pink and blue bars represent the signal-strength shifts Δr, while the nuisance parameter varies up and down by 1σ. The nuisance parameters are ordered by their impact on signal strength, from top to bottom.

Figure 5.3-5.8 are impact plots at three T mass points (600 GeV, 900 GeV and 1100 GeV), generated with Asimov datasets assuming a signal strength equal to 0 or 1. In the plots with Asimov datasets generated with a signal strength of 0, the impacts from signal-modeling systematic uncertainties are suppressed as expected. In the plots with Asimov datasets assuming a signal strength of 1, the nuisance parameters with the greatest impact are from the background modeling and statistics, especially at the lowest T mass points, 600 GeV. When moving to higher T mass points, such as 1100 GeV, the impact from background modeling decreases, and the effect from the jet energy scale uncertainties in the signal modeling becomes more visible. In general, the Δr bars are symmetric, and the pulls meet expectations, indicating good fit quality.

Figure 5.9 and 5.10 are impact plots after unblinding at 600 GeV and 1200 GeV mass points. The largest impacts are from background modeling in both plots, which agree with the Asimov impact plots. At the 600 GeV mass point, the background normalization parameters have a leading effect on the signal strength, since the background is relatively high in the corresponding mass region. The background shape parameters and the jet energy scale uncertainty from signal modeling show big impacts at 600 GeV. At

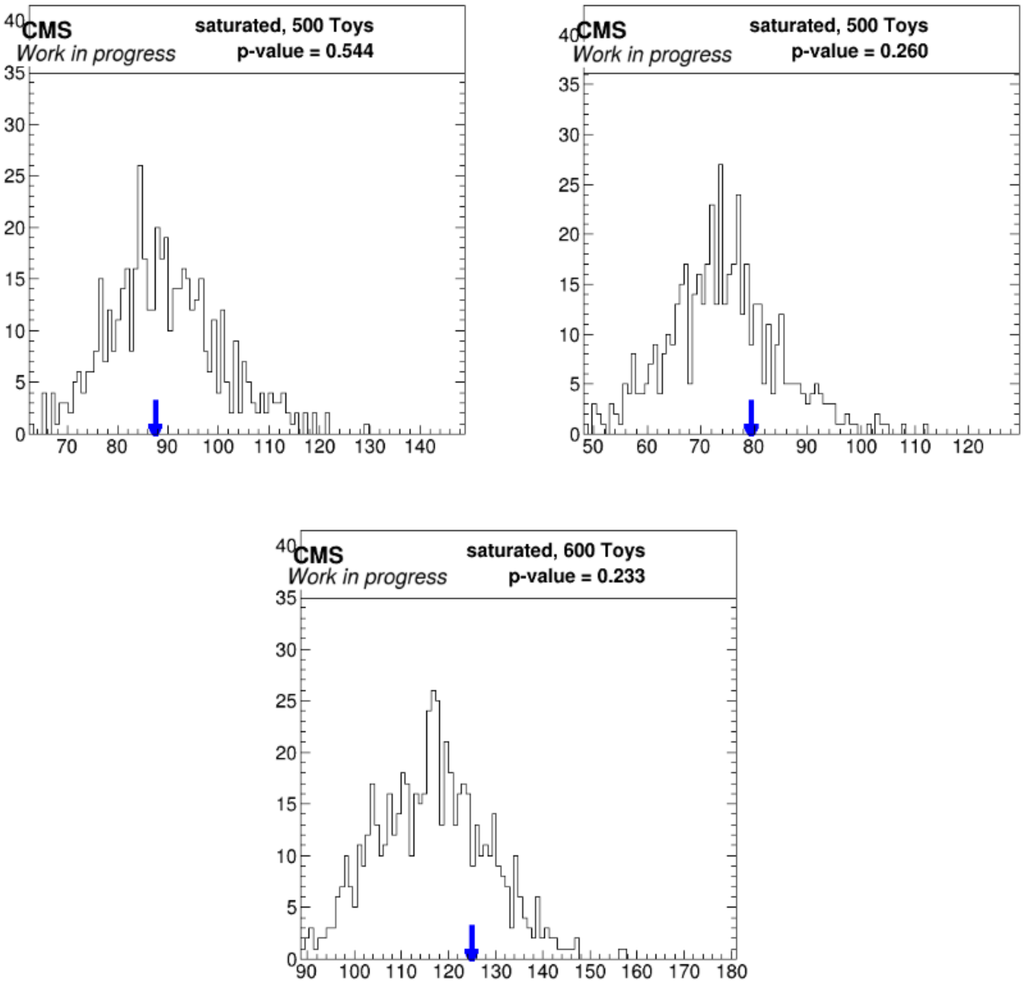

Figure 5.2: GOF tests with signal region data in Run 2, showing $\mu\mu$, $e\mu$, and ee channels from left to right.

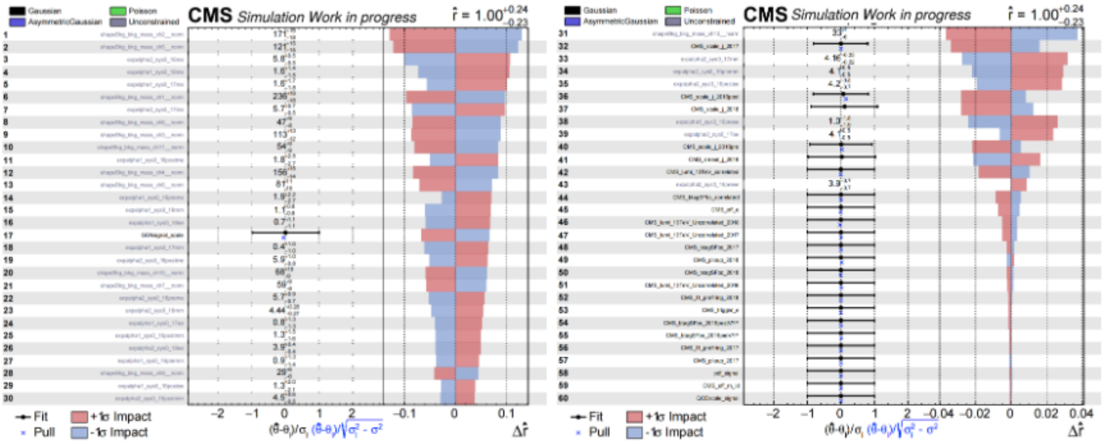

Figure 5.3: Impact plots with Asimov dataset (signal strength r=1) at 600 GeV mass point combining full Run 2.

the 1200 GeV mass point, the background is lower, so that the background shape parameters and the JES in signal become the major impacts, while the impact from the background normalization factors is slightly reduced. Among the background modeling nuisance parameters, those from the $e\mu$ channel have the largest impact, indicating that the $e\mu$ channel has the best sensitivity compared to the ee and $\mu\mu$ channels. No asymmetric Δr bars or unexpected pulls are observed in these plots.

5.2 The m_{tH} distribution in the signal region

The signal region data distributions of the main variable, m_{tH}, are presented after unblinding, associated with the fitting results under the null hypothesis. In Figure 5.11, three plots show the distributions and fitting results in three channels from ee, $\mu\mu$ and $e\mu$ in Run 2. In the upper panel of each plot, the data points are from data, and the green curves show the background shapes from the background-only fits. The lower panels show the pull, which is defined as $(n_{Data} - n_{fit})/\sigma$, where $\sigma^2 = \sigma^2_{Data} - \sigma^2_{fit}$. Additionally, signal shapes under T-mass hypotheses of 600 GeV, 900 GeV, and 1200 GeV are shown in the upper panels as reference. Each signal shape is from MC with all corrections applied and is normalized to a cross-section of 1 Pb. In general, the data shows very good agreement with the background-only hypothesis, as previously shown in GOF tests.

5.3 Limits on VLQ production

No significant excess of events is observed at any of the investigated mass points from 600 GeV to 1200 GeV. The data support the null-hypothesis (background-only), which

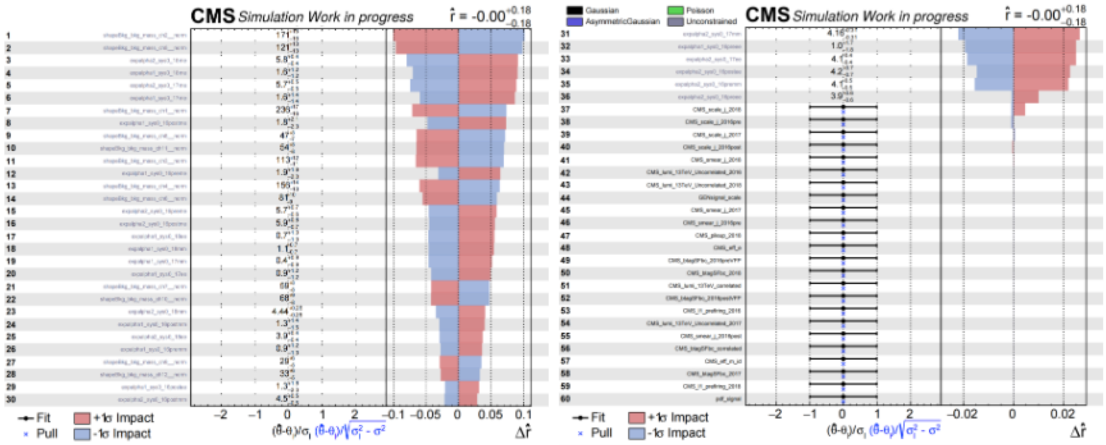

Figure 5.4: Impact plots with Asimov dataset (signal strength r=0) at 600 GeV mass point combining full Run 2.

