SEARCH FOR DARK MATTER PRODUCED IN ASSOCIATION WITH BOTTOM QUARKS IN $\sqrt{s} = 13$ TeV PROTON-PROTON COLLISIONS WITH THE CMS DETECTOR

A Thesis
Submitted For the Degree of
MASTER OF SCIENCE

by
Spandan Mondal
Undergraduate Programme,
Indian Institute of Science

under the supervision of
Dr. Jyothsna Rani Komaragiri
Centre for High Energy Physics,
Indian Institute of Science,
BANGALORE - 560012.

APRIL, 2018

DECLARATION

I hereby declare that I am the sole author of this thesis. I authorize Indian Institute of Science to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Candidate's Signature:

Date:

Place:

Acknowledgements

I am grateful to several people for their support throughout the duration of this thesis. I would like to mention some people who have particularly helped me in overcoming the challenges I faced during the production of this work.

Firstly, I would like to thank my supervisor, Dr. Jyothsna Rani Komaragiri, for guiding me throughout the project and helping me sort out several problems I faced. I would also like to thank Dr. Raman Khurana (National Central University, Taiwan) who actively guided me through every technical aspect of the analysis presented in this thesis.

Secondly, I would like to thank my fellow group members, Praveen Chandra Tiwari and Deepak Kumar, for their suggestions and involvement in studies related to the analysis. I would also like to thank my other colleagues, Lata Panwar and Zoheb Abai, for helping maintain a healthy working and learning environment.

I am thankful to the conveners of the MET+X subgroup at CMS for providing feedback on the analysis and endorsing my thesis on behalf of CMS. I would also like to thank the conveners and other members of the BTV group at CMS for their assistance and valuable feedback on the cMVAv2 retraining work described in this thesis.

I am indebted to my parents and family for their unconditional support and encouragement in my work and in every other aspect of life. I am also thankful to my closest friends for their support and encouragement.

Lastly, I would like to thank all the faculty members associated with the Undergraduate Programme, Indian Institute of Science, for providing a launchpad for exciting career opportunities in research.

Abstract

Dark matter is hypothesized to constitute about 80% of the total matter in the universe. While existence and effects of dark matter are quite evident in astrophysical observations, it has not yet been experimentally observed. Besides direct and indirect searches for dark matter, collider experiments also present us with a possibility to produce and detect dark matter in the laboratory.

A search for dark matter produced in association with bottom quarks in proton-proton collisions at a center of mass energy of $\sqrt{s} = 13$ TeV is presented in this thesis. The data collected by the Compact Muon Solenoid (CMS) experiment at CERN in 2016 is used for the analysis. We look at a final state containing high missing transverse energy and one or two jets produced from bottom quarks. Several control regions are defined to model the major backgrounds present in the signal regions. Hence, the signal is extracted through simultaneous fits on the signal and control regions. The results are interpreted by finding Asimov limits to the production cross section of dark matter in this channel for various mass points of the mediator, for both scalar and pseudoscalar mediators, using the simplified dark matter model.

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

SM	Standard Model
LHC	Large Hadron Collider
CMS	Compact Muon Solenoid
MC	Monte Carlo
MET	Missing Transverse Energy
pT	Transverse component of Momentum (p_T)
Eta	Pseudorapidity (η)
DM	Dark Matter
bb/tt	Bottom/Top quark-antiquark pair ($\mathrm{b\bar{b}}/\mathrm{t\bar{t}}$)
DY	Drell-Yan
SR	Signal Region
CR	Control Region
LO	Leading Order
NLO	Next-to-Leading Order
PU	Pile-up

1 Introduction

Mankind's inquisitiveness of what fundamentally constitutes the universe culminated into the proposition of the Standard Model (SM) of particle physics, which is arguably the most elaborate theory ever developed by mankind. The Standard Model theorizes the existence of multiple elementary particles and describes the several possible interactions between them. While most of the predictions of the SM have been found to be in agreement with results of experiments spanning over the last four decades, several observations point towards the existence of physics beyond the SM. One such fundamentally interesting observation is that of dark matter.

Many large scale collaborative experiments have been designed to probe into the most fundamental interactions of elementary particles. These experiments are usually conducted in large accelerators that collide particles at high energies. The Large Hadron Collider (LHC) at CERN is one such accelerator that collides protons with protons at the highest achievable energy till date. The Compact Muon Solenoid (CMS) is one of the two large general-purpose experiments at the LHC that uses a set of complex detectors to investigate a wide range of topics in particle physics.

The experiments at the LHC may be able to detect dark matter particles that are produced in proton-proton collisions. Since dark matter particles would not interact with any of the detectors, they would be detected as a large amount of missing energy. Thus, any missing energy detected in excess of the usual Standard Model backgrounds could be a potential signal for dark matter.

1.1 Project Statement

A search for dark matter produced in association with bottom quarks is carried out using $35.5\,\text{fb}^{-1}$ of data collected by the Compact Muon Solenoid (CMS) experiment in 2016, in proton-proton collisions at a center of mass energy of $\sqrt{s} = 13$ TeV. We look for a final state containing high missing transverse energy (E_T^{miss}) and one or two jets arising from bottom quarks.

1.1.1 Project Motivation

Even though numerous astrophysical observations have provided overwhelming evidence towards the ubiquity of non-luminous matter in the universe, both direct and indirect searches for dark matter have failed to conclusively detect new particles that could be potential dark matter candidates. Under the assumption that a certain interaction exists between SM particles and the DM candidate particle, proton-proton collisions at high center of mass energies are likely to produce dark matter particles at the LHC. Thus, searches for dark matter produced in colliders with collision data recorded by the CMS experiment could provide us with a first-hand detection of the elusive dark matter candidates.

1.1.2 Aims of the Project

In this project, we analyze about 35.5 fb^{-1} of CMS 2016 data to study a final state of high missing transverse energy and one or two jets arising from bottom quarks (b-jets) and no leptons. We define two distinct signal regions with one and two b-jets respectively. We aim to estimate the backgrounds in the two signal regions by modelling 12 mutually exclusive control regions. Further, the signal efficiencies of the dark matter signal in each of these two regions are calculated for various combinations of dark matter mass and mediator mass. Finally, we aim to calculate the uncertainties and derive limits for the production cross section of dark matter with b-quarks by performing a simultaneous fit on the signal and control regions.

To summarize, the main objectives of the project are:

1. Isolating events matching the signal region criteria from CMS 2016 data and MC simulation, for 2 signal regions.
2. Isolating events matching the control region criteria from CMS 2016 data and MC simulation, for 12 control regions.
3. Calculating the signal efficiencies of the dark matter signal using MC simulation, for various combinations of dark matter and mediator masses.
4. Calculating systematic uncertainties and deriving limits for the production cross section of dark matter with b-quarks.

1.2 Thesis Outline

The thesis is organized under several sections. This section serves as a general introduction. Section 2 provides the theoretical background for the project, such as a brief description of dark matter and its production in colliders. Section 3 provides the experimental background including description of the Large Hadron Collider and the CMS detector, focusing on the detector aspects relevant to this thesis. Sections 4 and 5 describe the data samples and signal and background MC samples we use in this analysis and show some preselection properties of the signal. Section 6 describes the selections we use to separate signal events from backgrounds and shows some signal region and signal efficiency plots. Section 7 deals with the control regions which we set up to model backgrounds in the signal region. In section 8 we look at the systematic uncertainties in our measurements. In section 9 we describe the signal extraction strategy and show some limits derived from this analysis. Section 10 describes retraining of the cMVAv2 b-tagger for CMS phase-2 upgrade, which is another of my contributions to the CMS experiment.

2 Theoretical Principles

2.1 Dark Matter

Several astrophysical observations have provided evidence for the existence of non-luminous, non-baryonic matter in the universe, called dark matter [1]. Even though cosmological calculations predict that nearly 80% of the matter in the universe is dark matter, it has not been directly observed and its nature remains unknown till date.

Several models and candidates for dark matter have been proposed. One common hypothesis for a dark matter candidate is that of weakly interacting massive particles (WIMPs) [2]. WIMPs are hypothesized to interact via gravity and perhaps other forces that may not be part of the standard model and are at least as weak as the weak nuclear force. Thus, when produced in collider experiments, WIMPs are expected to leave all the instruments undetected, thereby creating an excess of missing energy in the event.

Simplified dark matter models [3] refer to models that assume the existence of a mediator particle that couples to both SM and DM sectors, as opposed to the Effective Field Theory approach that considers an effective vertex mediated by kinematically inaccessible particles. The mediator can be either a scalar or a pseudoscalar particle that is not a part of the SM. Simplified models are typically characterized by the following free parameters:

1. mass of the mediator (M_{med} or M_{ϕ}),
2. mass of the DM candidate particle (M_{DM} or M_{χ}),
3. coupling of the mediator particle to the SM sector (g_{SM}), and,
4. coupling of the mediator particle to the DM sector (g_{DM}).

2.2 Dark Matter production in association with heavy flavour quarks

This thesis focuses on the production of dark matter in association with bottom quarks in proton-proton collisions at the LHC. We look at a simplified dark matter model describing the production of a dark matter candidate particle pair ($\chi\bar{\chi}$) through a scalar or pseudoscalar mediator (ϕ), along with a pair of bottom quarks. The process is similar to dark matter production with a pair of top quarks. The leading order Feynman diagram for both these processes are shown in figure 1.

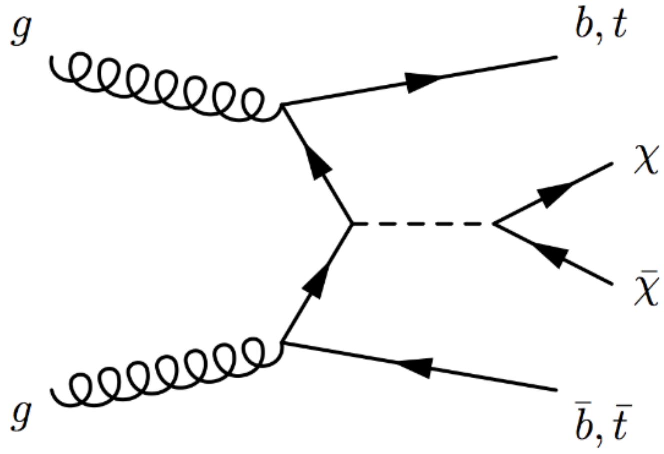


Figure 1: Leading order Feynman diagram for the production of a pair of DM candidate particles (χ) with heavy-flavor (top or bottom) quark pair via a scalar or pseudoscalar mediator (dashed line).

The final state contains the dark matter particle pair which would leave the detector undetected, along with a pair of bottom quarks recoiling against the dark matter pair. The two bottom quarks hadronize to two B-hadrons which have a small lifetime (order of 10^{-12} s) and further decay within the detector itself and appear as jets in the final state.

The kinematics of the final state particles are expected to depend on the model parameters. In general, the dark matter particles are expected to be more boosted for a higher mass of the mediator particle. This is expected to appear as a larger amount of missing energy in the detectors.

The Feynman diagram in figure 1 is constructed with the 4-flavour (4F) scheme, which assumes that the colliding protons are composed of only the four lightest quarks, thereby allowing only u, d, c and s quarks and gluons in the initial state. One could additionally have a 5-flavour (5F) scheme in which the initial state has a b-quark along with a gluon, and the final state contains dark matter particles produced in association with one bottom quark. This search focuses only on the 4F scheme of dark matter production with a pair of bottom quarks.

3 Experimental Background

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a particle collider built by the European Organization for Nuclear Research (CERN) in collaboration with over 10,000 scientists and engineers from all over world in between 1998 and 2008. The LHC is designed to collide protons with protons at a maximum center-of-mass energy of 14 TeV, which is the highest achievable energy till date. Located near Geneva, Switzerland, the LHC lies in a 27 kilometer tunnel situated in between 45 to 175 meters beneath the surface.

The collider was built with the aim to test the predictions of various theories in particle physics and find evidence for the existence of the Higgs boson, as predicted by the SM. It also serves to find evidence for theories beyond the SM, such as particles predicted by supersymmetric theories, dark matter searches, antimatter research, etc.

There are a total of seven experiments at the LHC that use detectors to analyze the particles produced by *p*-*p* collisions at the LHC. The largest of these, the ATLAS [4] and the CMS experiments [5], are general-purpose experiments that investigate a wide range of topics. These experiments use two independent detectors to probe into the events and cross-examine results. ALICE [6] and LHCb [7] are two other experiments that focus on the properties of quark-gluon plasma and b-physics respectively. The TOTEM [8] and LHCf [9] experiments study the forward particles (particles thrust forward by collisions). The MoEDAL [10] experiment was approved in 2010 to search for the hypothetical magnetic monopole.

3.2 The CMS Detector

The Compact Muon Solenoid is a general-purpose detector designed to detect various particles with a large angular coverage. The detector is about 21 meters long and 15 meters in diameter and has multiple components arranged in concentric layers with a cylindrical symmetry about the axis of the accelerator. The central feature of the detector is a 12.5 meter long superconducting solenoid providing a uniform magnetic field of 3.8 T [5].

3.2.1 Layers of the CMS

The CMS detector has multiple concentric layers around the central axis where the collision of two counter-rotating proton beams take place. A short description of the layers (in order of increasing distance from the central axis) are as follows:

1. Tracker: The silicon tracker records the path taken by a charged particle in a magnetic field. The curvature of the path gives information about the momentum of the particle: a more curved path indicating a lower momentum.
2. Electromagnetic Calorimeter: The Electromagnetic Calorimeter (ECAL) is designed to measure the energies of photons and electrons with high accuracy.
3. Hadronic Calorimeter: The Hadron Calorimeter (HCAL) measures the energy of hadrons (bound state of quarks, such as, protons, neutrons, pions, and other mesons and baryons).
4. Magnet: The superconducting solenoid is designed to provide a uniform and strong magnetic field of 3.8 T. Such high magnetic fields ensure a high resolution in the silicon tracker.
5. Muon chambers: To identify muons, three types of detectors, viz., drift tubes (DT), cathode strip chambers (CSC) and resistive plate chambers (RPC), are employed at the radial edge of the detector. They measure the presence and momenta of the muons, which are particles capable of penetrating through the calorimeters and the solenoid.

3.2.2 Trigger system

The proton-proton collision frequency at the LHC is around 40 MHz, which roughly corresponds to 40 Terabytes of raw data being generated per second. Since data storage capacities are not even nearly enough, it is necessary to save only a small fraction of the generated data that would be "interesting" for subsequent analysis. Moreover, the filtration for "interesting" events must be performed rapidly to ensure that the next stream of events are processed in the same way.

This gives rise to the necessity for a very fast, real-time filtration mechanism that can filter events with interesting indicative features and save them to the drive for further analysis. This is accomplished in two stages:

- Level-1 trigger: While the information of the event is held in buffers within the detectors, presence of interesting features are determined through fast, approximate hardware-level algorithms using certain key information from the buffer data. This lowers the event frequency by an order of magnitude of 3.
- High-level trigger: After the interesting events are filtered by the first level trigger, the significantly smaller subset of filtered events is processed with more sophisticated computer softwares, further reducing the frequency by an order of magnitude of 2. Finally, the data is stored on drives with a manageable frequency of a few KHz or less.

In general, several trigger paths are deployed during a run and an event that passes through any one of these triggers is saved for analysis. The information of which trigger path(s) selected the event is also stored for every event. This trigger information is also utilized in an analysis.

3.2.3 Event Reconstruction

Particles produced in an event are recognized from the signals they leave behind in the detectors while passing through them. For example, charged particles like electrons leave signals in the silicon tracker and exhibit curved paths owing to their interaction with the magnetic field. The curvature gives a measure of their momenta with good precision. Electrons and photons are absorbed in the ECAL giving a fairly precise measure of their energies. Neutral hadrons are not detected at the tracker but are fully absorbed at the HCAL and thus their energies can be measured too. Muons are the only detectable particles that reach the muon detectors, and thus their presence and momenta can be measured as well. Neutrinos, however, pass through all the layers and hence cannot be detected directly.

Muon reconstruction Muons leave tracks at the silicon detector and simultaneously at the muon detectors. Since muons are the only detectable particles that reach the muon detectors, they can be identified by matching tracks inside the silicon tracker with tracks in the muon detectors.

Jet reconstruction Proton-proton collisions result in a large number of jets. Whenever a single quark or gluon is produced from a hard collision, it hadronizes shortly to form a number of hadrons, which together continue to move with the initial momentum of the quark/gluon, thereby forming a cone-shaped cluster of hadrons called jets. Jets are reconstructed from the detected hadrons using software-based jet reconstruction algorithms, such as k_t, anti-k_t, Cambridge/Aachen. In this thesis, the jets were reconstructed using the anti-k_t algorithm [11] with a distance parameter of $R = 0.4$.

b-jet Tagging Immediately after the p-p hard collision at the primary vertex, the resulting quarks and gluons hadronize to form bound states of hadrons. Although a resulting jet is recorded by the detectors, it is, in general, not possible to conclude which kind particle resulted in the jet. However, B-hadrons formed from bottom quarks have just enough lifetime (order of 10^{-12} s) to travel a noticeable distance before decaying within the detector itself. Thus, presence of a secondary vertex, besides the primary vertex, is a characteristic of b-jets. Since the B-hadron is usually far heavier than its decay products, b-jets tend to be wider with higher jet multiplicities. Additionally, the B-hadron sometimes has a lepton in its decay chain. Based on these characteristics of a b-jet, several algorithms have been developed that assign a discriminator value to jets. The higher the value of this discriminator, the higher is the probability of the jet originating from a b-quark. In this analysis, we use the Combined Secondary Vertex version 2 (CSVv2) b-tagger with its medium working point (0.8484) which corresponds to 1% mistag rate for light jets.

Missing Transverse Energy and Neutrinos Weakly interacting particles, such as neutrinos or potential dark matter candidate particles, escape the instruments undetected. The only way to indirectly detect these particles is to conserve momenta of the initial and final states of the collision. However, since momentum along the axis of collision is unknown, only the initial and final transverse momenta can be measured and conserved. The missing energy in the transverse direction, $\vec{E}_T^{\text{miss}}$, would then be equal to the negative vectorial sum of the transverse energies of all escaping particles:

$$\vec{E}_T^{\text{miss}} = -\sum_i \vec{E}_T^i. \tag{1}$$

If there is only one missing particle, say a neutrino, we can approximately write

$$E_T^{\nu} = |\vec{E}_T^{\text{miss}}|. \tag{2}$$

3.2.4 Monte-Carlo Simulation of Collider Events

To estimate the distributions of the signal and several SM processes after applying a set of requirements, it is necessary to have a simulation of the signal and all background processes. A simulation of events also gives us an estimate of selection efficiency of signal after the same cuts are applied to the real data.

For this purpose, several Monte Carlo samples simulating Standard Model processes as well as dark matter production are used in this analysis. The CMS detector is simulated in a software toolkit called `GEANT4` [12, 13]. The simulated events are generated and hadronized using tools like MadGraph [14], pythia [15] and powheg [16]. To make the simulation at par with real CMS data, raw simulations are processed using the same reconstruction algorithm used to reconstruct collision data.

4 Signal

4.1 Signal Samples

Signal events in this analysis refer to events with a pair of dark matter candidate particles being produced in association with a pair of bottom quarks. Monte Carlo simulations are used to simulate the final state of these events as observed by the detector.

The signal MC samples used in this analysis conform to simplified dark matter models. All the datasets were produced in the `RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2` campaign. These samples have been generated for both scalar and pseudoscalar mediator hypotheses, for several mediator (ϕ) and dark matter candidate (χ) masses. Both the SM coupling (g_{SM}) and DM coupling (g_{DM}) of the mediator are taken as 1. These samples were produced in both Leading Order (LO) and Next-to-Leading Order (NLO) accuracies.

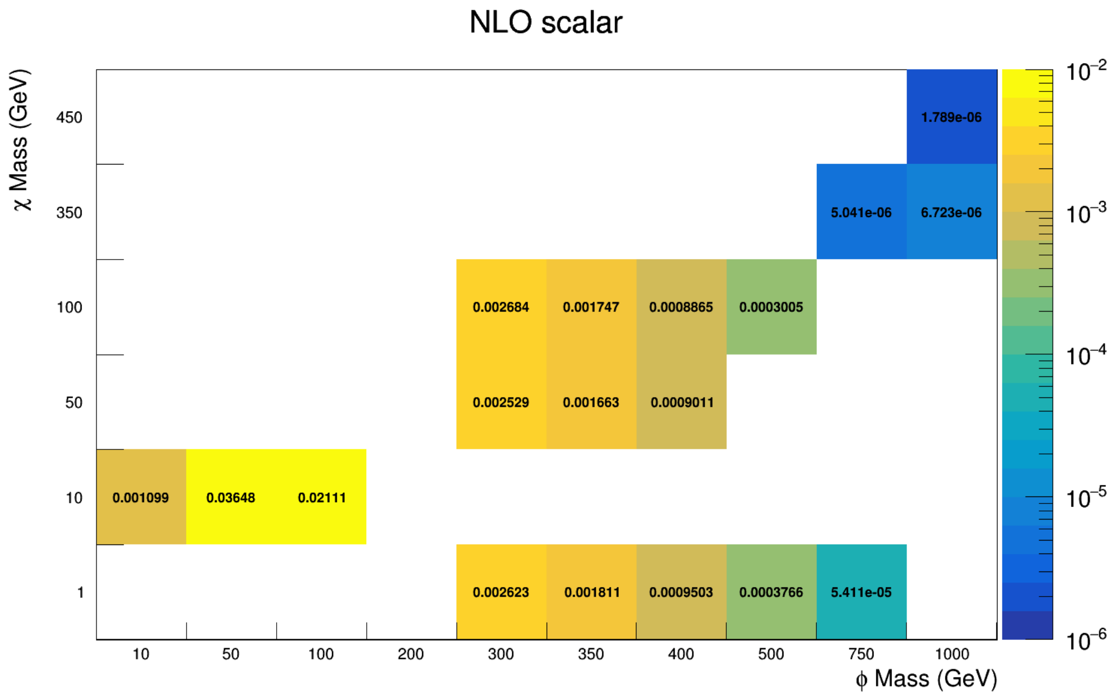
We shall mainly focus on the NLO samples in this analysis while presenting comparisons with the LO samples wherever necessary.

4.1.1 Cross sections

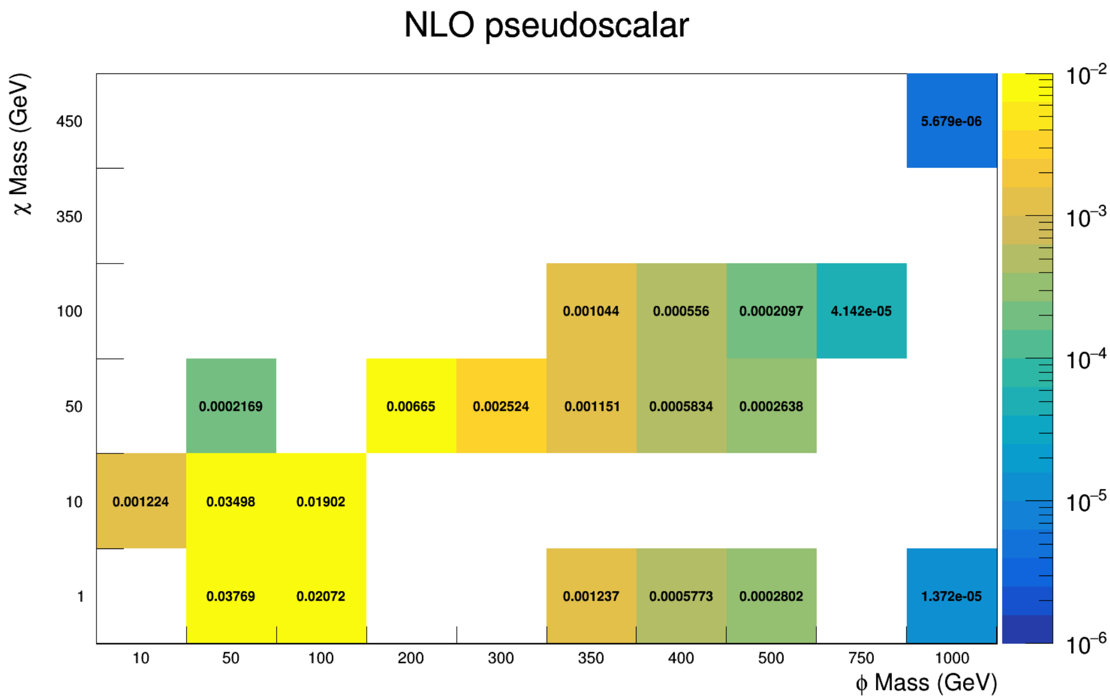
The cross sections of the various NLO signal samples being considered in this study for various mass points of the dark matter candidate and the mediator, for both scalar and pseudoscalar mediators, are represented graphically in figure 2.

4.1.2 Preselection distributions

The distributions of several event parameters for various mass points of dark matter particle and mediator are studied for both scalar and pseudoscalar mediators. The distributions of MET (E_T^{miss}), number of jets, number of b-tagged jets and lead jet p_T for $M_\chi = 1$ GeV are presented in figure 3.



(a)



(b)

Figure 2: Production cross section of dark matter candidate (χ) through mediator (ϕ) for various masses of χ and ϕ in picobarns for (a) scalar mediator and (b) pseudoscalar mediator, to Next-to-Leading Order (NLO) accuracy.

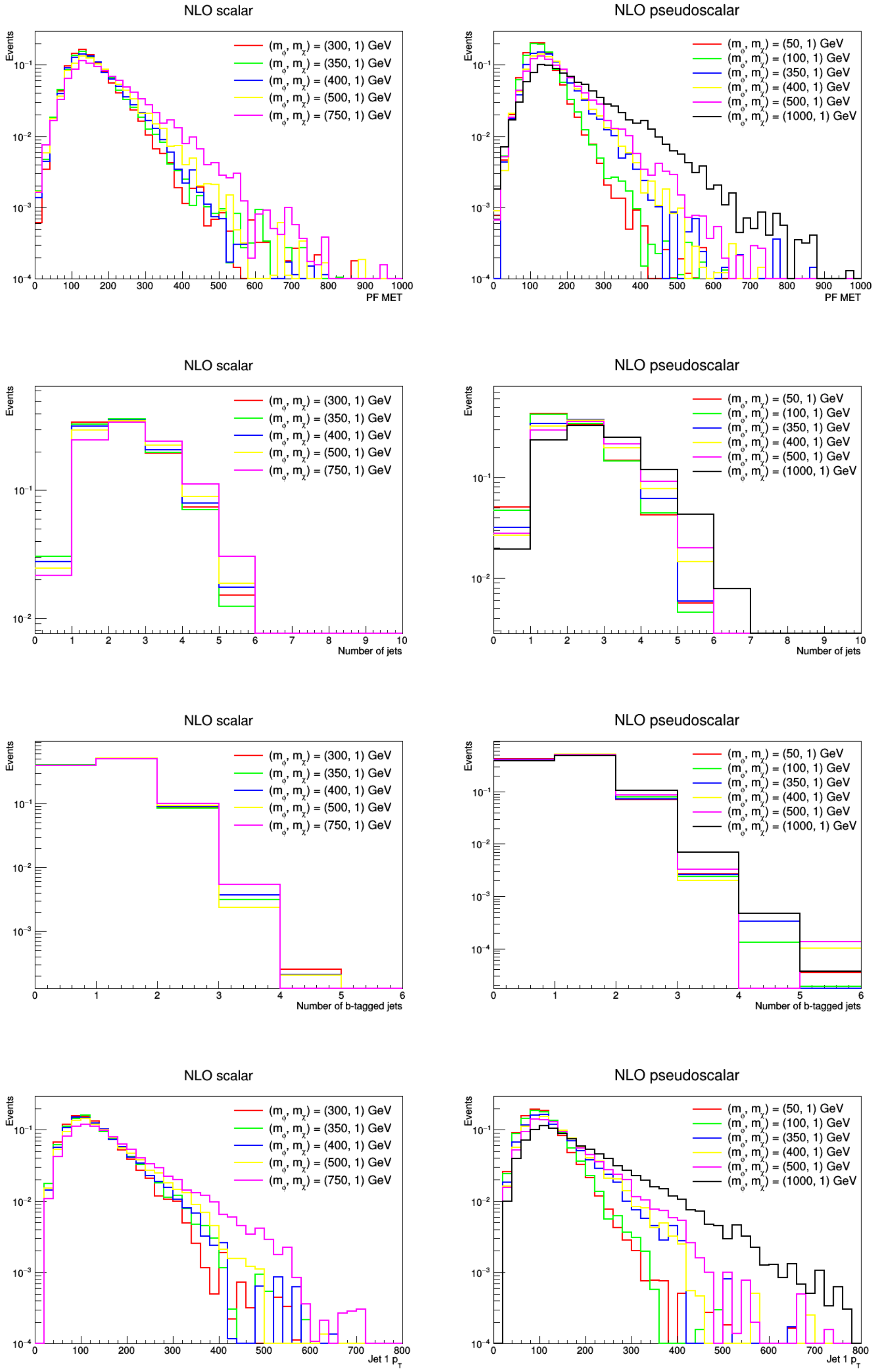


Figure 3: Preselection distributions of MET, number of jets, number of b-tagged jets and transverse momentum of leading jet in NLO dark matter signal samples, for $M_\chi = 1$ GeV and various mass points for ϕ, for scalar (left) and pseudoscalar (right) mediators. The total number of events is normalized to 1.

5 Data and MC Samples

5.1 MC Simulated Samples

All Standard Model processes that could potentially yield final states with high missing energy and one or two b-jets are considered as possible sources of background for this analysis. The major backgrounds relevant to dark matter production in association with heavy flavour quarks are Z and W boson production with jets, and top quark pair production. Single top, diboson and QCD multijet events also contribute to the background in small amounts.

- $Z(\nu\nu)$ + jets: Events with Z production with jets where the Z-boson decays into two neutrinos are the main irreducible background in the signal regions, owing to high missing transverse energy arising from the neutrinos.
- $W(l\nu)$ + jets: Events with jets along with a W-boson decaying to one charged lepton and a neutrino would contribute to the background for cases where the lepton is not detected.
- $t\bar{t}$ and single-top: Top decays are another source of background as the final states contain b-jets and W-bosons, resulting in missing energy.
- Diboson: Production of two vector bosons is rare, but has similar kinematics to that of the signal.
- γ + jets: Events with photons and jets in the final state where the photon is not detected are expected to be a minor background.
- QCD Multijet: The requirement of high missing energy gets rid of the majority of the multijet background, but it has some contribution in some control regions. We also define a control region for multijet to find the scale factor to correctly normalize multijet events in other signal and control regions.
- $Z(ll)$ + jets: DY events with two muons or electrons do not contribute to the signal region due to lepton veto, but are used in the control regions to model $Z(\nu\nu)$ + jets events.

5.2 Data Samples

Data samples used in this analysis were collected by CMS during 2016B through 2016H runs, at a center-of-mass energy of 13 TeV, in 25ns runs and with the magnet field enabled.