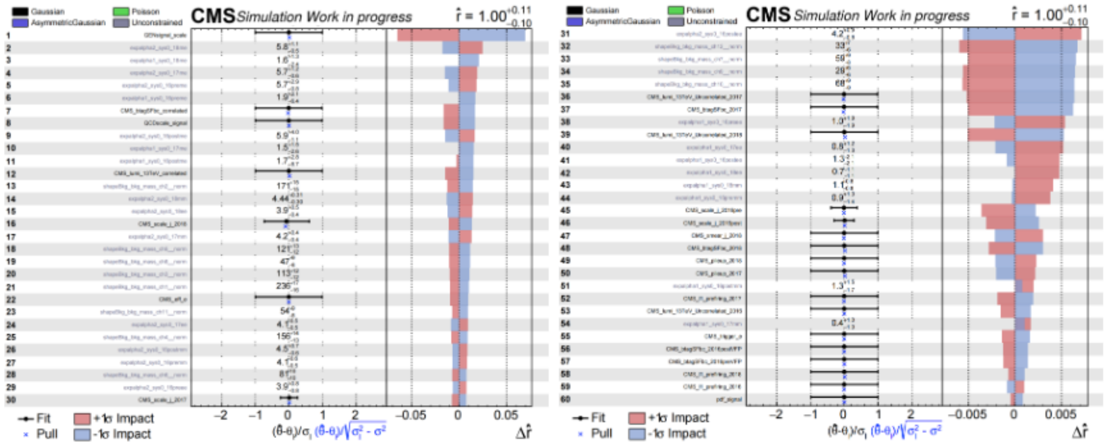

Figure 5.5: Impact plots with Asimov dataset (signal strength r=1) at 900 GeV mass point combining full Run 2.

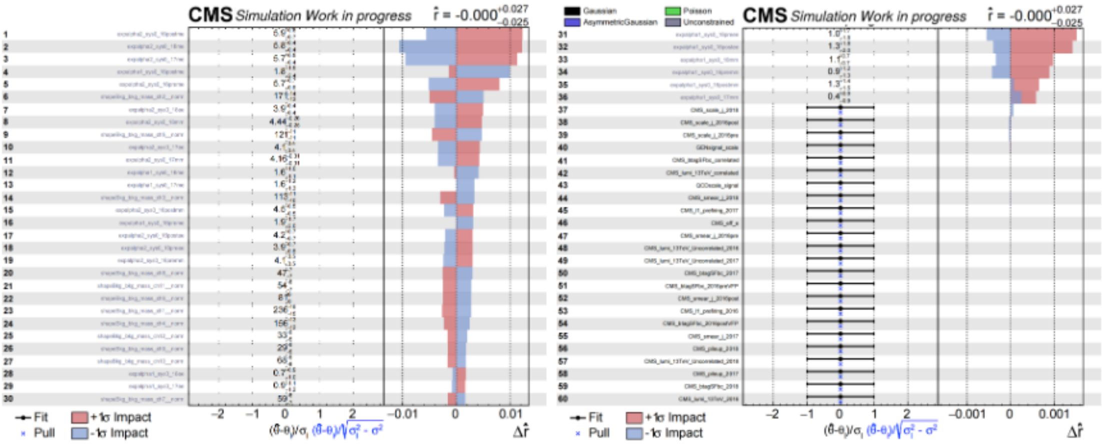

Figure 5.6: Impact plots with Asimov dataset (signal strength r=0) at 900 GeV mass point combining full Run 2.

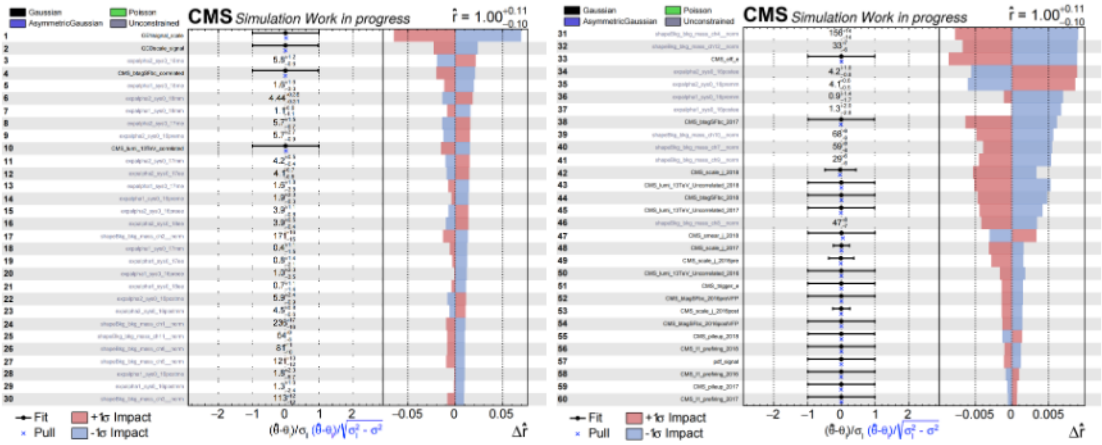

Figure 5.7: Impact plots with Asimov dataset (signal strength r=1) at 1100 GeV mass point combining full Run 2.

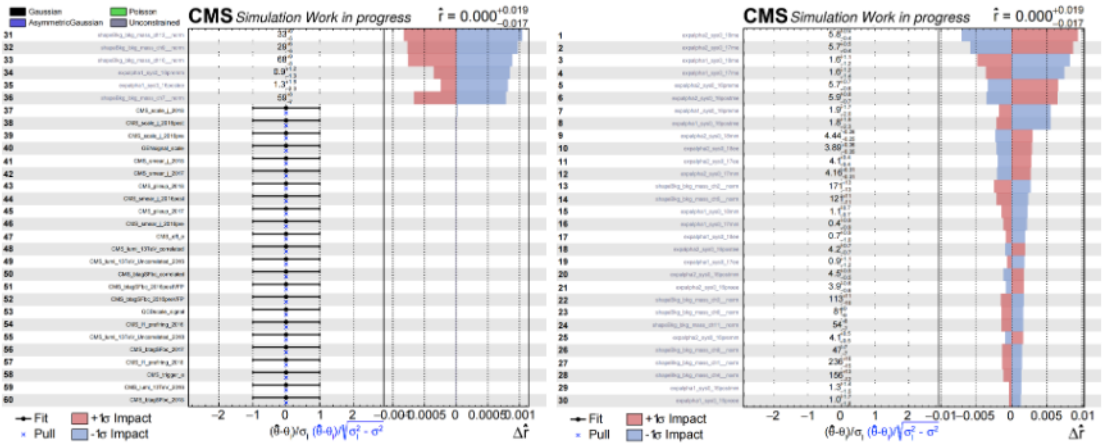

Figure 5.8: Impact plots with Asimov dataset (signal strength r=0) at 1100 GeV mass point combining full Run 2.

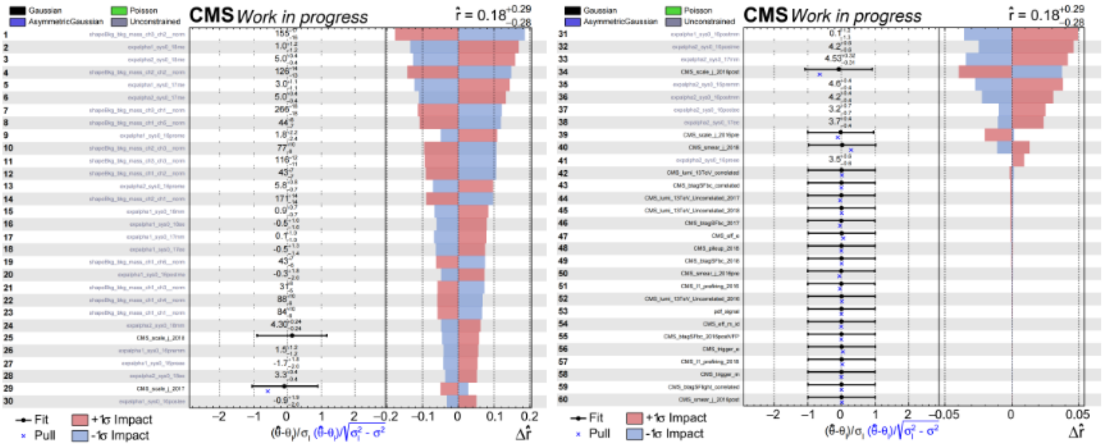

Figure 5.9: Impact plots after unblinding at 600 GeV mass point combining full Run 2.

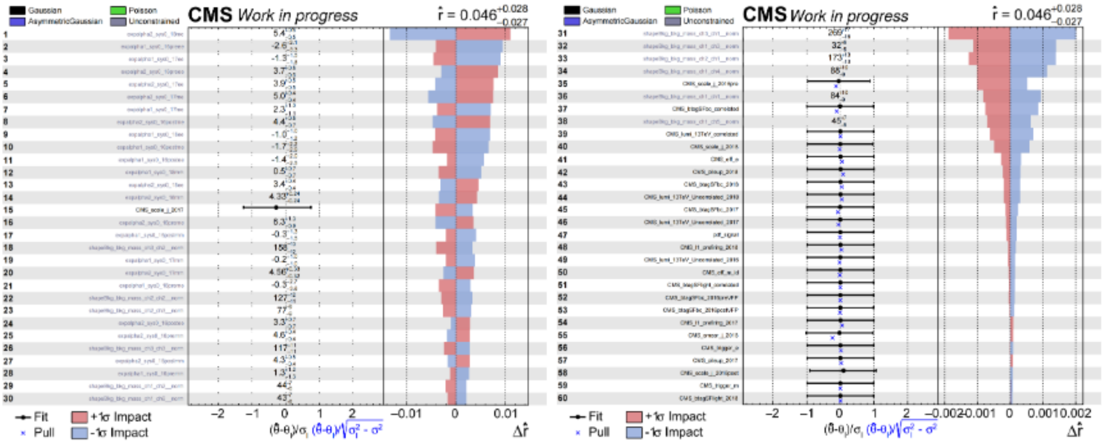

Figure 5.10: Impact plots after unblinding at 1200 GeV mass point combining full Run 2.

is used to set 95% CL upper limits on the T production cross section times branching ratio $\sigma(\mathrm{pp} \rightarrow \mathrm{T}) \cdot \mathscr{B}(\mathrm{T} \rightarrow \mathrm{tH})$, assuming SM Higgs branching fractions. Upper limits are computed using the statistical method discussed in Section 4.16, with the likelihood model constructed as described in Section 4.6.