- The signal regions and the QCD multijet control regions are modelled using the MET primary dataset.
- The control regions involving muon triggers were modelled with both the MET and the SingleMuon datasets separately, and the results were found to be similar. The plots presented in this thesis have been generated using the MET dataset.
- The control regions involving electron triggers were modelled with the SingleElectron primary dataset.

Only the runs and lumisections that are certified as "good" are used in this analysis. The total integrated luminosity used is about 35.545 fb^{-1} which corresponds to nearly the entire 2016 data recorded by CMS.

5.3 Triggers

The events used in the analysis were selected by several different trigger paths depending on the region. The list of triggers used in each region is summarized in table 1.

Region	**Triggers (logical OR)**
SRs and QCD multijet CRs	HLT_PFMETNoMu90_PFMHTNoMu90_IDTight_v* HLT_PFMETNoMu120_PFMHTNoMu120_IDTight_v*
$Z \to \mu\mu$, $W \to \mu\nu$ and Top CRs	HLT_IsoMu24_v* HLT_IsoTkMu24_v*
$Z \to ee$, $W \to e\nu$ CRs	HLT_Ele27_WPLoose_Gsf HLT_Ele27_WPTight_Gsf

Table 1: Table showing the triggers used in various regions. Mutliple triggers in each region indicate that the logical OR of these triggers were used, i.e., the event is considered "triggered" for a region if it passes through any of the triggers corresponding to that region.

6 Event Selection

In this section, the event selection method for the signal regions is described. Distributions of some parameters for events selected in the signal region are also presented.

6.1 Physics Objects

6.1.1 Physics Object Selections

The physics objects used in this analysis are selected to meet certain criteria recommended by various CMS groups, to ensure accuracy of the objects. The selection criteria for each object is outlined in this section.

Jets Only AK4 (anti-k_t with $R = 0.4$) jets reconstructed by the Particle Flow algorithm [19, 20] are considered in this analysis. They are required to have a corrected p_T larger than 30 GeV and to lie within the $|\eta| < 4.5$ range. Further, the jets are required to pass the `Loose` jet identification criteria recommended for Run 2 analyses by the corresponding CMS group.

b-tagging The b-tagging algorithm used in this analysis is Combined Secondary Vertex version 2 (CSVv2). The medium working point of CSVv2 (0.8484) is used and it corresponds to 1% mistag rate for light jets. Further, jets are tagged as b-jets only if they lie in the tracker range of $|\eta| < 2.4$, as b-tagging performance is reliable only within the tracker range.

b-tagging performance is known to be different in data and MC. Hence, b-tag event weights corresponding to the medium working point as recommended by the respective CMS group are applied to every event to achieve better data-MC agreement.

Muons Muons used in this analysis are reconstructed using the Particle Flow algorithm. An object in the muon collection is considered a muon only when it has a p_T greater than 10 GeV, lies within $|\eta| < 2.4$ range, satisfies the `Loose` Muon ID requirement and has an isolation less than 0.25.

Muon isolation is defined as

$$I_{rel} = \left[\sum p_T^{\text{ch had}} + \max\left(0,\ \sum p_T^{\text{neu had}} + \sum p_T^{\gamma} - 0.5 \cdot \sum p_T^{\text{pu ch had}}\right)\right] / p_T^{l}$$

where $\sum p_T^{\text{ch had}}$ is the sum of the transverse momenta of the charged hadrons, $\sum p_T^{\text{neu had}}$ is the sum of transverse energies of the neutral hadrons, $\sum p_T^{\gamma}$ is the sum of the transverse energies of particle flow photons and $\sum p_T^{\text{pu ch had}}$ is the sum of transverse momenta of the charged particles in the cone of interest but with particles not originating from the primary vertex (for pileup corrections).

Electrons Electrons used in this analysis satisfy the Particle Flow criteria. Electrons are required to have a p_T greater than 10 GeV, lie in the ECAL pseudorapidity fiducial range of $|\eta| < 2.5$, and pass the `Loose` Electron ID requirement. No additional isolation requirement is required as the cut-based ID definition includes the isolation cuts as well.

Taus Taus are required to have p_T greater than 18 GeV and to lie within $|\eta| < 2.3$ range. A requirement to pass two quality cuts, viz., `DecayModeFinding` and `Loose` Tau ID, is also imposed.

Pile-up Usually several primary vertices are reconstructed in an event due to pileup. The MC simulations used in this analysis are generated simulating the expected PU conditions. However, the PU conditions in the MC do not exactly match those in data, and therefore an event-wise reweighting of MC events is applied to achieve better agreement with data.

6.1.2 Physics Object Cleaning

The physics objects that are selected using the methods and cuts outlined in the previous subsection are not always accurate. For instance, a fraction of the taus selected using the aforementioned criteria are actually electrons that were erroneously added to the tau collection. To mitigate such cases, we apply the following methods to clean the physics objects from potential fakes.

Tau cleaning Muons and electrons often enter the tau collection. In such cases, the object is found in both the muon/electron collection and the tau collection, with slight differences. To correct for this, we parse over all the selected electrons and muons in a given event and see if there is an object in the tau collection inside an $R = 0.4$ cone around each of the electrons and muons. If we do find one, we assume that this object was misidentified as a tau and remove it from the tau collection.

Jet cleaning Electrons are also known to enter the jet collection. To mitigate this effect, we parse over all the objects inside the jet collection in an event and see if there is an object in the electron collection within an $R = 0.4$ cone around the jet axis. If such an object is found, we assume that this object was in fact an electron and remove it from the jet collection.

6.2 Preselection

The analysis is performed in two steps. In the first step, some preliminary selection cuts are applied to vastly reduce the amount of data that needs to be processed in the final analysis programme in the second step. This step is called preselection.

To satisfy the preselection conditions for the signal region, events are required to have a reconstructed $E_T^{\text{miss}} > 200$ GeV, and at least one jet that satisfies the jet selection criteria (without cleaning against electrons). All events satisfying this preliminary cut are accepted and further analyzed in the main analysis routine.

6.3 Signal Region

Events in the signal region are required to be selected by at least one of the triggers mentioned in the relevant row in table 1. Further, the requirement of an offline reconstructed $E_T^{\text{miss}} > 200$ GeV is imposed. This cut gets rid of most of the QCD multijet background. Events containing one or more isolated electron and/or muon that satisfy the corresponding physics object selection outlined in section 6.1.1 are not accepted in the signal regions. This helps in removing background events with high E_T^{miss} and leptons, mostly $W \to l\nu$ events.

Two mutually exclusive regions, depending on the number of b-tagged jets in each event, are defined, as shown in figure 4.

- **SR1:** In the single b-tagged category, we require exactly one b-tagged jet in the event. We allow at most one additional jet satisfying the jet selection criteria outlined in section 6.1.1, in the event. The leading jet is required to satisfy $p_T > 50$ GeV and tight quality selections (charged hadron fraction > 0.1, neutral hadron fraction < 0.8). The sub-leading jet, if any, is simply required to satisfy the jet selection criteria.
- **SR2:** In the double b-tagged category, we require exactly two b-tagged jets in the event. We allow at most one additional jet satisfying the jet selection criteria outlined in section 6.1.1, in the event. The leading jet is required to satisfy

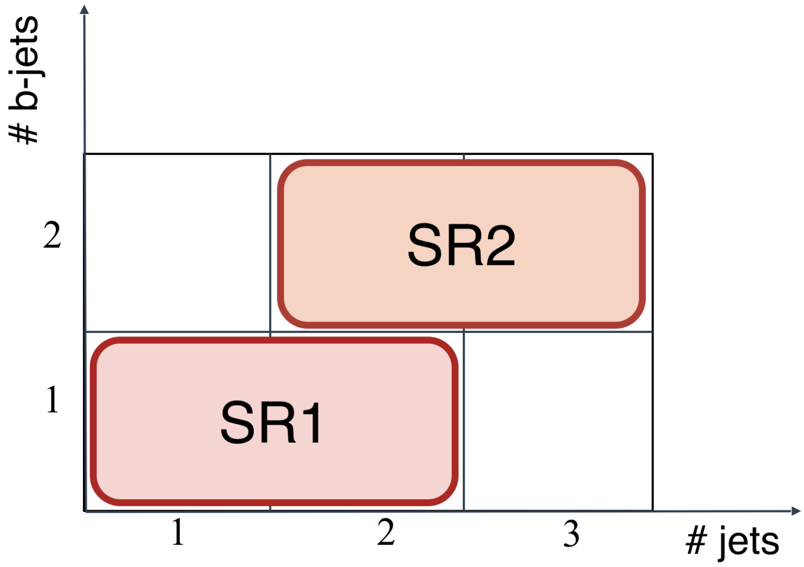
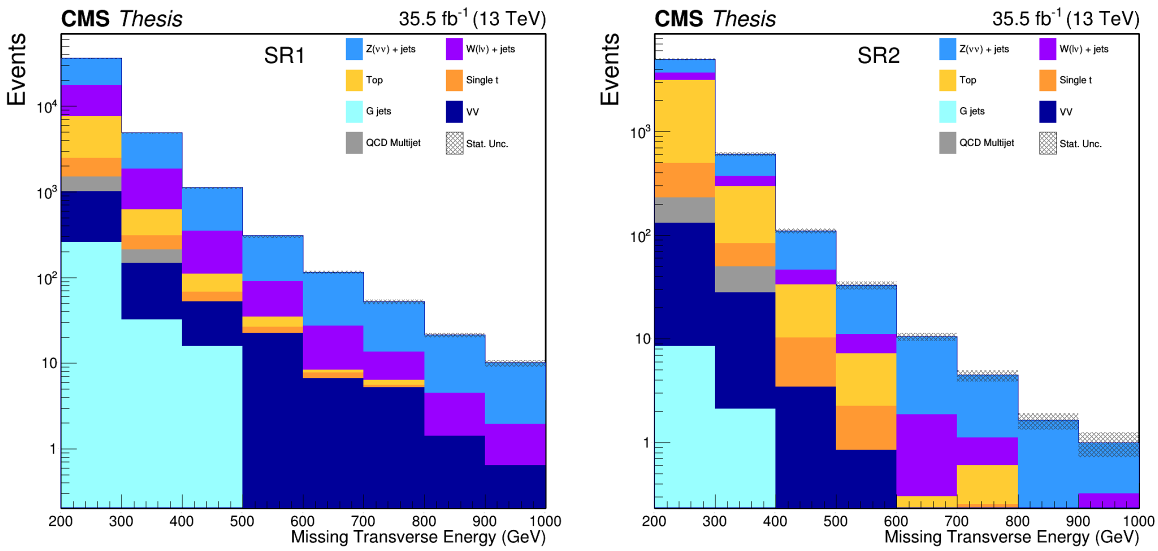
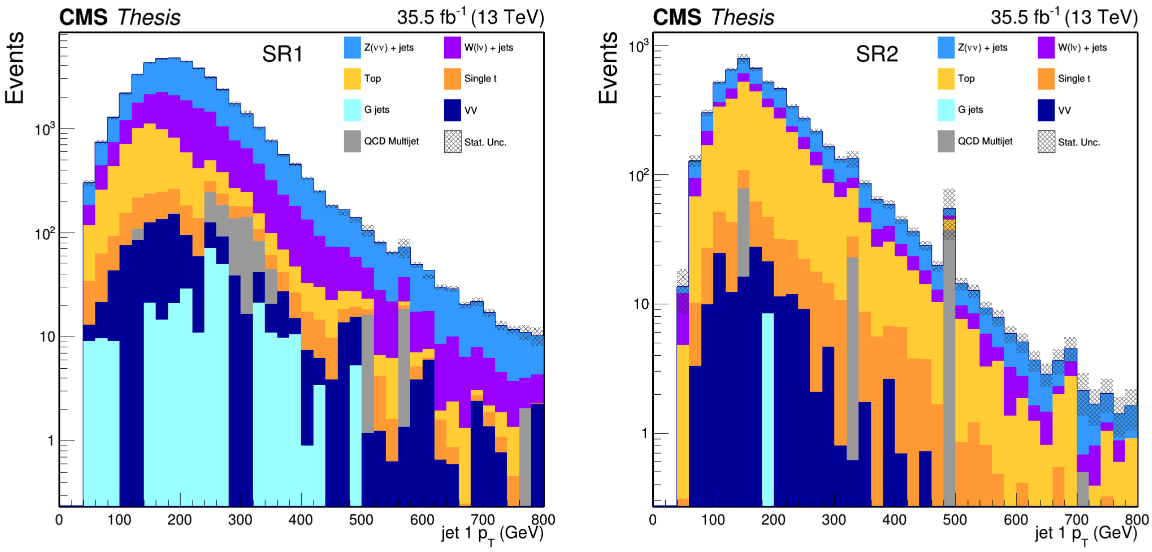


Figure 4: Figure showing the categories defined in the signal regions. The categories depend on the total number of jets and the number of b-tagged jets in an event.



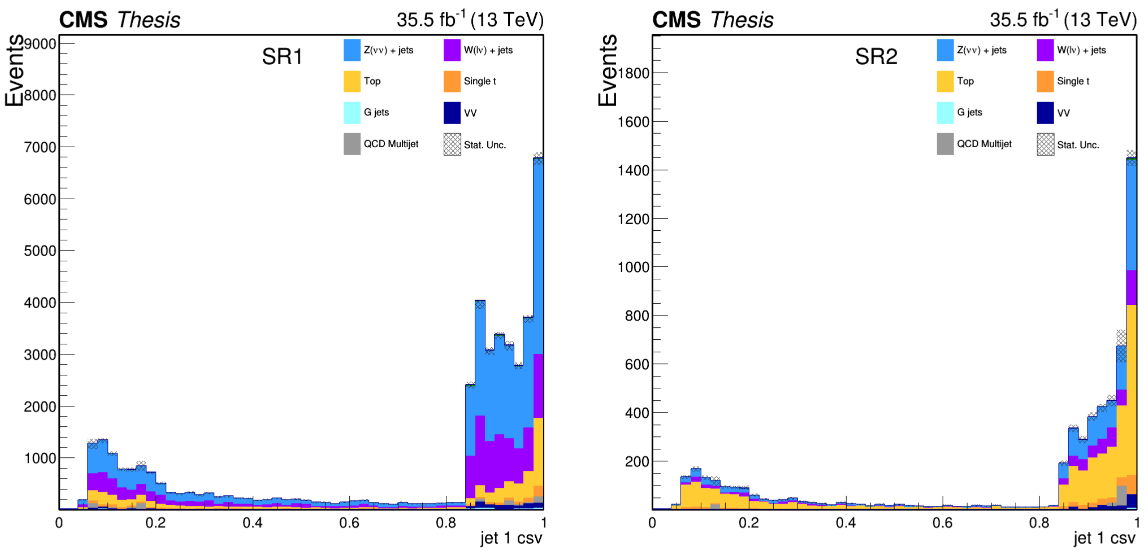
(a) (b)

Figure 5: Missing transverse energy distribution in the selected events in (a) 1 b-tagged category (SR1), and (b) 2 b-tagged category (SR2).



(a) (b)

Figure 6: Distribution of the transverse momentum of the leading jet in the selected events in (a) 1 b-tagged category (SR1), and (b) 2 b-tagged category (SR2).



(a) (b)

Figure 7: Distribution of the CSVv2 discriminator value of the leading jet in the selected events in (a) 1 b-tagged category (SR1), and (b) 2 b-tagged category (SR2).

	Category 1: SR1	**Category 2: SR2**
trigger	`HLT_PFMETNoMu90_PFMHTNoMu90_IDTight_v*` `HLT_PFMETNoMu120_PFMHTNoMu120_IDTight_v*`	
E_T^{miss}	> 200 GeV	
Jet 1	$p_T > 50$ GeV, $\Delta\phi(\text{jet}, E_T^{\text{miss}}) > 0.5$, charged hadron fraction > 0.1, neutral hadron fraction < 0.8	
Jet 2	(if present) $p_T > 30$ GeV, $\Delta\phi(\text{jet}, E_T^{\text{miss}}) > 0.5$	$p_T > 50$ GeV, $\Delta\phi(\text{jet}, E_T^{\text{miss}}) > 0.5$
Jet 3	veto	(if present) $p_T > 30$ GeV, $\Delta\phi(\text{jet}, E_T^{\text{miss}}) > 0.5$
b-tag	1 with CSV medium WP	2 with CSV medium WP
leptons	veto	

Table 2: Selection applied on the each of the two categories in the signal region, SR1 and SR2.

$p_T > 50$ GeV and same quality selections as that in SR1. The sub-leading jet is required to have a $p_T > 50$ GeV. The third jet, if any, is only required to satisfy the jet selection criteria.

In addition to the above, all jets in both the signal regions are required to have $\Delta\phi(\text{jet}, E_T^{\text{miss}}) > 0.5$ to further reduce QCD multijet background. The selections for each of the categories is summarized in table 2.

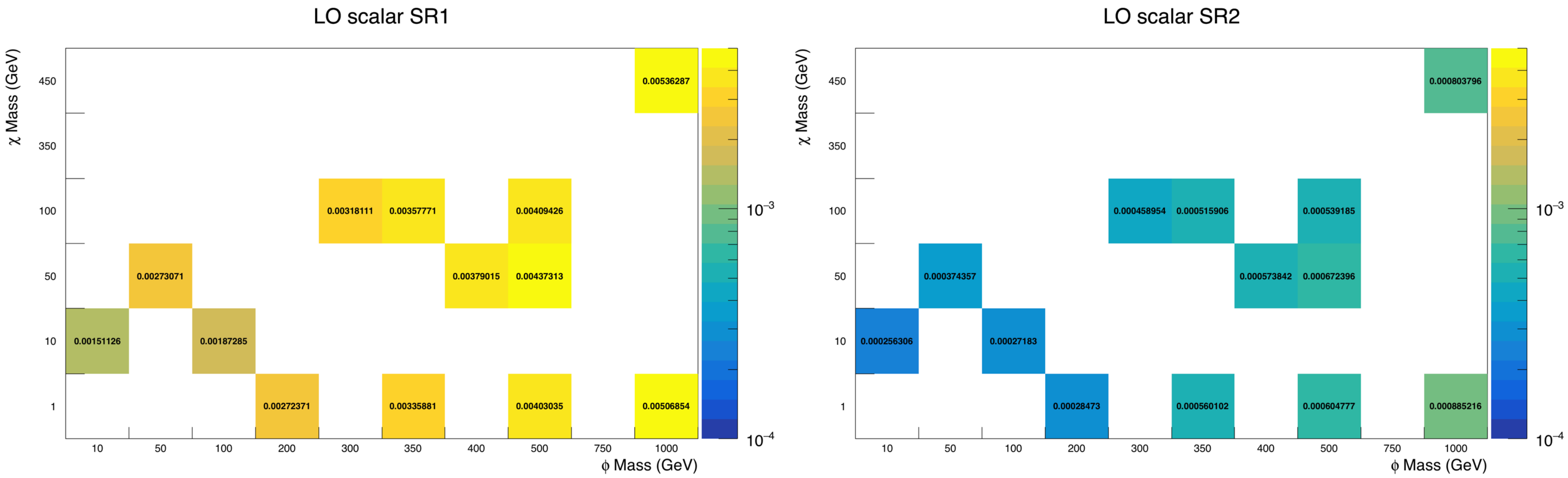
The E_T^{miss} (MET) distribution of the two SR categories are presented in figure 5. The distributions of the transverse momentum and CSVv2 discriminator value of the leading jet are shown in figures 6 and 7 respectively.

6.4 Signal Efficiencies

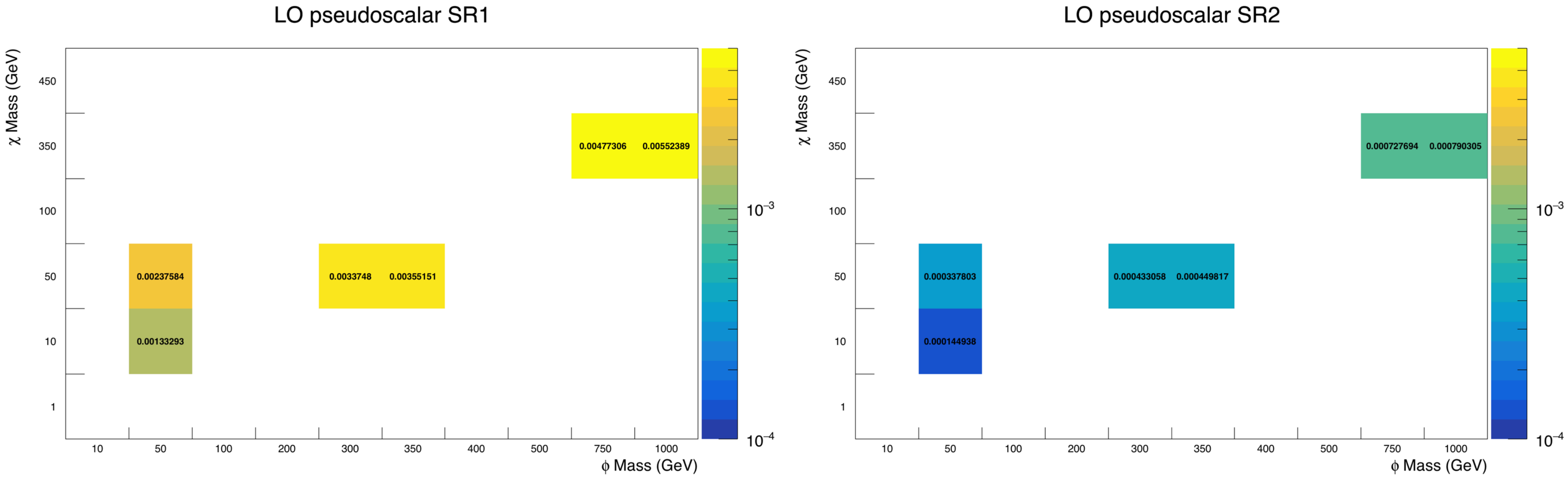
Signal efficiency is defined as the fraction of events of a signal sample passing the selection criteria. In this subsection, we look at the signal efficiency of various signal models across various mass points of the dark matter and the mediator, separately for scalar and pseudoscalar mediators, by applying the signal region selections to the signal samples. We also present a comparison of signal efficiency of bb+DM samples with that of tt+DM samples under the same selection cuts.

6.4.1 LO

The signal samples used in this section are accurate to the Leading Order (LO). The signal efficiencies across various mass points of the dark matter and the mediator,



(a)



(b)

Figure 8: Signal efficiencies of the Leading Order (LO) bb+DM model for various mass points of the dark matter and the mediator, for (a) scalar mediator, and (b) pseudoscalar mediator, in SR1 (left) and SR2 (right).

separately for scalar and pseudoscalar mediators, for each of the signal region categories, are plotted in 2D histograms in figure 8.

The efficiencies are comparable for the two mediators. A strong dependence of signal efficiency on the mass of the mediator (ϕ) is observed, and a weaker dependence is observed on the mass of the dark matter particle (χ). The efficiency is overall higher in SR1 than in SR2, which can be attributed to the stringent b-jet requirements in SR2.

6.4.2 NLO

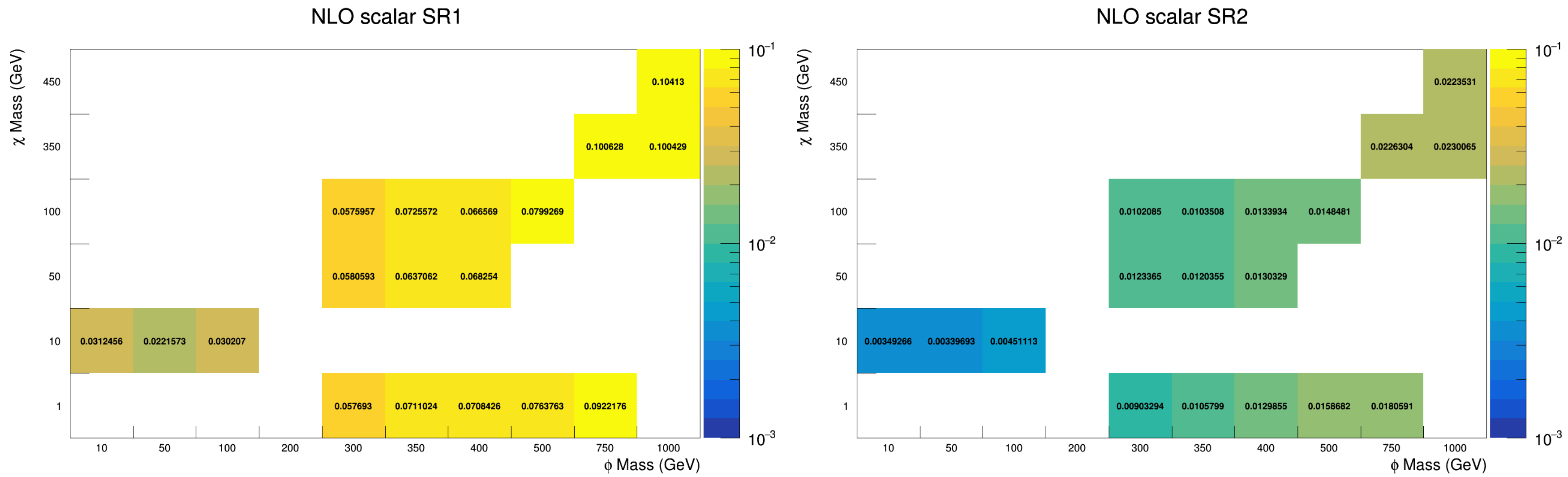
The signal samples used in this section are accurate to the Next-to-Leading Order (NLO). The signal efficiencies across various mass points of the dark matter and the mediator, separately for scalar and pseudoscalar mediators, for each of the signal region categories, are plotted in 2D histograms in figure 9.

Similar to the LO samples, efficiencies are comparable for scalar and pseudoscalar mediators, and the dependence on the mass of the mediator (ϕ) is strong. However, the overall signal efficiency is higher in the case of NLO samples compared to LO ones.

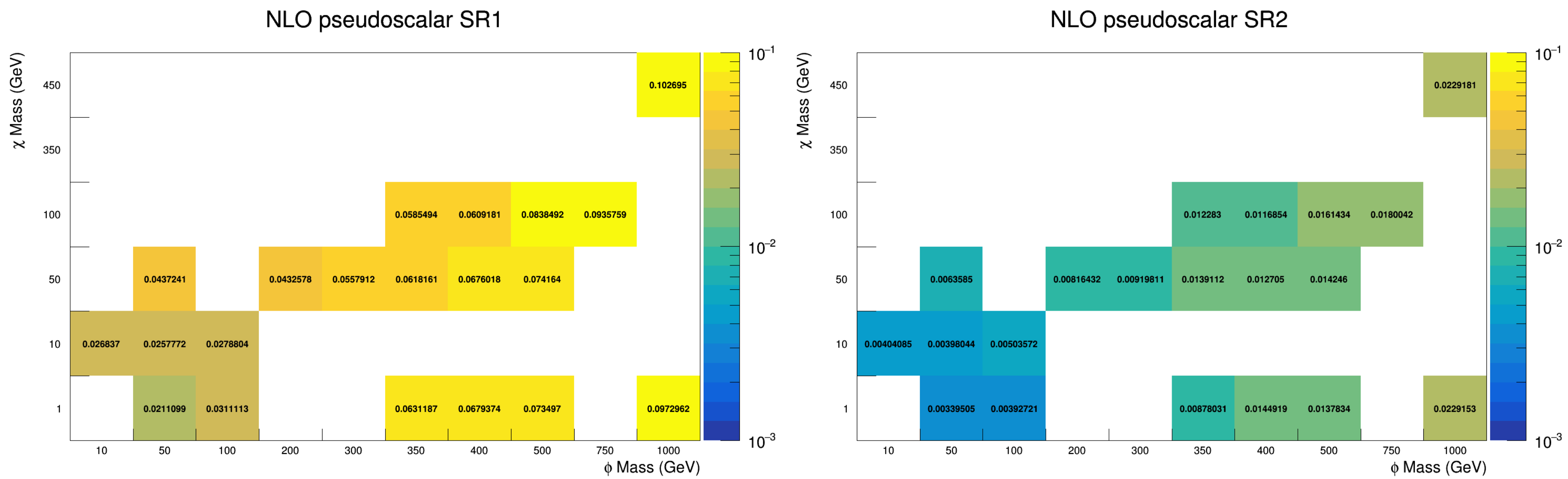
6.4.3 Comparison with ttDM

We present a comparison of the signal efficiencies of the bb+DM model with that of the tt+DM models since the final states are similar. The samples used are accurate to the Leading Order (LO). The signal efficiencies across various mass points of the dark matter and the mediator, separately for scalar and pseudoscalar mediators, for each of the signal region categories, are plotted in 2D histograms in figure 8.

We find similar dependence on the mediator and dark matter masses as in bb+DM samples, but the signal efficiencies are smaller for each mass point pair compared to bb+DM. This is because of the limit to the maximum number of jets we allow in the selection criteria in SR1 and SR2: the $t\bar{t}$ pair in tt+DM production would give rise to a larger number of jets in the final state compared to a $b\bar{b}$ pair, and hence a smaller fraction of these events are selected by our selection criteria.



(a)



(b)

Figure 9: Signal efficiencies of the Next-to-Leading Order (NLO) bb+DM model for various mass points of the dark matter and the mediator, for (a) scalar mediator, and (b) pseudoscalar mediator, in SR1 (left) and SR2 (right).