Figure 5.12 shows the upper limits for the cross section $\sigma(\mathrm{pp} \rightarrow \mathrm{T}) \cdot \mathscr{B}(\mathrm{T} \rightarrow \mathrm{tH})$ in three OS dilepton channels (ee, $\mu\mu$, and $e\mu$). In each plot, there is an "Expected" dashed line with a yellow and blue error band, an "Observed" black line, and a gray dashed line with an error band. The "Expected" line is computed with the Asimov dataset, showing the median upper limits at a CL of 95%. The yellow and blue bands show the 1- and 2-standard-deviation intervals around the median upper limit values. The "Observed" black points are from signal-region data and show the upper limit results. All upper limit results from the observation are within the 2 standard deviation band, except one mild excess at 700 GeV in the ee channel due to statistical fluctuation. The gray dashed line shows the theoretical prediction assuming $\Gamma/m_T = 5\%$ and $Br(T \rightarrow tH) = 25\%$. The strongest limits are obtained from the $e\mu$ channel, due to flavor combinatorics as well as the reduction of Drell–Yan background.

Figure 5.13 shows the upper limit results combining all channels in full Run 2. All results from observation agree with the null hypothesis within 2 standard deviations. Table 5.1 summarizes the upper limit results in Run 2, including the observed limits, the expected limits with error bands, and the theoretical predictions. The combined limits range from 2 pb at 600 GeV to about 0.1 pb at 1000 GeV. The observed limits are in good agreement with the expected limits within the uncertainty bands.

Figure 5.14 from [10] presents the published results of searching for a single vector-like top quark and their combination in Run 2, setting 95% CL upper limits on the single T production cross section. This analysis presents the first results from a search for T → tH

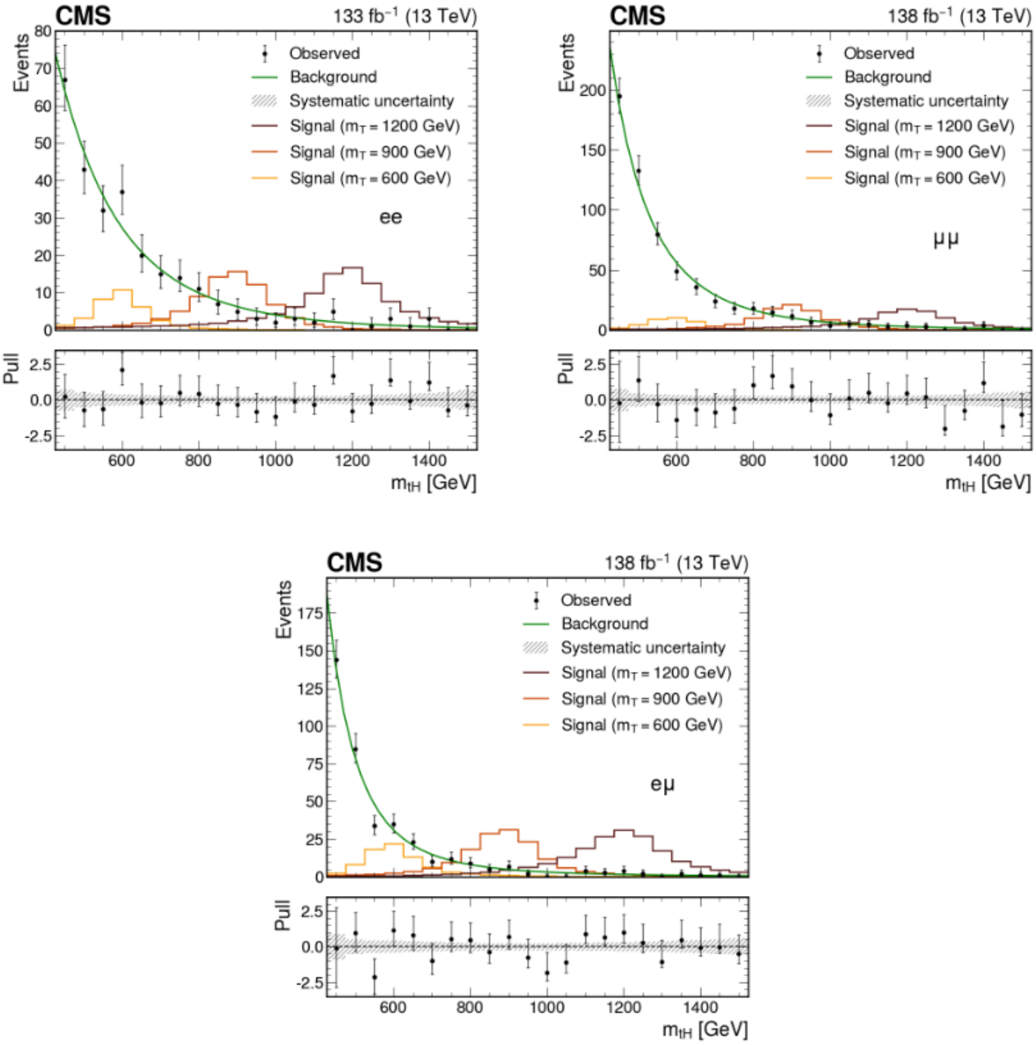

Figure 5.11: Post-fit distributions of the invariant mass m_{tH} in the ee (upper left), $\mu\mu$ (upper right) and $e\mu$ (bottom) channels in Run 2.

in opposite-sign dilepton final states, where electrons and muons are treated as leptons. The results will be combined with other T-decay channels. The overall sensitivity of the analysis is comparable to that of other similar channels searching for T → tH.

Mass points (GeV)		600	700	800	900	1000	1100	1200
$\mu\mu, e\mu, ee$ in Run 2	Observed	2.007	0.738	0.633	0.304	0.127	0.262	0.287
	Expected	1.625	0.727	0.450	0.303	0.225	0.183	0.177
	1 σ up	2.331	1.038	0.649	0.430	0.326	0.262	0.254
	2 σ up	3.167	1.428	0.902	0.597	0.451	0.371	0.349
	1 σ down	1.155	0.515	0.303	0.213	0.158	0.129	0.126
	2 σ down	0.863	0.384	0.220	0.148	0.112	0.094	0.093
$e\mu$ in Run 2	Observed	2.343	0.860	0.599	0.336	0.162	0.323	0.319
	Expected	2.129	0.863	0.534	0.384	0.277	0.222	0.215
	1 σ up	2.986	1.232	0.778	0.565	0.409	0.334	0.306
	2 σ up	4.034	1.705	1.094	0.788	0.570	0.477	0.427
	1 σ down	1.523	0.618	0.360	0.271	0.193	0.155	0.152
	2 σ down	1.147	0.465	0.252	0.195	0.139	0.113	0.113
$\mu\mu$ in Run 2	Observed	2.906	1.649	1.984	1.255	0.734	0.540	0.466
	Expected	4.189	1.863	1.206	0.716	0.606	0.495	0.499
	1 σ up	5.675	2.649	1.735	1.044	0.876	0.726	0.731
	2 σ up	7.358	3.627	2.376	1.456	1.212	1.040	1.029
	1 σ down	3.024	1.319	0.851	0.500	0.422	0.344	0.349
	2 σ down	1.996	0.982	0.631	0.366	0.312	0.253	0.256
ee in Run 2	Observed	6.149	5.153	1.644	0.549	0.407	0.602	0.784
	Expected	3.680	2.298	1.523	0.807	0.606	0.499	0.524
	1 σ up	5.132	3.297	2.148	1.141	0.886	0.743	0.751
	2 σ up	6.906	4.537	2.934	1.561	1.263	1.049	1.017
	1 σ down	2.668	1.642	1.090	0.575	0.426	0.346	0.363
	2 σ down	2.041	1.221	0.821	0.432	0.315	0.258	0.264
Theory	Nominal	0.875	0.443	0.229	0.126	0.073	0.043	0.027
	1 σ up	1.106	0.567	0.296	0.164	0.095	0.057	0.036
	1 σ down	0.707	0.355	0.182	0.099	0.057	0.034	0.021

Table 5.1: Observed and expected limit values on the production cross section $\sigma(\mathrm{pp} \rightarrow \mathrm{T})\mathscr{B}(\mathrm{T} \rightarrow \mathrm{tH})$ in Run 2.