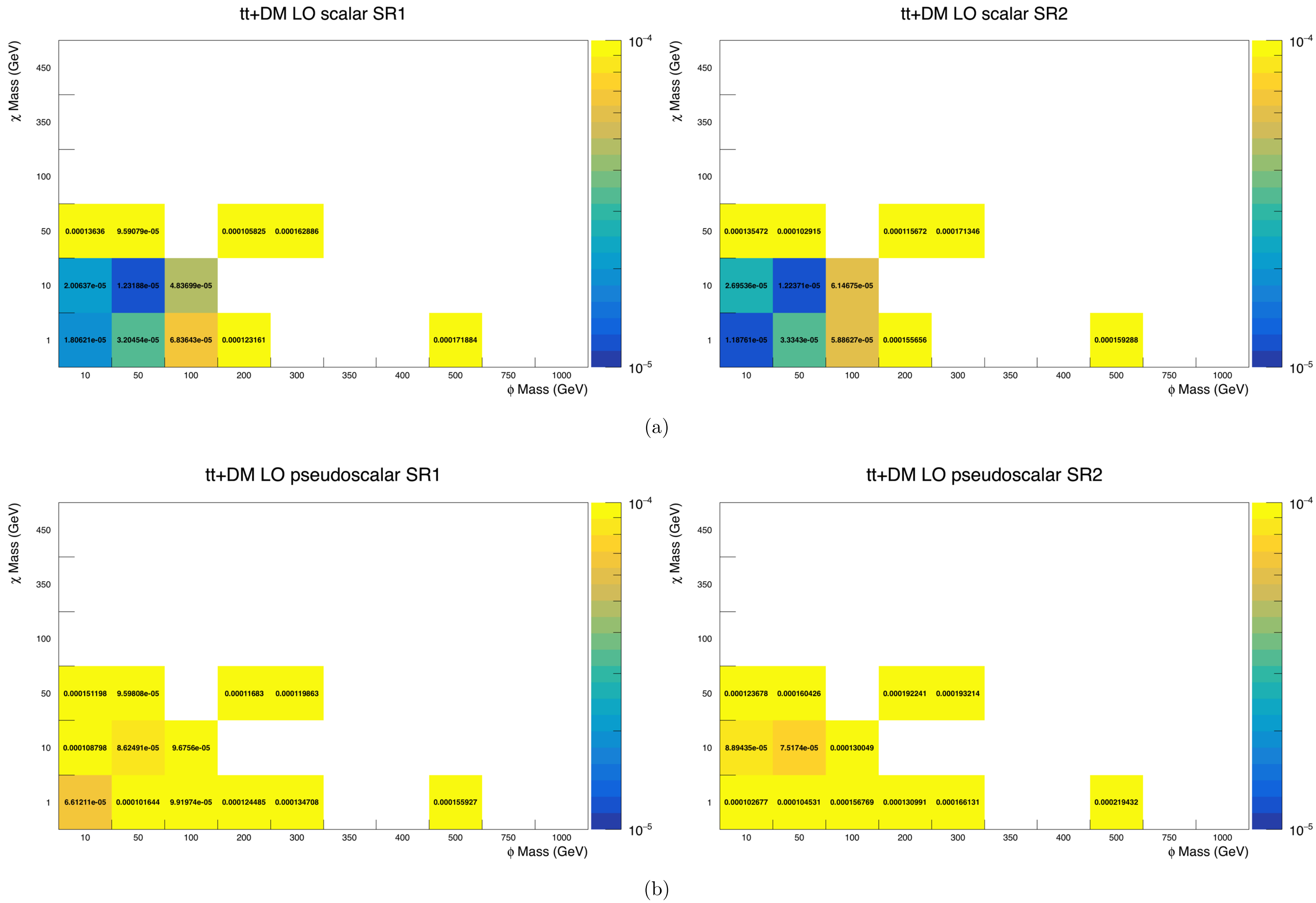


Figure 10: Signal efficiencies of the Leading Order (LO) tt+DM model for various mass points of the dark matter and the mediator, for (a) scalar mediator, and (b) pseudoscalar mediator, in SR1 (left) and SR2 (right).

7 Control Regions

We define several control regions to derive the normalization and determine the shapes of the dominant backgrounds in a data driven way. The major backgrounds, $Z \to \nu\nu$ + jets, $W \to l\nu$ + jets and $\mathrm{t\bar{t}}$, are modelled using kinematically similar processes with different and well-understood final states, by applying requirements identical to those in SR1 and SR2.

- $Z \to \nu\nu$ events are modelled using $Z \to ll$ events as they have similar kinematics. Moreover, reconstruction of the Z p_T spectrum and isolation of events using a cut on the Z mass is possible.
- $W \to l\nu$ events appear as backgrounds in the signal regions when the lepton is not detected. Such events are modelled using $W \to l\nu$ events where the lepton is well-measured.
- The $\mathrm{t\bar{t}}$ events where the leptons are lost also appear as backgrounds in the signal regions. These are modelled using $\mathrm{t\bar{t}}$ events where one electron and one muon are well-measured. The requirement of different flavours of leptons for this control region ensures mutual exclusivity of all the control regions.

The 10 mutually exclusive control regions are depicted in figure 11.

To model the quantity E_T^{miss} of the signal regions, we define the quantity "hadronic recoil" in the control regions. For each event in the control regions, we define hadronic recoil

$$\overrightarrow{\mathcal{U}} = -(\overrightarrow{E}_T^{\text{miss}} + \sum \overrightarrow{p}_T^{lep}), \tag{3}$$

where $\overrightarrow{E}_T^{\text{miss}}$ represents the actual missing energy in the control region event and $\sum \overrightarrow{p}_T^{lep}$ represents the sum of the transverse momenta of all the leptons in the event. The quantity, $\overrightarrow{\mathcal{U}}$, acts as a model for the value of $\overrightarrow{E}_T^{\text{miss}}$ in the corresponding modelled signal region, had the lepton(s) gone undetected.

Since the control regions need to be as close to the signal regions as possible, we require a cut of $\mathcal{U} > 200$ GeV in the control region events to emulate the $E_T^{\text{miss}} > 200$ GeV cut in the signal regions. Since the direction of the modelled E_T^{miss} is also different now, we look at a more inclusive selection in the control regions by relaxing the $\Delta\phi(\text{jet}, E_T^{\text{miss}}) > 0.5$ requirement for jets in the events selected in the control region.

In addition to these, we define a QCD multijet control region by having the same cuts as the SRs, but inverting the $\Delta\phi(\text{jet}, E_T^{\text{miss}})$ cut for at least one jet, i.e., by requiring $\min(\Delta\phi(\text{jet}, E_T^{\text{miss}})) < 0.5$. Since this region is expected to be dominated by QCD multijet events, this CR could be used to find a correct normalization for QCD events.

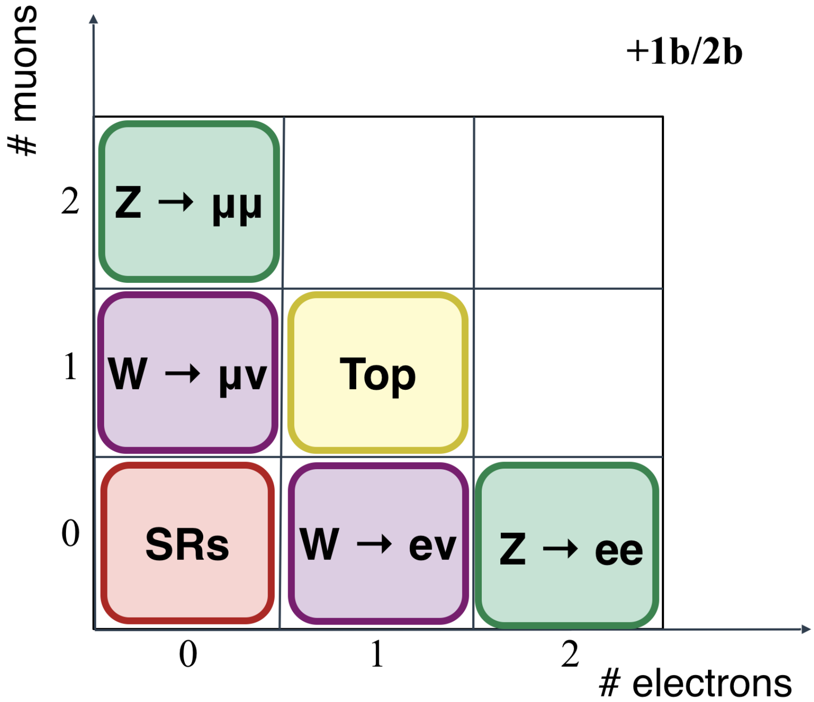


Figure 11: Figure showing the the control regions and the position of the signal regions in the electron-muon space. The categories depend on the number and flavour of leptons in an event. Each of the blocks represents two exclusive regions, corresponding to one and two b-tagged jets respectively.

To implement the control regions, the preselection step is modified to allow events containing one or two leptons that satisfy the physics object selections outlined in section 6.1.1. Such events satisfying $\mathcal{U} > 200$ GeV are kept for analysis, along with events with $E_T^{\text{miss}} > 200$ GeV.

7.1 Z Control Regions

The main background of the analysis is $Z \to \nu\nu$. This process is difficult to study and to normalize, since its only signature is a large missing energy. However, the other Z decay modes into leptons are useful as leptons can be easily triggered and measured, and a full reconstruction of the Z boson kinematics is possible. The decays into a pair of charged leptons and a pair of neutrinos of a Z boson are very similar from a theoretical point of view. Experimentally, one can then remove the charged leptons in $Z \to \mu\mu$ and $Z \to ee$ events in order get a proxy to $Z \to \nu\nu$ events. Since the neutrino would be detected as missing energy had the Z decayed to neutrinos, the E_T^{miss} of the $Z \to \nu\nu$ process can be estimated by the adding the lepton p_T to the E_T^{miss} of the $Z \to ll$ event according to equation (3).

7.1.1 1 b-tagged

The selection cuts for Z+1b control regions are summarized in table 3.

	Muon Region	**Electron Region**
Trigger	`HLT_IsoMu24_v*` `HLT_IsoTkMu24_v*`	`HLT_Ele27_WPLoose_Gsf` `HLT_Ele27_WPTight_Gsf`
Leptons	2 muons, opposite charges	2 electrons, opposite charges
Leading l	$p_T > 30$ GeV, tight ID, Iso < 0.15	$p_T > 30$ GeV, tight ID
Sub-leading l	$p_T > 10$ GeV, loose ID, Iso < 0.25	$p_T > 10$ GeV, loose ID
Z mass	70 GeV $< m_{ll} <$ 110 GeV	
Hadronic recoil	$>$ 200 GeV	
Jets	same as in SR1	
Other leptons	veto	

Table 3: Selection cuts for the Z+1b control regions, showing Z+1b muon region in the second column and Z+1b electron region in the third column.

Some distributions for events selected in the Z+1b control regions are shown in figures 12 and 13.

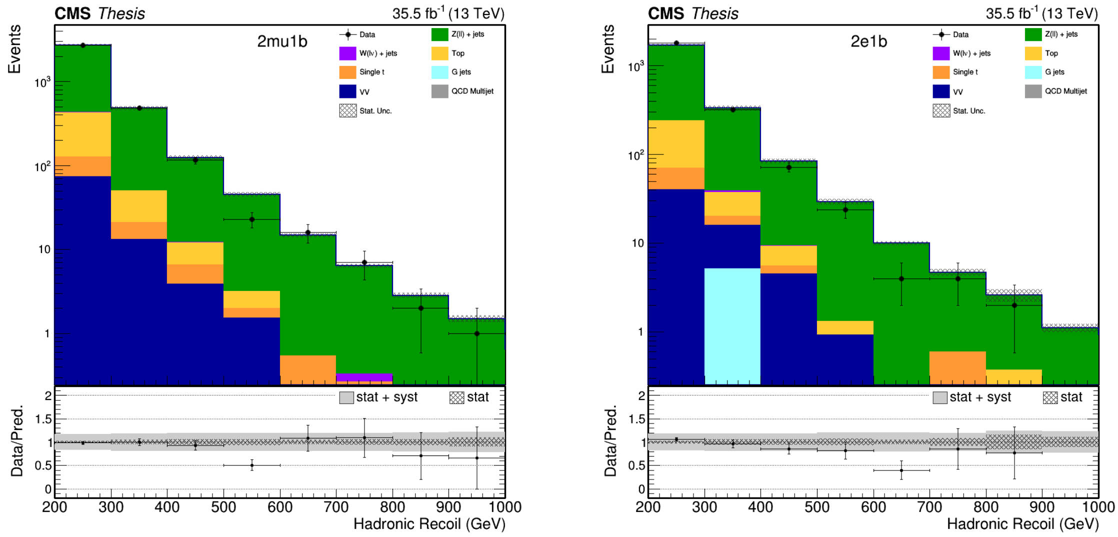
7.1.2 2 b-tagged

The selection cuts for Z+2b control regions are summarized in table 4.

	Muon Region	**Electron Region**
Trigger	`HLT_IsoMu24_v*` `HLT_IsoTkMu24_v*`	`HLT_Ele27_WPLoose_Gsf` `HLT_Ele27_WPTight_Gsf`
Leptons	2 muons, opposite charges	2 electrons, opposite charges
Leading l	$p_T > 30$ GeV, tight ID, Iso < 0.15	$p_T > 30$ GeV, tight ID
Sub-leading l	$p_T > 10$ GeV, loose ID, Iso < 0.25	$p_T > 10$ GeV, loose ID
Z mass	70 GeV $< m_{ll} <$ 110 GeV	
Hadronic recoil	$>$ 200 GeV	
Jets	same as in SR2	
Other leptons	veto	

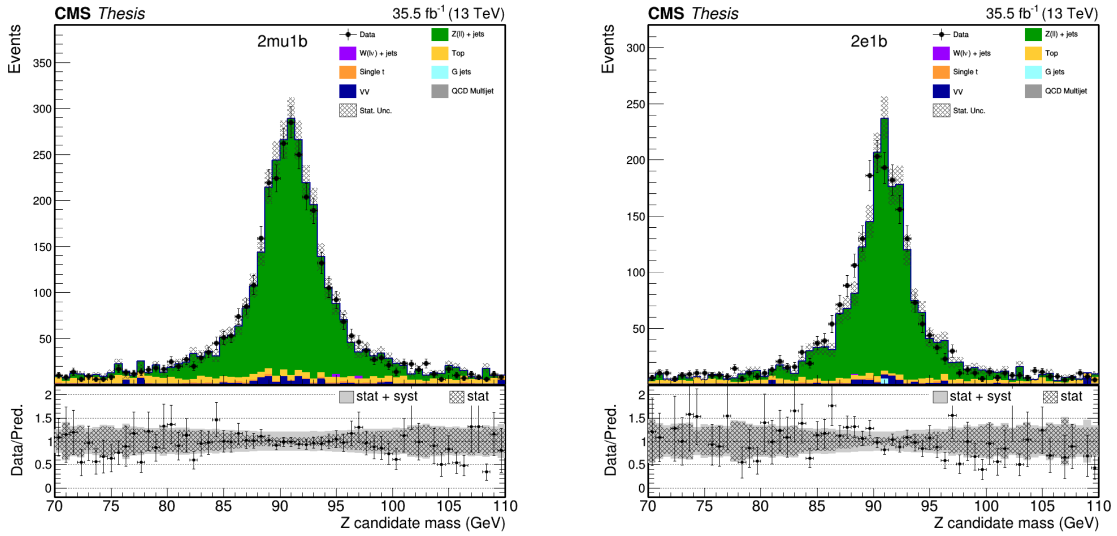
Table 4: Selection cuts for the Z+2b control regions, showing Z+2b muon region in the second column and Z+2b electron region in the third column.

Some distributions for events selected in the Z+2b control regions are shown in figures 14 and 15.



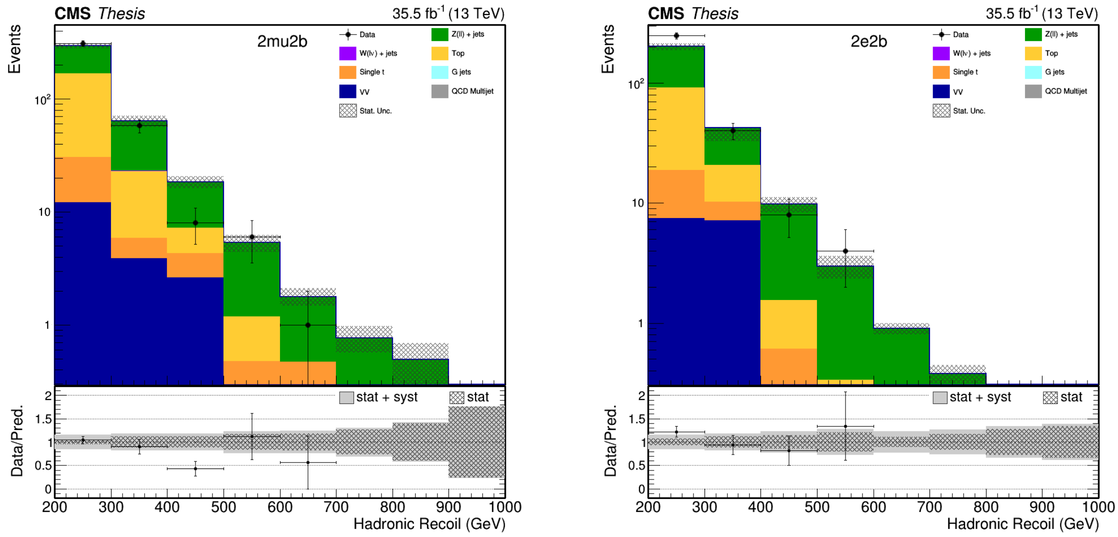
(a) (b)

Figure 12: Distribution of hadronic recoil in events selected in the (a) Z+1b muon and (b) Z+1b electron control regions.



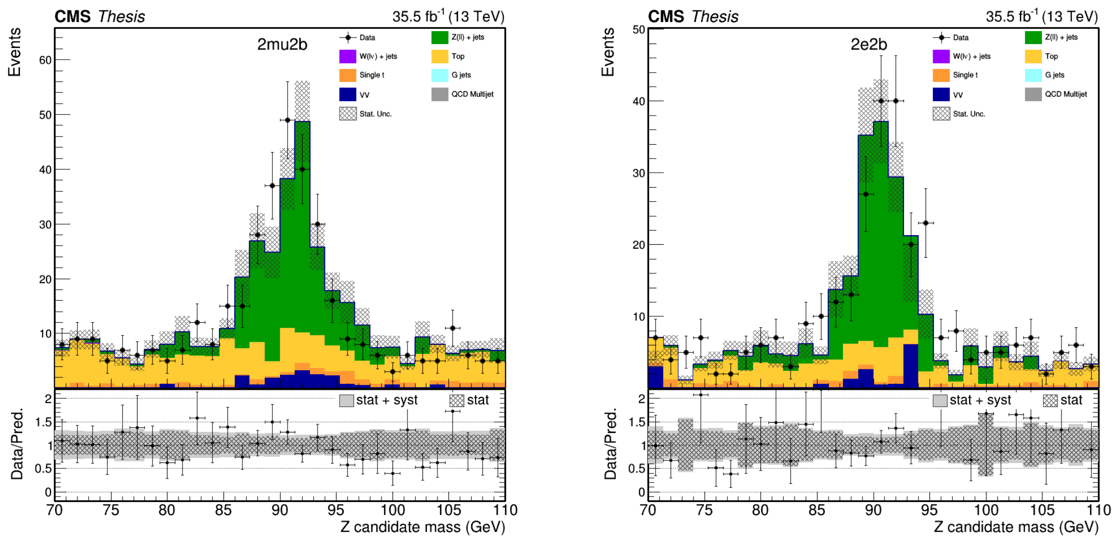
(a) (b)

Figure 13: Distribution of dilepton invariant mass in events selected in the (a) Z+1b muon and (b) Z+1b electron control regions.



(a) (b)

Figure 14: Distribution of hadronic recoil in events selected in the (a) Z+2b muon and (b) Z+2b electron control regions.



(a) (b)

Figure 15: Distribution of dilepton invariant mass in events selected in the (a) Z+2b muon and (b) Z+2b electron control regions.

7.2 W Control Regions

Events with $W \to l\nu$ + jets, where the lepton is not detected, form another major background in this analysis. These background events can be modelled using the same principle used in the Z control regions. We look into $W \to l\nu$ events where the l is well measured and add the evaluate the hadronic recoil to model the E_T^{miss} of $W \to l\nu$ events where the lepton is lost.

7.2.1 1 b-tagged

The selection cuts for W+1b control regions are summarized in table 5.

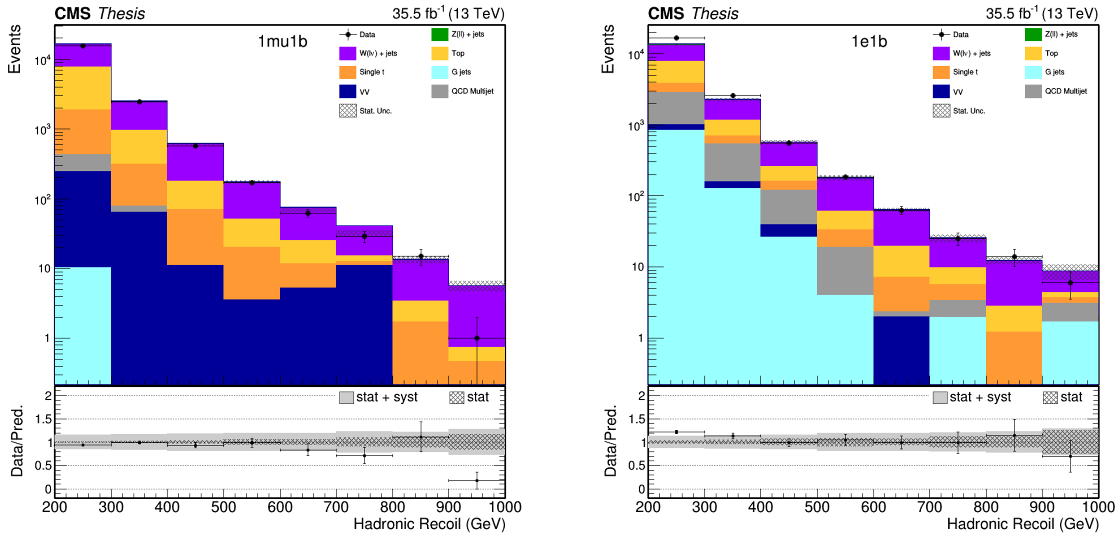
	Muon Region	**Electron Region**
Trigger	`HLT_IsoMu24_v*` `HLT_IsoTkMu24_v*`	`HLT_Ele27_WPLoose_Gsf` `HLT_Ele27_WPTight_Gsf`
Leptons	1 muon	1 electron
Leading l	$p_T > 30$ GeV, tight ID, Iso < 0.15	$p_T > 30$ GeV, tight ID
W transverse mass	50 GeV $< m_T^W <$ 160 GeV	
Hadronic recoil	$>$ 200 GeV	
Jets	same as in SR1	
Other leptons	veto	

Table 5: Selection cuts for the W+1b control regions, showing W+1b muon region in the second column and W+1b electron region in the third column.

Some distributions for events selected in the W+1b control regions are shown in figures 16 and 17.

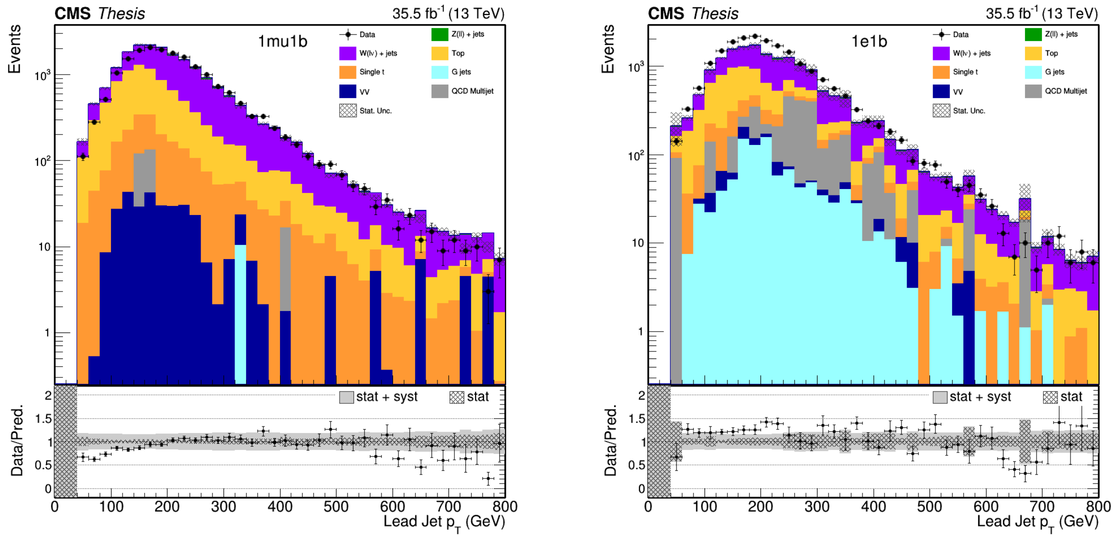
7.2.2 2 b-tagged

The selection cuts for W+1b control regions are summarized in table 6.



(a) (b)

Figure 16: Distribution of hadronic recoil in events selected in the (a) W+1b muon and (b) W+1b electron control regions.



(a) (b)

Figure 17: Distribution of transverse momentum of leading jet in events selected in the (a) W+1b muon and (b) W+1b electron control regions.

	Muon Region	**Electron Region**
Trigger	`HLT_IsoMu24_v*` `HLT_IsoTkMu24_v*`	`HLT_Ele27_WPLoose_Gsf` `HLT_Ele27_WPTight_Gsf`
Leptons	1 muon	1 electron
Leading l	$p_T > 30$ GeV, tight ID, Iso < 0.15	$p_T > 30$ GeV, tight ID
W transverse mass	50 GeV $< m_T^W <$ 160 GeV	
Hadronic recoil	> 200 GeV	
Jets	same as in SR2	
Other leptons	veto	

Table 6: Selection cuts for the W+2b control regions, showing W+2b muon region in the second column and W+2b electron region in the third column.

Some distributions for events selected in the W+1b control regions are shown in figures 18 and 19.

7.3 Top Control Regions

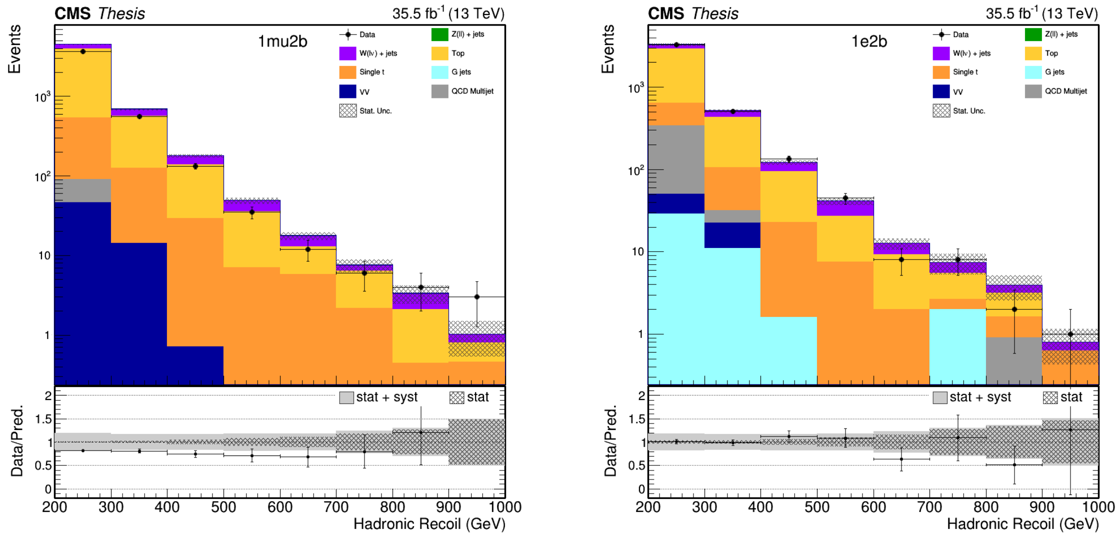
The $\mathrm{t\bar{t}}$ control region is modelled using the dileptonic top decays that are mutually exclusive to both Z and W control regions. The only final state satisfying this is $\mathrm{t\bar{t}} \rightarrow \mu\nu_\mu e\nu_e \mathrm{b\bar{b}}$. The hadronic recoil for these events are calculated in the same way as in Z and W CRs.

The selection cuts for Top control regions are summarized in table 7.

	Top+1b Region	**Top+2b Region**
Trigger	`HLT_IsoMu24_v*` `HLT_IsoTkMu24_v*`	
Leptons	1 muon, 1 electron, opposite charges	
Leading μ	$p_T > 30$ GeV, tight ID, Iso < 0.15 (if muon)	
Leading e	$p_T > 30$ GeV, tight ID	
Hadronic recoil	> 200 GeV	
Jets	same as in SR1	same as in SR2
Other leptons	veto	

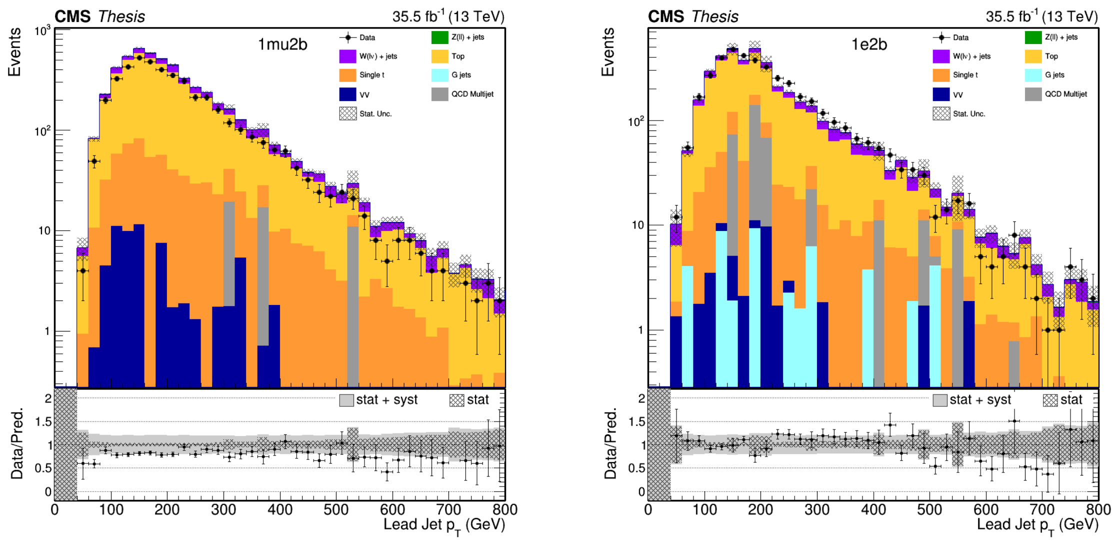
Table 7: Selection cuts for the Top control regions, showing Top+1b region in the second column and Top+2b region in the third column.

Some distributions for events selected in the Top control regions are shown in figures 20, 21 and 22.



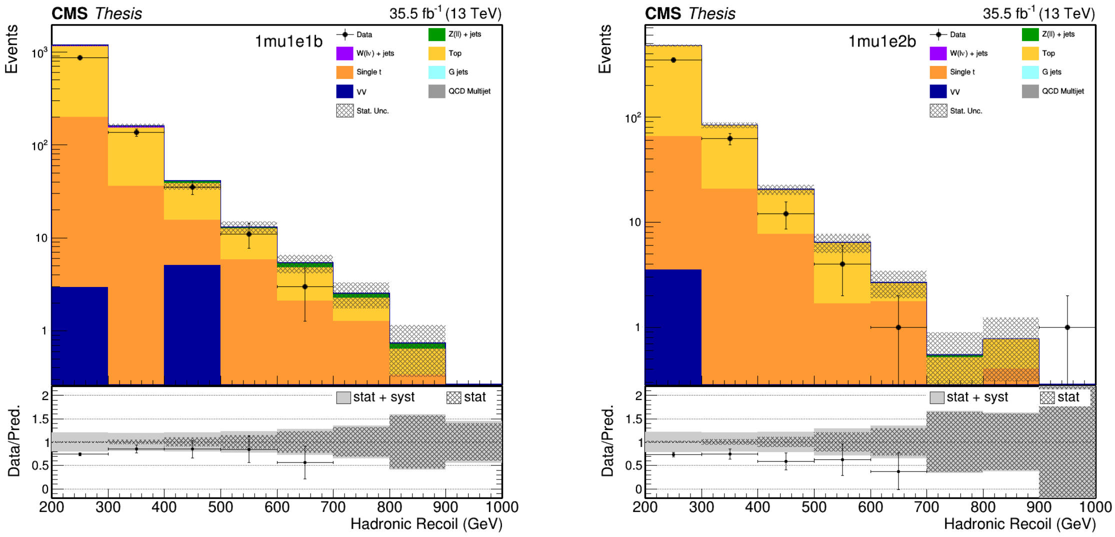
(a) (b)

Figure 18: Distribution of hadronic recoil in events selected in the (a) W+2b muon and (b) W+2b electron control regions.