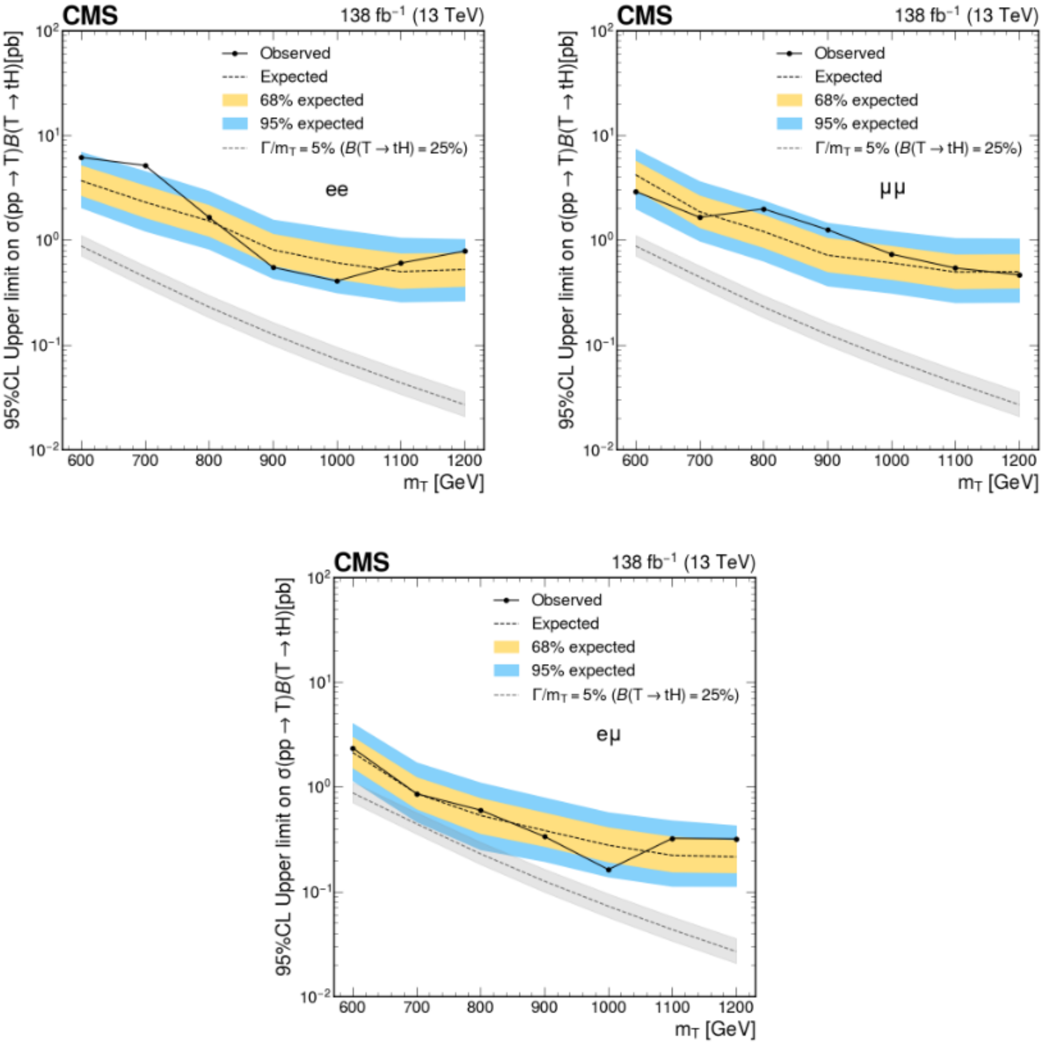

Figure 5.12: Observed upper limits at 95% CL on the production cross section $\sigma(\mathrm{pp} \rightarrow \mathrm{T})\mathscr{B}(\mathrm{T} \rightarrow \mathrm{tH})$, for the three OS dilepton channels ee (upper left), $\mu\mu$ (upper right) and $e\mu$ (bottom) in the whole Run 2 dataset.

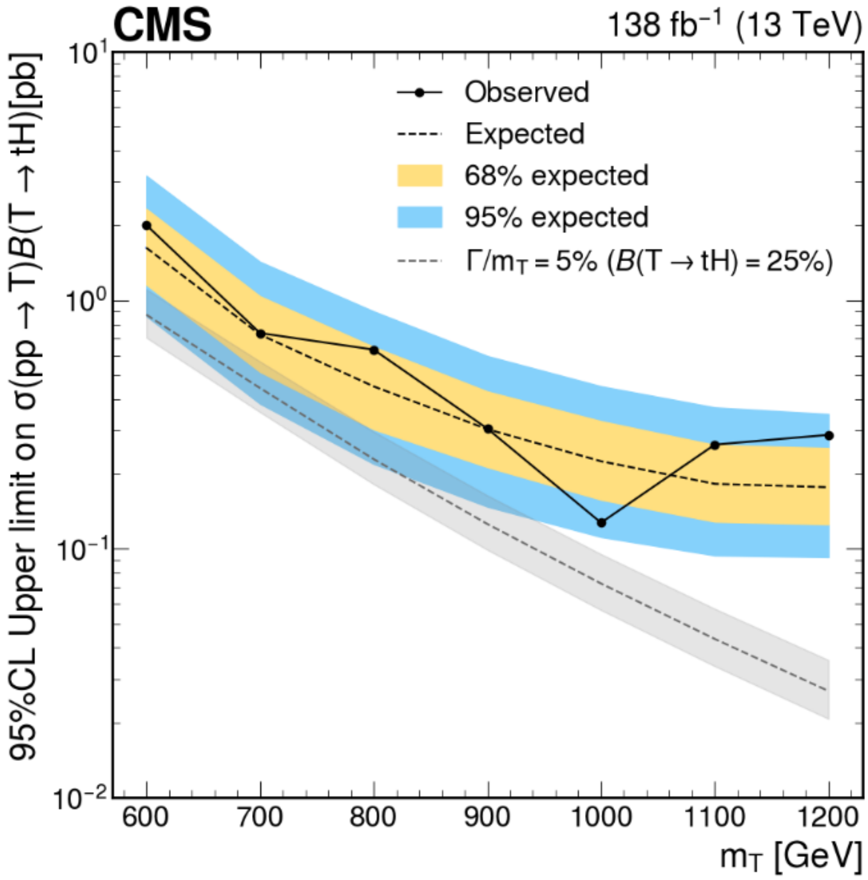

Figure 5.13: Observed upper limits at 95% CL on the production cross section $\sigma(\mathrm{pp} \to \mathrm{T})\mathscr{B}(\mathrm{T} \to \mathrm{tH})$, combining all OS dilepton channels in the Run 2 dataset.

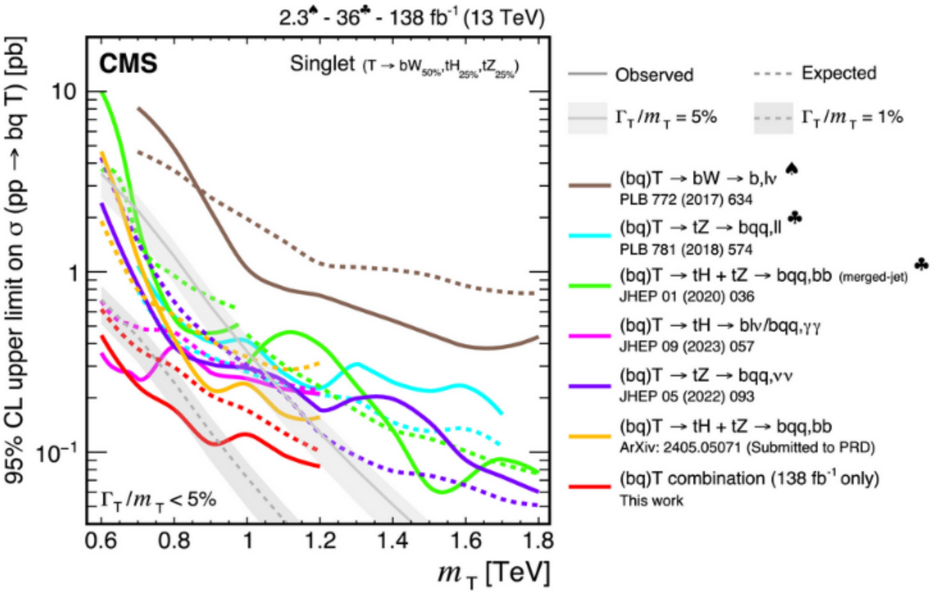

Figure 5.14: The 95% CL observed upper limits on the T single production cross section combining all channels in Run 2. The red curves show the combined results, the other curves show the individual results. Figures from Ref [10]

CHAPTER 6

Summary

The thesis presents an analysis searching for a vector-like top quark decaying into a top quark and a Higgs boson, in a final state containing two OS leptons, jets, and missing transverse momentum. The T search focuses on the T mass hypotheses from 600 GeV to 1200 GeV, assuming that the T signal has a narrow width. The integrated luminosity of the data utilized by the analysis is 138 fb^{-1}, collected by the CMS experiment during the LHC Run 2. Cut-based selection criteria are optimized to suppress the SM background. The main discriminating variable, the reconstructed T mass, is computed using a partial reconstruction algorithm under the neutrino kinematic assumptions. No significant signal is found, and the observed data agree well with the null hypothesis. Upper limits at 95 % confidence level on the T production cross section times the branching fraction of $\mathrm{T} \to \mathrm{tH}$ are set, ranging from 2 pb at a T mass of 600 GeV to about 0.1 pb at 1000 GeV.

The T decay channel in this study is the first time to be searched for at CMS, and the upper limit results are comparable to those of T's search in other decay modes. The analysis combines three channels, $\mu\mu$, $e\mu$, and ee, while the $e\mu$ channel provides the best sensitivity due to flavor combinations and reduced DY background. The results aim to be combined with other channels searching for T to improve the sensitivity of VLQ search. The current single T search in the OS dilepton final state has already shown good performance, and the analysis still has significant potential for future study. For instance, more data can be included to reduce statistical uncertainty, and a machine learning approach to event selection can improve discrimination between signal and background.

Furthermore, VLQ searches are a promising topic at the HL-LHC [10], where the expected integrated luminosity will reach 3000 fb^{-1} and particle identification technology, especially for boosted objects with high momenta, will be improved. According to the simulation-based study in Ref. [75], the pair production of T in HL-LHC could be discovered up to a T mass of 1440 GeV with a significance of five standard deviations at CMS. If such a model does not exist, the T mass below 1775 GeV from pair production will be excluded at 95% CL. In any case, the future VLQ study will enrich the knowledge of particle physics.