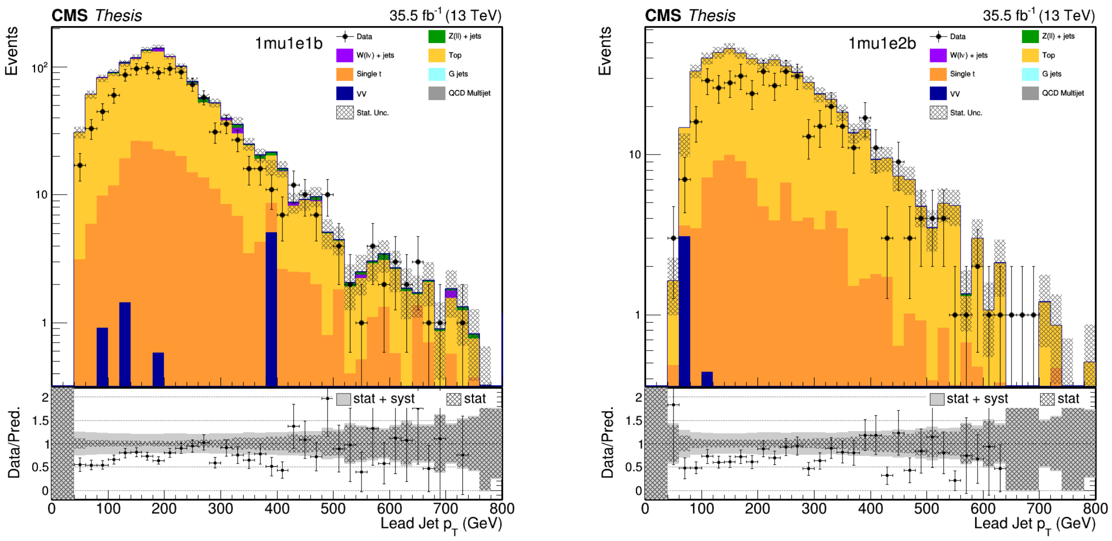
(a) (b)

Figure 19: Distribution of transverse momentum of leading jet in events selected in the (a) W+2b muon and (b) W+2b electron control regions.



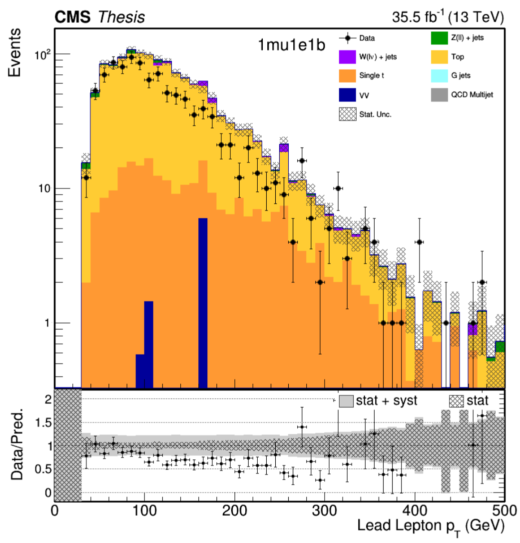
(a) (b)

Figure 20: Distribution of hadronic recoil in events selected in the (a) Top+1b and (b) Top+2b control regions.

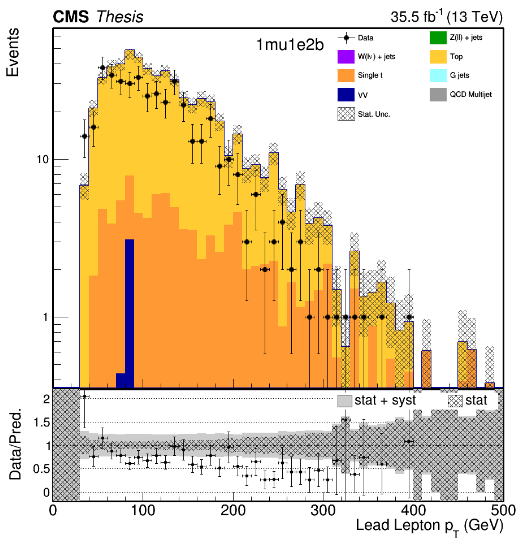


(a) (b)

Figure 21: Distribution of transverse momentum of leading jet in events selected in the (a) Top+1b and (b) Top+2b control regions.



(a)



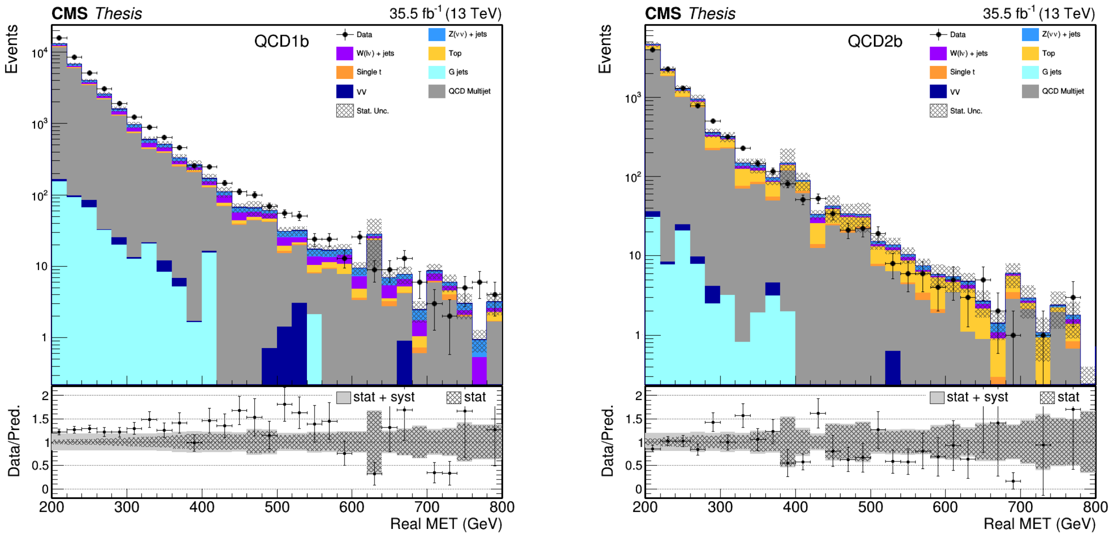
(b)

Figure 22: Distribution of transverse momentum of leading lepton in events selected in the (a) Top+1b and (b) Top+2b control regions.

7.4 QCD Multijet Control Regions

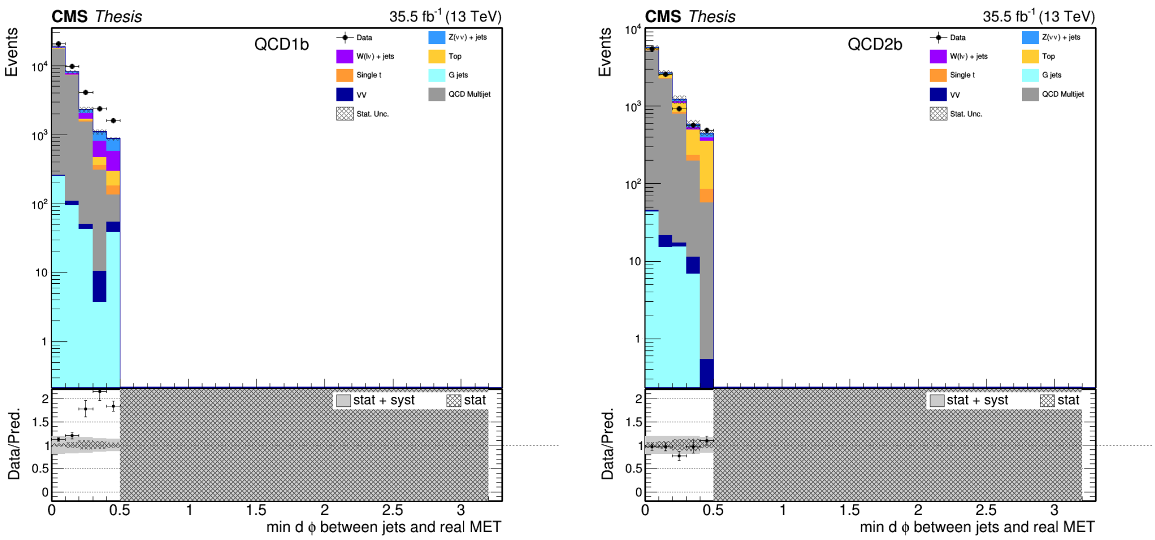
QCD multijet does not form a major background in our analysis, but since QCD simulation is challenging, we define another pair of control regions to derive the normalization for QCD multijet events.

The selection cuts are identical to those in the SRs, except for the $\Delta\phi(\text{jet}, E_T^{\text{miss}})$ cut. This cut is inverted for at least one jet, i.e., we require $\min(\Delta\phi(\text{jet}, E_T^{\text{miss}})) < 0.5$ to be satisfied for an event to enter the QCD control regions. Some distributions for events selected in the QCD+1b and QCD+2b are shown in figures 23 and 24.



(a) (b)

Figure 23: Distribution of E_T^{miss} in events selected in the (a) QCD+1b and (b) QCD+2b control regions.



(a) (b)

Figure 24: Distribution of $\min(\Delta\phi(\text{jet}, E_T^{\text{miss}}))$ in events selected in the (a) QCD+1b and (b) QCD+2b control regions.

8 Systematic Uncertainties

The background and signal predictions are affected by systematic uncertainties that have to be estimated and taken into account in the signal fit. This section includes the major systematic uncertainties for this analysis and how they are estimated.

8.1 b-tagging

b-tagging is known to have significant difference in performance for data and simulations. We apply b-tag scale factors to the simulated events to achieve better data-MC agreement. The uncertainty in the scale factors give rise to b-tagging uncertainty which is the major source of uncertainty in this analysis. To find the b-tagging systematic uncertainty, we repeat the b-tag correction procedure twice with the scale factors shifted up and down by one standard deviation.

8.2 MET trigger

Not all data events with an offline reconstructed MET $>$ 200 GeV are selected by the trigger used in the analysis, owing to limited trigger efficiency. To account for this fact and to get better data-MC agreement, the trigger efficiency is multiplied to the simulated events as a function of the reconstructed MET of the event. The measurement of trigger efficiency has some uncertainty and this method of MET trigger reweighting introduces this uncertainty to the analysis.

8.3 Lepton

The systematic uncertainties associated with lepton identification and isolation are evaluated by moving the scale factors up and down for identification and isolation by their errors as provided by respective POGs.

8.4 Luminosity

The uncertainty in the integrated luminosity across all runs of the data used in this analysis is measured to be about 2.5%. Hence, an additional uncertainty of 2.5% is introduced in the normalization of the MC samples.

8.5 Pile-up

An additional source of systematic error is the limited knowledge of the total inelastic cross section at 13 TeV, used to get the expected primary vertices distribution used for pile-up reweighting. An additional source of uncertainty arises from this.

8.6 Theory and MC simulation

PDF Systematic uncertainties coming from uncertainties in the Parton Distribution Function are an additional source of uncertainty. An uncertainty of 1% is considered for PDF in this analysis.

Electroweak reweighting Monte Carlo simulation of events with electroweak decays of Z and W bosons are known to have differences with those in data. To improve data-MC agreements for events with electroweak Z or W decay, we apply a scale factor that is precalculated as a function of the generator-level transverse momentum of the respective gauge boson. The uncertainties in this scale factor add to the systematic uncertainties.

Top reweighting Like corrections in events with electroweak decays, correction factors are applied to $\mathrm{t\bar{t}}$ events using a similar method. The uncertainty associated with this scale factor is another source of systematic uncertainties.

Heavy flavour Events with heavy flavour quarks are particularly difficult to simulate. An uncertainty of 3% has been assumed for simulations involving heavy flavour quarks.

8.7 Summary

The systematic uncertainties are summarized in table 8.

Nuisance Parameter	Uncertainty
Luminosity	2.5%
Pile-up	0.1%
b-tagging	5%
PDF	1%
Electron SFs	1%
Muon SFs	1%
Heavy Flavour	3%
MET trigger SF	5%

Table 8: Summary of systematic uncertainties considered in this analysis

9 Results

9.1 Signal Extraction Strategy

The data and MC yields obtained in the signal and control regions can be interpreted in a meaningful way by extracting the signal through a combined fit of the signal and control regions. The fit can be performed in three ways:

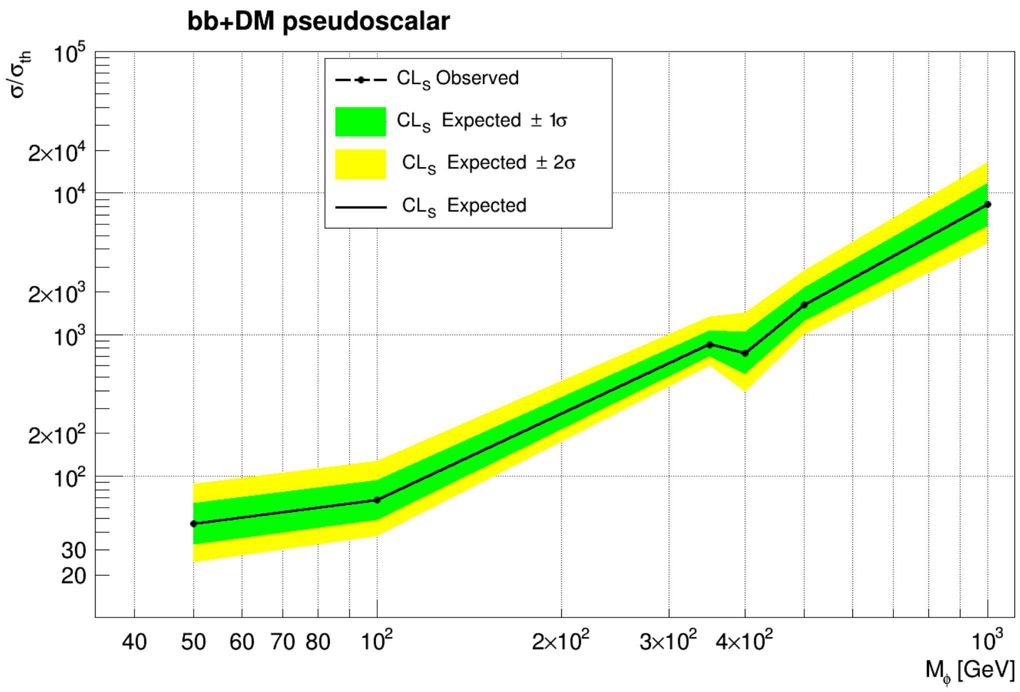
1. Cut-and-count: The fit is performed by simply considering the number of data and MC events in each signal and each control region after the cuts of the respective regions have been applied.
2. Cut-and-count++: Each of the regions is divided into a number of E_T^{miss} bins and the number of data and MC events in each bin is calculated. The fit is then performed simultaneously on all the bins. Since the E_T^{miss} bins are assumed to be uncorrelated, all hadronic recoil bins in the control regions must have enough events to obtain bin-by-bin background normalization values with sufficiently low statistical error for a reasonable constraint.
3. Shape: In addition to the previous method, one can take into account the correlation between bins while performing the fit. The normalization for backgrounds is calculated from the normalization in the control regions and is propagated to the higher E_T^{miss} bins with full bin-by-bin uncertainties.