Acknowledgements

It has been a great pleasure to work and study in this excellent group, where we collaborate to explore particle physics and enjoy research. First of all, I would like to thank my supervisors, Dr. Rainer Mankel and Prof.Dr. Elisabetta Gallo, for offering me the great opportunity to stay in this group and constantly supporting me during the past three years. We always get together to discuss the research project, from the technical details to the overall strategy. All the scientific discussions we had together have become valuable and delightful memories of my life. I would like to thank my colleagues and friends from the Lyon group, Dr. Susan Gascon-Shotkin, Jieun, and Benjamin, for their cooperation on the analysis project and their warm welcome during my visit to Lyon. I would like to thank my office mates and friends in this group, Daina, Aliya, Stepan, Jacopo, Haoyu, Mathilde, Lakshmi, Suman, and many more. We had a lot of fun in the coffee breaks, the group dinners, and the Christmas market. I would like to send my deepest thanks to my high school and college friends, who have shown me that friendship does not decrease with physical distance. Finally, I would like to thank my family, especially my mother, my uncle, and my aunts, for always believing in me, loving me, and raising me into a strong and independent person.

CHAPTER 7

Appendix

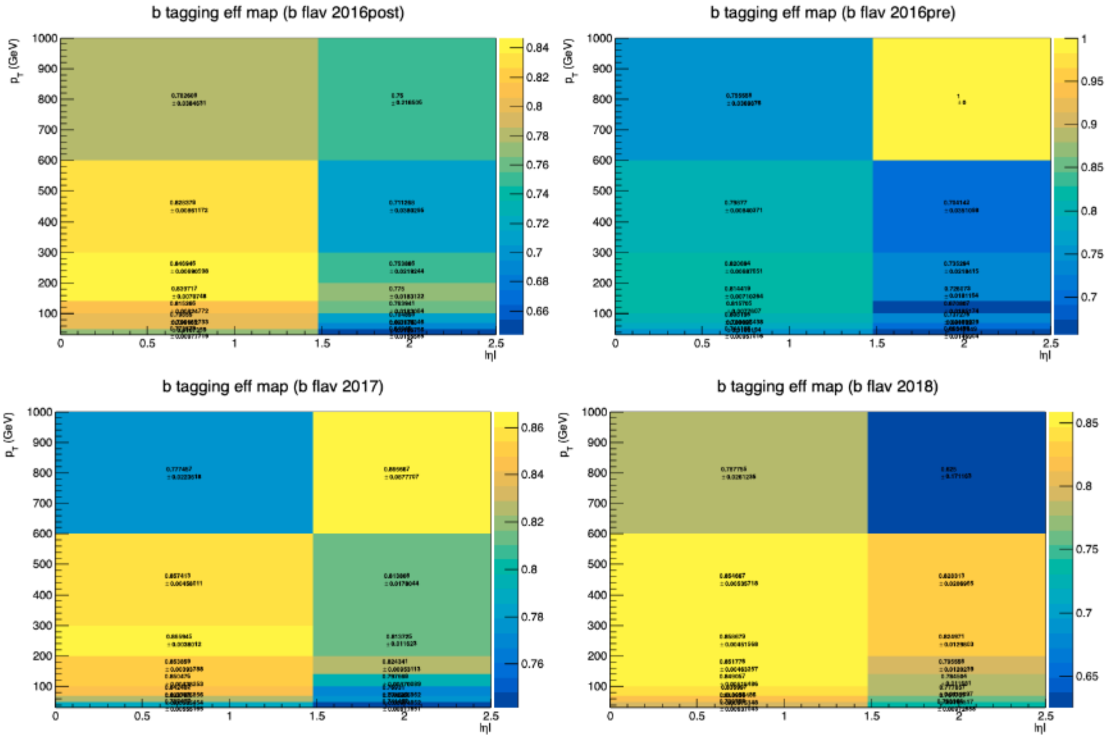

Figure 7.1: B-tagging efficiency maps for each data-taking era. Jets in the maps have hadron flavor 5.

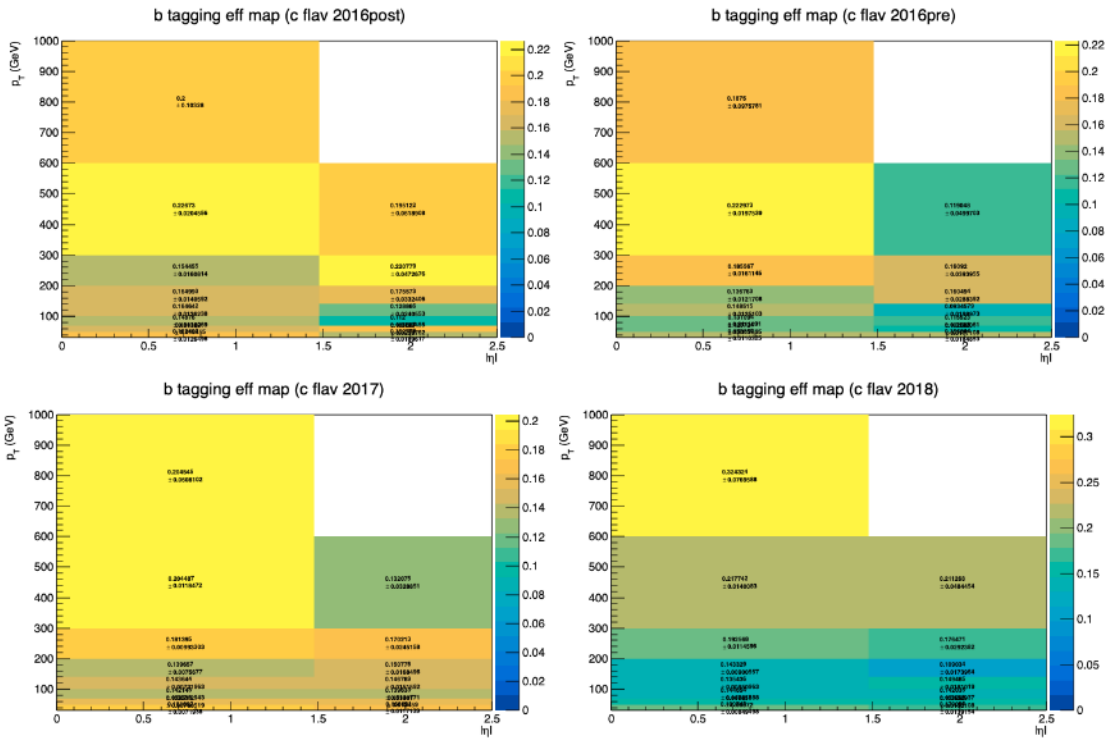

Figure 7.2: B-tagging efficiency maps for each data-taking era. Jets in the maps have hadron flavor 4.

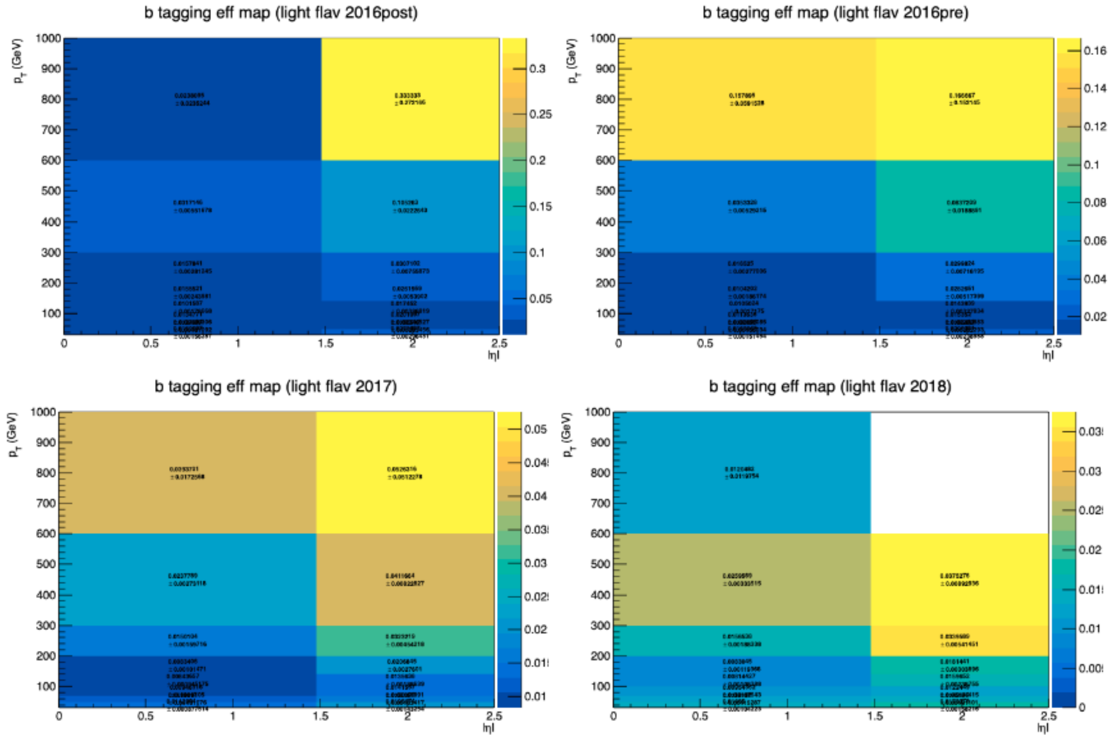

Figure 7.3: B-tagging efficiency maps for each data-taking era. Jets in the maps have hadron flavor 0.

PROCEEDINGS OF SCIENCE

Searches for vector-like quarks and leptons at CMS

Di Wang[a,*] **on behalf of the CMS Collaboration**

[a] *Deutsches Elektronen-Synchrotron (DESY),*
Notkestrasse 85, Hamburg, Germany

E-mail: di.wang@cern.ch

New fermions, referred to as vector-like quarks and vector-like leptons, are hypothesized to address multiple questions ranging beyond the standard model, such as the hierarchy problem. These models are the subject of extensive searches by the CMS experiment. To date, no statistically significant excess has been observed. Upper limits are set on the production cross sections. This talk presents the latest analyses searching for vector-like quarks and vector-like leptons, utilizing data collected by the CMS experiment at the CERN LHC, in proton-proton collisions at $\sqrt{s} = 13$ TeV.

The European Physical Society Conference on High Energy Physics (EPS-HEP2025)
7-11 July 2025
Marseille, France

[*] Speaker

1. Introduction

The standard model (SM) of particle physics has been proven to be a successful theory by many experiments. Nevertheless, there are many open questions that cannot be answered within the SM framework, such as the hierarchy problem and the origin of neutrino masses. To address these questions, fermions beyond those of the SM are hypothesized, including vector-like quarks (VLQs) and vector-like leptons (VLLs).

The VLQs are hypothetical quarks for which the left- and right-handed chiral components transform in the same way under the SM electroweak group. Their production cross section depends on the coupling between the VLQ and the SM particle. They can be produced singly or in pairs at the LHC and have been extensively searched for by the CMS experiment [1]. The pair production of VLQs proceeds typically through the strong interaction, as shown in the upper plots in Figure 1. The single production of VLQ can occur through a t-channel exchange of a W or Z boson via electroweak interaction, as shown in the lower left plot in Figure 1. In presence of vector bosons beyond the SM, the single VLQ production might also proceed through a W' boson via new interactions, as shown in the lower right plot in Figure 1.There are usually four types of VLQs considered, named T, B, X, and Y. The allowed decay modes for each of them are:

$$\begin{gathered}\mathrm{T} \to \mathrm{bW^+},\ \mathrm{T} \to \mathrm{tZ},\ \mathrm{T} \to \mathrm{tH} \\ \mathrm{B} \to \mathrm{tW^-},\ \mathrm{B} \to \mathrm{bZ},\ \mathrm{B} \to \mathrm{bH} \\ \mathrm{X}_{5/3} \to \mathrm{tW^+},\ \mathrm{Y}_{4/3} \to \mathrm{bW^-}\end{gathered} \tag{1}$$

The VLLs are color-singlet counterparts of the VLQs. They can appear as either SU(2) doublets, (E,N), or as singlets, E, where E is the electrically charged and N is the neutral state. In the singlet models, the VLLs can be produced in pairs via the electroweak interaction, $\mathrm{pp} \to \mathrm{E\bar{E}}$. In the doublet models, they can be produced in pairs through the process $\mathrm{pp} \to \mathrm{E\bar{E}}/\mathrm{N\bar{N}}$, or produced in association, $\mathrm{pp} \to \mathrm{E\bar{N}}/\mathrm{N\bar{E}}$. An E or N particle decays into a SM neutral boson and a lepton, following the decay modes below:

$$\begin{gathered}\mathrm{E} \to \mathrm{Z}\ell,\ \mathrm{E} \to \mathrm{H}\ell \\ \mathrm{N} \to \mathrm{W}\ell\end{gathered} \tag{2}$$

We present the latest analyses searching for VLQ and VLL in proton-proton collisions at a center-of-mass energy of 13 TeV at the CMS experiment in Run 2 (from 2016 to 2018). Section 2 introduces a VLL search with long-lived particle decays, Section 3 presents three analyses searching for singly produced VLQs, and Section 4 shows the combination results.

2. Search for VLLs

The analysis searches for pair-produced vector-like τ leptons named τ' [3], which decay into SM tau leptons (τ) and long-lived pseudoscalars (a_τ). The a_τs are beyond standard model (BSM) particles. The τ' are produced via s-channel electroweak process with a Z boson or a photon. In the signal process, a τ' decays into τa_τ, then the a_τ decays into two photons via a τ' loop diagram after a few meters of traveling from the interaction point (Figure 2). The final state consists of

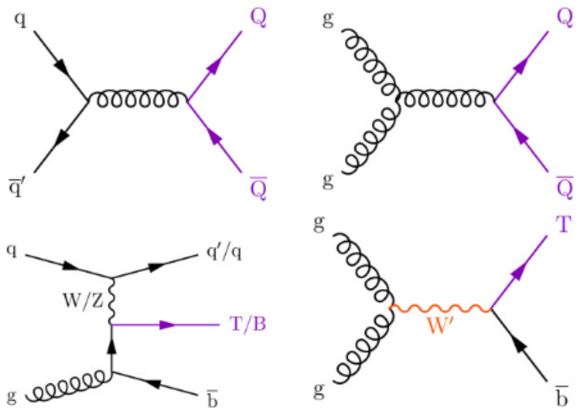

Figure 1: Feynman diagrams showing the productions of VLQ. Figure from Ref. [8].

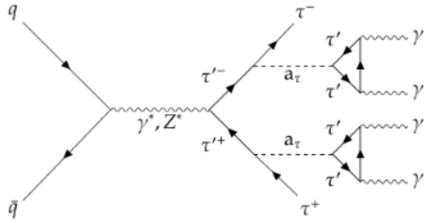

Figure 2: Feynman diagrams showing the productions of VLL, decaying into a_τ and τ. Figure from Ref. [3].

hadronically decaying tau leptons τ_h and displaced photons, which create particle showers in the CMS muon detector system. The discriminating variable is $\mathrm{N_{hits}}$, which refers to the number of reconstructed muon detector hits in a cluster representing the shower. The signal clusters produce a large number of reconstructed hits compared to the background clusters, as can be seen in the left plot of Figure 3. The analysis is categorized into drift tube and cathode strip chambers, according to the two parts of the CMS muon detector in which the clusters are searched for. No excess of data above SM predictions is observed. Upper limits at the 95% confidence level (CL) are set on the VLL production cross section, as a function of the VLL mass (m_{VLL}) and the a_τ lifetime ($c\tau_a$). The mass of a_τ is assumed to be 2 GeV. As shown in the right plot of Figure 3, the observed VLL mass exclusion is set at around 690 GeV, depending on the a_τ lifetime assumption.

3. Search for VLQs

The first VLQ analysis searches for single production of T or Y, decaying into a W boson and a bottom (b) quark, where the W boson decays into a lepton and a neutrino [4]. Six event categories are defined based on the combination of b-tagged jets and the lepton information in the final states. The discriminating variable is the VLQ invariant mass reconstructed using the recursive jigsaw algorithm [5]. A neural network multiclass classifier is developed to suppress backgrounds dominated by W+jets, $t\bar{t}$, and single top production. Each background shape is modeled by fitting the MC simulated background distributions. The left plot in Figure 4 shows the VLQ invariant mass distribution in one of the six categories. No signal is observed. Upper limits at the 95% CL on the coupling parameter κ_W are set, as shown in the right plot of Figure 4. The results provide the

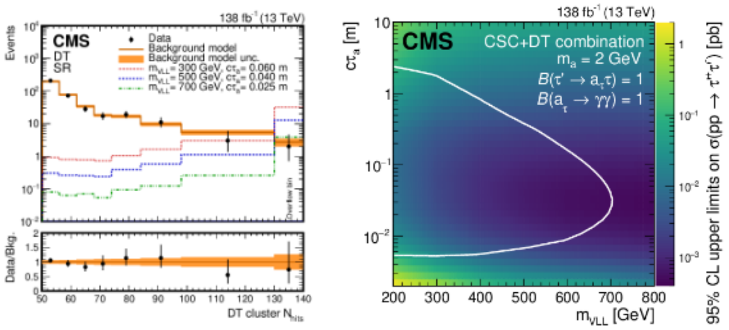

Figure 3: Left: Number of reconstructed hits in signal region of the drift tube category in signal region; Right: 95% CL observed upper limits on the VLL production cross section. Figures from Ref. [3].

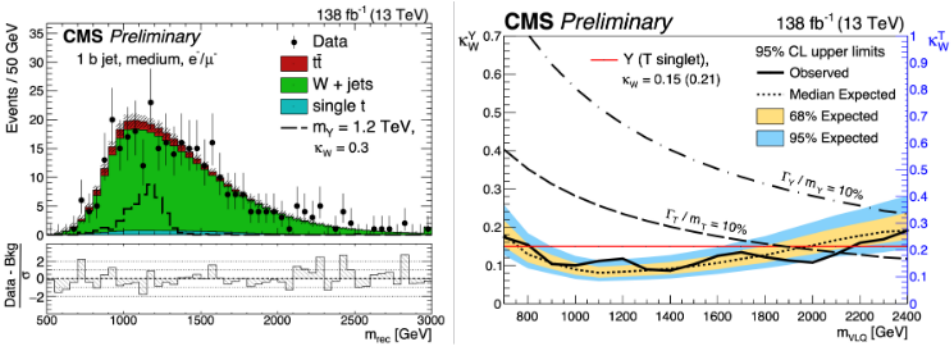

Figure 4: Left: m_{rec} distribution in signal region after the maximum likelihood fit; Right: The 95% CL observed upper limits on the VLQ coupling parameter. Figures from Ref. [4].

most stringent limits on single production of Y/T → bW, and exclude the hypothesis from Ref. [9] for the coupling in the (B, Y) doublet by the preferred electroweak fit.

The second VLQ analysis targets single production of T decaying to a top quark and a neutral scalar boson [6], where the top quark decays leptonically and the neutral scalar boson decays into a b quark-antiquark pair. The neutral scalar boson can be an SM Higgs boson or a new BSM particle labeled as ϕ. The main variable for signal extraction is the reconstructed T mass, where the ϕ boson is reconstructed as a large-radius jet. 16 regions, including 8 signal regions, are defined based on the top quark tagging, the b-jet tagging, and the number of forward jets. The left plot in Figure 5 shows the distributions of the reconstructed mass of the T candidate in one of the signal region categories, where the data shows good agreement with the background-only hypothesis. A multiclass boosted decision tree is trained for top tagging to discriminate the signal from the main background, including $t\bar{t}$, W+jets, single top, and QCD multi-jets. No excess is observed in the signal region, and the data agree with the background-only hypothesis. Upper limits are set at the 95% CL on the T production cross-section times the branching ratio of T→ tϕ → b$\ell\nu$bb, as a function of T mass and ϕ mass, as shown in the right plot of Figure 5. In case the neutral scalar boson is an SM Higgs boson, the T mass hypothesis below 1200 GeV is excluded assuming a singlet T quark of resonance width Γ of 5% of the mass.

The third VLQ analysis searches for T → tH/Z in all-hadronic final states [7]. The main

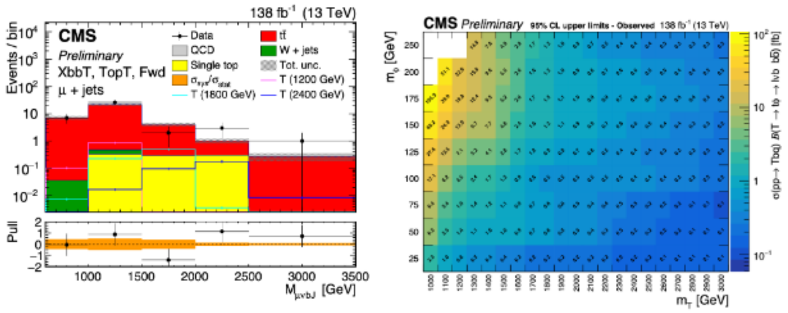

Figure 5: Left: Reconstructed mass of the T candidate in the signal region from one of the categories; Right: The 95% CL observed upper limits on the T production cross section as a function of the T and ϕ masses. Figures from Ref. [6].

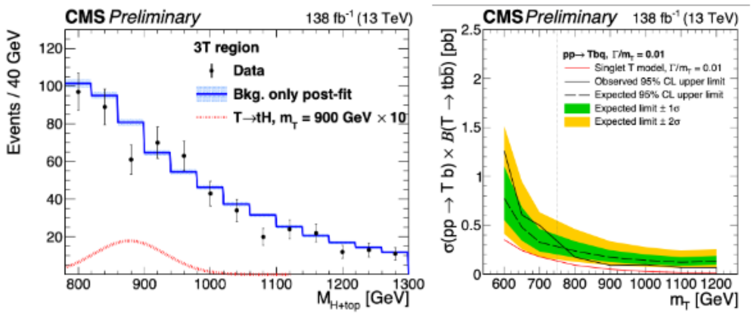

Figure 6: Left: Reconstructed mass of the T candidate in the signal region; Right: The 95% CL observed upper limits on the T production cross section combining all channels in Run 2. Figures from Ref. [7].

observable is the T candidate mass reconstructed from five jets, which are selected by matching their reconstructed mass to the known top quark and Higgs boson mass values. The cut-based event selection criteria are optimized for the low-mass and high-mass regions, respectively, suppressing background mainly rarising from QCD multijet and $t\bar{t}$ production. The background model is derived from data in the relaxed b-tagging region. The left plot in Figure 6 shows the reconstructed mass distributions of the T candidate with results from fits under the background-only hypothesis. No statistically significant signal is found. Upper limits are set at 95% CL on the T production cross-section combining the all-hadronic channels from the T → tH and T → tZ modes in Run 2, as shown in the right plot of Figure 6.

4. Combination of VLQ and VLL searches

The review paper [8] summarizes the current CMS searches for new fermions, including VLQs, VLLs, and heavy neutral leptons, in Run 2. The left plot in Figure 7 shows the combination of several searches targeting pair produced vector-like b quarks, setting 95% CL upper limits on the production cross section of $B\bar{B}$, assuming a B → tH branching fraction of 100%. The right plot in Figure 7 shows the combination of several single vector-like top quarks searches in three T decay modes, setting upper limits on the T production cross-section. Additionally, the discovery potential for new fermions at the High-Luminosity LHC is discussed in the paper.

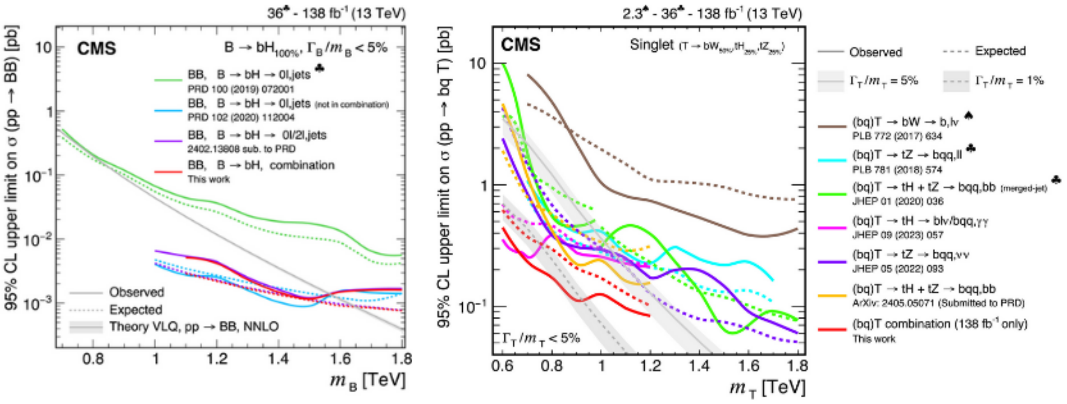

Figure 7: Left: Reconstructed T mass in signal region; Right: The 95% CL observed upper limits on the T production cross section combining all channels in Run 2. The red curves show the combined results. Figures from Ref. [8].

References

[1] CMS Collaboration, "The CMS experiment at the CERN LHC", JINST 3, S08004 (2008).

[2] CMS Collaboration, A. Hayrapetyan et al., "Review of searches for vector-like quarks, vector-like leptons, and heavy neutral leptons in proton–proton collisions at $\sqrt{s}$ =13 TeV at the CMS experiment." Phys. Rept. 1115 (2025) 570, [arXiv:2405.17605].

[3] CMS Collaboration, V. Chekhovsky et al., "Search for vector-like leptons with long-lived particle decays in the CMS muon system in proton-proton collisions at $\sqrt{s}$ =13 TeV." JHEP 08 (2025) 156, [arXiv:2503.16699].

[4] CMS Collaboration, "Search for single production of vector-like quarks decaying into a W boson and a b quark using the single-lepton final states in proton-proton collisions at 13 TeV".

[5] Jackson and C. Rogan, "Recursive jigsaw reconstruction: HEP event analysis in the presence of kinematic and combinatoric ambiguities", Phys. Rev. D 96 (2017) 112007.

[6] CMS Collaboration, "Search for single production of a vector-like T quark decaying to a top quark and a neutral scalar boson in lepton+jets final states at $\sqrt{s}$ =13 TeV".

[7] CMS Collaboration, "Search for production of single vector-like quarks decaying to tH or tZ in the all-hadronic final state in pp collisions at sqrt(s) = 13 TeV".

[8] CMS Collaboration, "Review of searches for vector-like quarks, vector-like leptons, and heavy neutral leptons in proton–proton collisions at $\sqrt{s}$ = 13 TeV at the CMS experiment." Phys. Rept. 1115 (2025) 570–677

[9] J. A. Aguilar-Saavedra, R. Benbrik, S. Heinemeyer, and M. Perez-Victoria, "A handbook of vectorlike quarks: mixing and single production", Phys. Rev. D 88 (2013) 094010

Bibliography

[1] CMS Collaboration, “Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC.” *Phys. Lett. B* **716** (2012) 30–61, [arXiv:1207.7235].

[2] ATLAS Collaboration, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC.” *Phys. Lett. B* **716** (2012) 1–29, [arXiv:1207.7214].

[3] Particle Data Group Collaboration, S. Navas et al., “Review of particle physics.” *Phys. Rev. D* **110** (2024), no. 3 030001. 11. Status of Higgs Boson Physics.

[4] N. Arkani-Hamed, A. G. Cohen, E. Katz, and A. E. Nelson, “The Littlest Higgs.” *JHEP* **07** (2002) 034, [hep-ph/0206021].

[5] J. P. H. Daza, *Composite Higgs models*, Ph.D. thesis, Sao Paulo U. (2019). arXiv:1908.10204.

[6] M. Perelstein, M. E. Peskin, and A. Pierce, “Top quarks and electroweak symmetry breaking in little Higgs models.” *Phys. Rev. D* **69** (2004) 075002, [hep-ph/0310039].

[7] R. Contino, L. Da Rold, and A. Pomarol, “Light custodians in natural composite Higgs models.” *Phys. Rev. D* **75** (2007) 055014, [hep-ph/0612048].

[8] O. Matsedonskyi, G. Panico, and A. Wulzer, “Light Top Partners for a Light Composite Higgs.” *JHEP* **01** (2013) 164, [arXiv:1204.6333].

[9] J. A. Aguilar-Saavedra, R. Benbrik, S. Heinemeyer, and M. Pérez-Victoria, “Handbook of vectorlike quarks: Mixing and single production.” *Phys. Rev. D* **88** (2013), no. 9 094010, [arXiv:1306.0572].

[10] CMS Collaboration, “Review of searches for vector-like quarks, vector-like leptons, and heavy neutral leptons in proton–proton collisions at $\sqrt{s}$=13 TeV at the CMS experiment.” *Phys. Rept.* **1115** (2025) 570–677, [arXiv:2405.17605].

[11] CMS Collaboration, A. Tumasyan et al., "Search for single production of a vector-like T quark decaying to a top quark and a Z boson in the final state with jets and missing transverse momentum at $\sqrt{s} = 13$ TeV." *JHEP* **05** (2022) 093, [arXiv:2201.02227].

[12] CMS Collaboration, A. M. Sirunyan et al., "Search for electroweak production of a vector-like T quark using fully hadronic final states." *JHEP* **01** (2020) 036, [arXiv:1909.04721].

[13] P. W. Higgs, "Spontaneous Symmetry Breakdown without Massless Bosons." *Phys. Rev.* **145** (1966) 1156–1163.

[14] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons." *Phys. Rev. Lett.* **13** (1964) 321–323.

[15] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons." *Phys. Rev. Lett.* **13** (1964) 508–509.

[16] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global Conservation Laws and Massless Particles." *Phys. Rev. Lett.* **13** (1964) 585–587.

[17] A. Bednyakov, "Quantum Field Theory and the Electroweak Standard Model." *CERN Yellow Rep. School Proc.* **6** (2019) 1–41, [arXiv:1812.10675].

[18] E. Noether, "Invariant Variation Problems." *Gott. Nachr.* **1918** (1918) 235–257, [physics/0503066].

[19] Particle Data Group Collaboration, S. Navas et al., "Review of particle physics." *Phys. Rev. D* **110** (2024), no. 3 030001. 12. CKM Quark–Mixing Matrix.

[20] J. Ellis, "Higgs Physics." arXiv:1312.5672. 52 pages, 45 figures, Lectures presented at the ESHEP 2013 School of High-Energy Physics, to appear as part of the proceedings in a CERN Yellow Report.

[21] S. L. Glashow, J. Iliopoulos, and L. Maiani, "Weak Interactions with Lepton-Hadron Symmetry." *Phys. Rev. D* **2** (1970) 1285–1292.

[22] M. Perelstein, "Little Higgs models and their phenomenology." *Prog. Part. Nucl. Phys.* **58** (2007) 247–291, [hep-ph/0512128].

[23] N. Arkani-Hamed, A. G. Cohen, and H. Georgi, "Electroweak symmetry breaking from dimensional deconstruction." *Phys. Lett. B* **513** (2001) 232–240, [hep-ph/0105239].

[24] A. Djouadi and A. Lenz, "Sealing the fate of a fourth generation of fermions." *Phys. Lett. B* **715** (2012) 310–314, [arXiv:1204.1252].

[25] O. Eberhardt, G. Herbert, H. Lacker, A. Lenz, A. Menzel, U. Nierste, and M. Wiebusch, "Joint analysis of Higgs decays and electroweak precision observables in the Standard Model with a sequential fourth generation." *Phys. Rev. D* **86** (2012) 013011, [`arXiv:1204.3872`].

[26] J. A. Aguilar-Saavedra, "Effects of mixing with quark singlets." *Phys. Rev. D* **67** (2003) 035003, [`hep-ph/0210112`]. [Erratum: Phys.Rev.D 69, 099901 (2004)].

[27] CMS Collaboration, "The CMS hadron calorimeter project: Technical Design Report.".

[28] L. Evans and P. Bryant, "LHC machine." *Journal of Instrumentation* **3** (aug, 2008) S08001–S08001.

[29] E. Lopienska, "The CERN accelerator complex, layout in 2022.". General Photo.

[30] CMS Collaboration, "CMS Luminosity Public Results."

[31] CMS Collaboration, "The CMS Experiment at the CERN LHC." *JINST* **3** (2008) S08004.

[32] CMS Collaboration, "CMS Detector Introduction."

[33] CMS Collaboration, "CMS corrdinate system."

[34] CMS Tracker Group Collaboration, "The CMS Phase-1 Pixel Detector Upgrade." *JINST* **16** (2021), no. 02 P02027, [`arXiv:2012.14304`].

[35] CMS Collaboration, "The CMS electromagnetic calorimeter project: Technical Design Report.".

[36] CMS Collaboration, A. Benaglia, "The CMS ECAL performance with examples." *JINST* **9** (2014) C02008.

[37] P. Adzic et al., "Energy resolution of the barrel of the CMS electromagnetic calorimeter." *JINST* **2** (2007) P04004.

[38] CMS Collaboration, *The CMS hadron calorimeter project.* Technical design report. CMS. CERN, Geneva (1997).

[39] CMS Collaboration, "Performance of the CMS Hadron Calorimeter with Cosmic Ray Muons and LHC Beam Data." *JINST* **5** (2010) T03012, [`arXiv:0911.4991`].

[40] CMS Collaboration, "Calibration of the CMS hadron calorimeter in Run 2." *JINST* **13** (2018), no. 03 C03025.

[41] CMS Collaboration, “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV.” *JINST* **13** (2018), no. 06 P06015, [arXiv:1804.04528].

[42] CMS Collaboration, “CMS Technical Design Report for the Muon Endcap GEM Upgrade.”.

[43] CMS Collaboration, “Performance of CMS Muon Reconstruction in pp Collision Events at $\sqrt{s} = 7$ TeV.” *JINST* **7** (2012) P10002, [arXiv:1206.4071].

[44] CMS Collaboration, M. Tosi, “The CMS trigger in Run 2.” *PoS* **EPS-HEP2017** (2017) 523.

[45] CMS Collaboration, “Particle-flow reconstruction and global event description with the CMS detector.” *JINST* **12** (2017), no. 10 P10003, [arXiv:1706.04965].

[46] T. Speer, W. Adam, R. Fruhwirth, A. Strandlie, T. Todorov, and M. Winkler, “Track reconstruction in the CMS tracker.” *Nucl. Instrum. Meth. A* **559** (2006) 143–147.

[47] CMS Collaboration, “Description and Performance of Track and Primary-Vertex Reconstruction with the CMS Tracker.” *JINST* **9** (2014), no. 10 P10009, [arXiv:1405.6569].

[48] K. Rose, “Deterministic annealing for clustering, compression, classification, regression, and related optimization problems.” *IEEE Proc.* **86** (1998), no. 11 2210–2239.

[49] S. A. S. c. Frühwirth, R., “Secondary vertex reconstruction.” *Pattern Recognition, Tracking and Vertex Reconstruction in Particle Detectors. Particle Acceleration and Detection* (2020).

[50] CMS Collaboration, “Muon identification using multivariate techniques in the CMS experiment in proton-proton collisions at sqrt(s) = 13 TeV.” *JINST* **19** (2024), no. 02 P02031, [arXiv:2310.03844].

[51] CMS Collaboration, “Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC.” *JINST* **16** (2021), no. 05 P05014, [arXiv:2012.06888].

[52] W. Adam, R. Fruhwirth, A. Strandlie, and T. Todorov, “Reconstruction of electrons with the Gaussian sum filter in the CMS tracker at LHC.” *eConf* **C0303241** (2003) TULT009, [physics/0306087].

[53] M. Cacciari, G. P. Salam, and G. Soyez, “The anti-k_t jet clustering algorithm.” *JHEP* **04** (2008) 063, [arXiv:0802.1189].

[54] CMS Collaboration, "Jet Performance in pp Collisions at 7 TeV.".

[55] CMS Collaboration, "Pileup Removal Algorithms.".

[56] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV." *JINST* **12** (2017), no. 02 P02014, [arXiv:1607.03663].

[57] CMS Collaboration, "Identification of b-Quark Jets with the CMS Experiment." *JINST* **8** (2013) P04013, [arXiv:1211.4462].

[58] E. Bols, J. Kieseler, M. Verzetti, M. Stoye, and A. Stakia, "Jet Flavour Classification Using DeepJet." *JINST* **15** (2020), no. 12 P12012, [arXiv:2008.10519].

[59] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations." *J. High Energy Phys.* **07** (2014) 079, [arXiv:1405.0301].

[60] S. Bashir and A. Oblakowska-Mucha, "Impact on Multiplicity of Particles by Changing Multiparton Interaction Parameters in PYTHIA 8.3 at LHC Energies." *Acta Phys. Polon. Supp.* **16** (2023), no. 3 2.

[61] GEANT4 Collaboration, S. Agostinelli et al., "GEANT4 - A Simulation Toolkit." *Nucl. Instrum. Meth. A* **506** (2003) 250–303.

[62] J. Allison et al., "Recent developments in Geant4." *Nucl. Instrum. Meth. A* **835** (2016) 186–225.

[63] CMS Collaboration, "Muon Trigger efficiency map."

[64] J. A. Aguilar-Saavedra, "Laboratory-frame tests of quantum entanglement in H→WW." *Phys. Rev. D* **107** (2023), no. 7 076016, [arXiv:2209.14033].

[65] J. Ellis, R. Fok, D. S. Hwang, V. Sanz, and T. You, "Distinguishing 'Higgs' spin hypotheses using $\gamma\gamma$ and WW^* decays." *Eur. Phys. J. C* **73** (2013) 2488, [arXiv:1210.5229].

[66] Particle Data Group Collaboration, R. M. Barnett et al., "Review of particle physics. Particle Data Group." *Phys. Rev. D* **54** (1996), no. 1 1–720.

[67] G. Punzi, "Sensitivity of searches for new signals and its optimization." *eConf* **C030908** (2003) MODT002, [physics/0308063].

[68] P. D. Dauncey, M. Kenzie, N. Wardle, and G. J. Davies, "Handling uncertainties in background shapes: the discrete profiling method." *JINST* **10** (2015), no. 04 P04015, [arXiv:1408.6865].

[69] T. Junk, "Confidence level computation for combining searches with small statistics." *Nucl. Instrum. Meth. A* **434** (1999) 435–443, [`hep-ex/9902006`].

[70] A. L. Read, "Presentation of search results: The CL_s technique." *J. Phys. G* **28** (2002) 2693–2704.

[71] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics." *Eur. Phys. J. C* **71** (2011) 1554, [`arXiv:1007.1727`]. [Erratum: Eur.Phys.J.C 73, 2501 (2013)].

[72] CMS Collaboration, "The CMS Statistical Analysis and Combination Tool: Combine." *Comput. Softw. Big Sci.* **8** (2024), no. 1 19, [`arXiv:2404.06614`].

[73] W. Verkerke and D. P. Kirkby, "The RooFit toolkit for data modeling." *eConf* **C0303241** (2003) MOLT007, [`physics/0306116`].

[74] L. Moneta, K. Belasco, K. S. Cranmer, S. Kreiss, A. Lazzaro, D. Piparo, G. Schott, W. Verkerke, and M. Wolf, "The RooStats Project." *PoS* **ACAT2010** (2010) 057, [`arXiv:1009.1003`].

[75] CMS Collaboration, K. K. Pal, "Search for Vector-Like Quark T Decaying to bW, tZ, tH in the Single Lepton Final State at the HL-LHC." *Springer Proc. Phys.* **304** (2024) 1287–1290.