In this analysis, we use the shape method to obtain limits. The `RooStats` [17] package and the CMS standard `combine` [18] tool are used to extract the limits on production cross sections with 1σ and 2σ bands. The limits are calculated for mass points with $M_\chi = 1$ GeV.

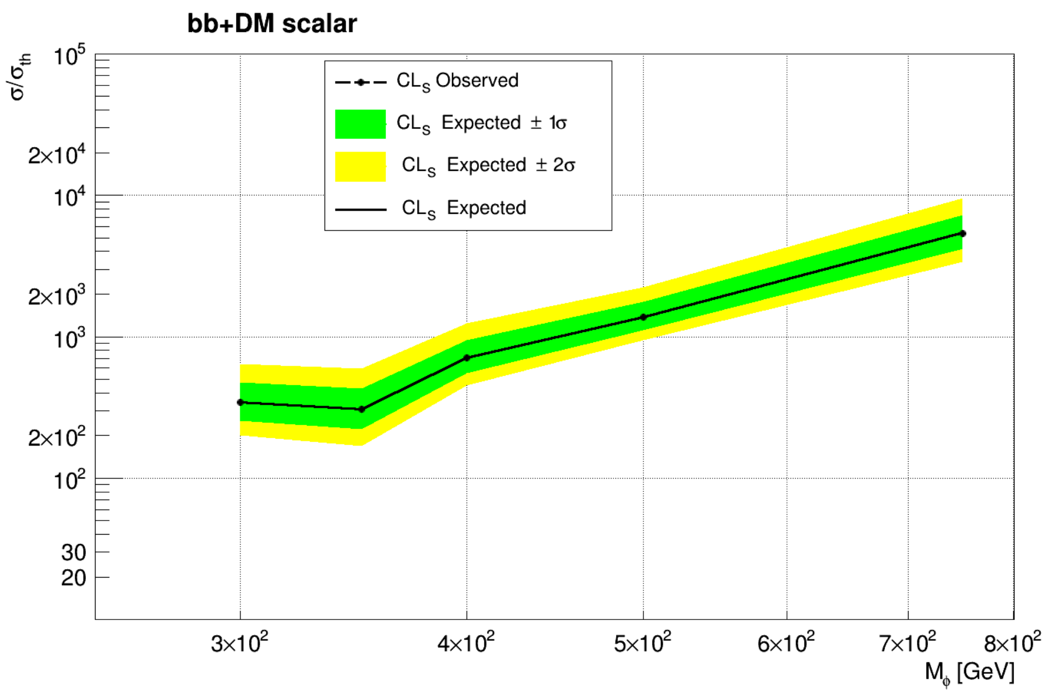
The limits are calculated using the Asimov method. This method does not take the number of events in data into consideration while performing the fit, and instead takes the sum of all the MC backgrounds in a given region as a proxy for the data in that region. As a result, the "expected" and "observed" values of cross section for a given mass point will be very close to each other. This method is adopted in this thesis as the data is still blinded in the current stage of the 2016 $b\bar{b}$+DM analysis.

9.2 Results and Limits

The calculated limits on production cross sections with $\pm1\sigma$ and $\pm2\sigma$ bands are shown in tables 9 and 10. The plots of the ratios of these values to the theoretically predicted cross sections are shown in figure 25.



(a)



(b)

Figure 25: Plots showing the ratio of the observed and expected limits to the theoretically predicted cross sections for bb+DM signal with $\pm 1\sigma$ and $\pm 2\sigma$ bands, for (a) pseudoscalar mediator, and (b) scalar mediator, for various mass points of the mediator. The mass of the dark matter particle is fixed at $M_\chi = 1$ GeV.

M_ϕ	50	100	350	400	500	1000
M_χ	1	1	1	1	1	1
Expected $-\ 2\sigma$	656.6792	608.4397	858.8628	221.4959	1229.447	63.806
Expected $-\ 1\sigma$	884.9971	808.0582	943.3203	296.4847	1267.2133	84.7397
Expected	1259.25	1124.625	1093.875	415.4062	1339.3125	117.9375
Expected $+\ 1\sigma$	1811.3867	1595.3193	1333.6877	589.2681	1472.7764	166.3582
Expected $+\ 2\sigma$	2512.8904	2186.8567	1652.8193	807.7661	1660.6581	228.6965
Observed	1260.7112	1125.8977	979.7634	414.8169	958.3258	117.9521

Table 9: Observed and expected limits for bb+DM signal with pseudoscalar mediator. All masses are in GeVs, and the limits are on cross sections in femtobarns (fb).

M_ϕ	300	350	400	500	750
M_χ	1	1	1	1	1
Expected $-\ 2\sigma$	466.5168	288.7405	424.7402	400.5104	184.3597
Expected $-\ 1\sigma$	598.9396	381.0498	513.6522	449.9994	224.9676
Expected	812.4375	529.875	657	538.2188	290.4375
Expected $+\ 1\sigma$	1126.5634	747.4218	869.1252	673.3766	386.5263
Expected $+\ 2\sigma$	1520.2959	1020.7495	1138.077	850.2266	508.7492
Observed	811.3579	529.3864	645.0002	504.9385	287.9466

Table 10: Observed and expected limits for bb+DM signal with scalar mediator. All masses are in GeVs, and the limits are on cross sections in femtobarns (fb).

9.3 Conclusions

The expected and observed cross section limits are very close and this is an artifact of using the Asimov method in the limit calculation, as data is blinded when Asimov limits are calculated. The observed values may change after the data is unblinded in this analysis.

However, we find that $\sigma/\sigma_{\text{th}}$ values along with their 2σ bands are greater than 1 for all mass points under consideration. This tells us that the analysis in this channel with current statistics is not sensitive enough to directly exclude any of the probed mass points in the simplified DM model parameter space.

10 Other activity with the CMS Collaboration: cMVAv2 Retraining for Phase-2

10.1 Introduction

b-tagging at the CMS experiment refers to determining whether a jet, as observed in the final state, originated from a bottom quark. Since the bottom quark is an important product for many known processes arising from proton-proton collisions at the LHC, it is crucial to have accurate b-tagging algorithms during event reconstruction and data analysis.

The current running of the CMS detector is referred to as the Phase-1 of the experiment. The detector is supposed to undergo a few major upgrades over the next few years and begin data collection in 2025 for Phase-2.

The combined MultiVariate Analysis version 2 (cMVAv2) is a commonly used b-tagging algorithm at CMS. In this section, results from retraining the cMVAv2 b-tagger for CMS for Phase-2 are presented. The training is performed using Monte Carlo event samples generated with the Phase-2 CMS detector simulation. Hence, a comparison of the performance of the retrained cMVAv2 with that of the existing training for Phase 0/1 is presented.

Two new variants of cMVAv2 that make use of DeepCSV, another existing b-tagging algorithm, are proposed. We show that the new cMVAv2 variants have a better performance than other b-tagging algorithms.

10.2 Experimental Background

10.2.1 b-tagging Algorithms used in CMS

There are a number of common b-tagging algorithms used in CMS analyses. They are:

JP and JBP: These make use of maximum likelihood methods on jet track information to output a discriminator value.

CSVv2: CSVv2 implements Neural Network based machine learning on secondary vertex and track information. Based on the algorithm used to reconstruct the secondary vertex, CSVv2 has two outputs: CSVAVR (Combined Secondary Vertex: Adaptive Vertex Reconstruction) and CSVIVF (Combined Secondary Vertex: Inclusive Vertex Finder).

DeepCSV: This algorithm uses Deep Neural Network based Deep Learning methods and makes use of secondary vertex and track information.

SoftMuon and SoftElectron: These make use of kinematic properties of particle flow muons and electrons whenever available inside the jet.

cMVAv2: This makes use of Boosted Decision Tree based machine learning methods to combine outputs of JP, JBP, CSVAVR, CSVIVF, SoftMu and SoftEl to give a super-combined discriminator output.

The aim of this project is to retrain the cMVAv2 tagger using CMS Phase-2 Monte Carlo samples, and test its performance and possible upgrades.

10.2.2 CMS Phase-2 Upgrade

The Phase-2 upgrade for CMS is planned to start operating from 2025. A number of detectors in CMS will be upgraded from Phase-0 and Phase-1 to cover more area and make more precise measurements. A few major upgrades are:

- More accurate tracking: 4 tracker layers instead of 3 in Phase-0/1.
- Greater tracker coverage: $|\eta| < 3.5$ instead of $|\eta| < 2.5$ in Phase-1 and $|\eta| < 2.4$ in Phase-0.
- Ability to collect data from proton-proton collisions at higher luminosities.

Consequently, all the algorithms need to be retrained with relevant information from extended detector configuration. In case of cMVAv2, the Phase-0 training (2015) was seen to be optimum for Phase-1 as well, and hence no retraining was performed for Phase-1. However, with more changes in detector configurations for Phase-2 and availability of other Phase-2 retrained b-taggers, it is expected that a cMVAv2 algorithm trained newly for Phase-2 would have a better performance in Phase-2 compared to the Phase-0 trained cMVAv2 that is in use now.

10.3 Monte Carlo Simulation of Collider Events

To design and test a tagger for CMS Phase-2, it is necessary to have a simulation of such Phase-2 events with all the raw properties that can be used as inputs to the training and testing algorithms. For this purpose, several Monte Carlo samples simulating Standard Model processes are used in this project. These samples were generated by simulating the CMS detector with all the upgrades planned for 2023. The sample used for training is:

- 50% of ttbar 91X PU200 (~350,000 events, ~2,000,000 jets):
`/TT_TuneCUETP8M1_14TeV-powheg-pythia8/`
`PhaseIITDRSpring17MiniAOD- PU200_91X_upgrade2023_realistic_`
`v3-v1/MINIAODSIM`

The samples used for validating the training are:

- (Other) 50% of ttbar 91X PU200 (~350,000 events, ~2,000,000 jets):
`/TT_TuneCUETP8M1_14TeV-powheg-pythia8/`
`PhaseIITDRSpring17MiniAOD- PU200_91X_upgrade2023_realistic_`
`v3-v1/MINIAODSIM`
- QCD 91X PU200 (~950k events, ~2,600,000 jets):
`/QCD_Flat_Pt-15to7000_TuneCUETP8M1_14TeV_pythia8/`
`PhaseIITDRSpring17MiniAOD-PU200_91X_upgrade2023_realistic_`
`v3-v1/MINIAODSIM`

10.4 cMVAv2 Retraining and Performance

10.4.1 Training using XGBoost

The retraining of cMVAv2 for CMS Phase-2 is performed using an open-source boosted decision tree based framework named XGBoost (eXtreme Gradient Boosting) [21] in a Python 2.7 environment, using jets from Phase-2 $t\bar{t}$ MC samples as inputs. The XGBoost hyperparameters used in the training are:

- learning_rate = 0.2
- max_depth = 3
- Number of boosting iterations = 500

XGBoost is used as a binary classifier. For each jet, the network is fed with the jet's JP, JBP, CSVAVR, CSVIVF, SoftMu and SoftEl discriminators and the true value of the jet's flavour (1 for a simulated b-jet, 0 for a udsg/c jet) as inputs. The network is expected to learn based on these inputs and then should be able to output a discriminator value between 0 and 1 for any input jet used in the subsequent validation process.

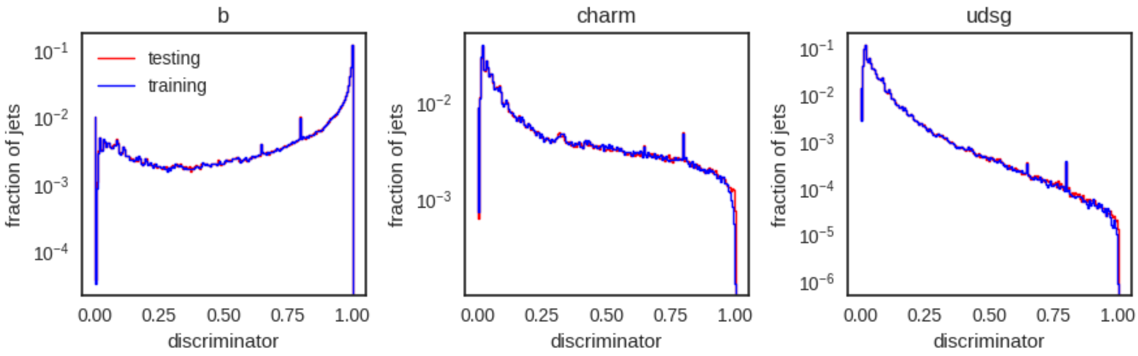


Figure 26: Plots showing the Phase-2 retrained cMVAv2 discriminator outputs for different flavour categories. As expected, a majority of jets have values closer to 1 for b-jets, and closer to 0 for c and udsg jets. Moreover, we see a good agreement between training (blue) and testing (red) data sets.

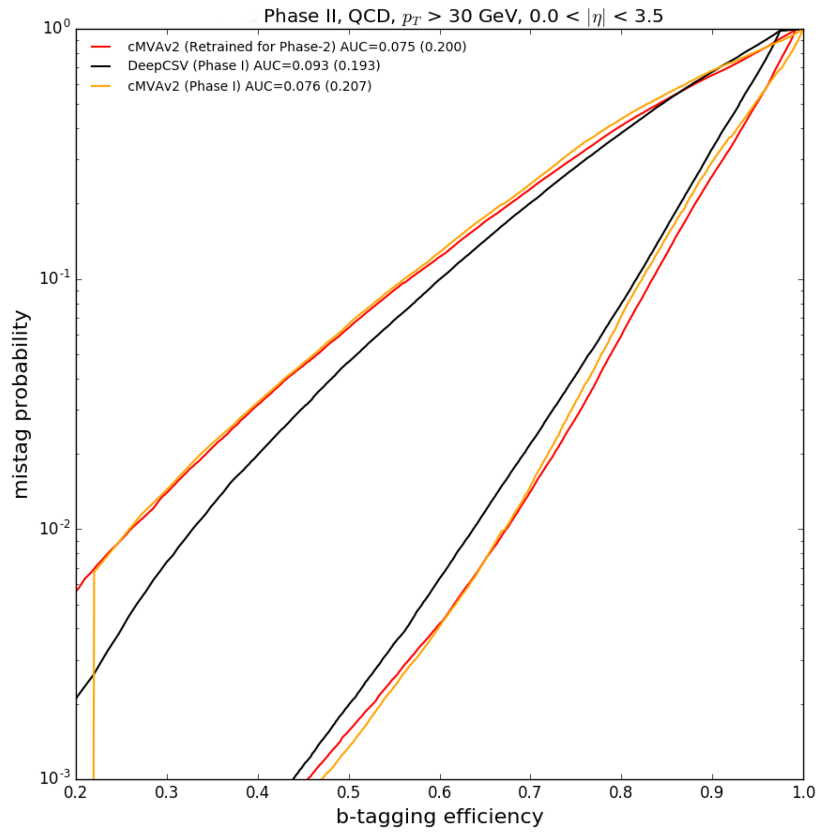


Figure 27: ROC curves for Phase-0/1 cMVAv2 (orange), Phase-2 retrained cMVAv2 (red) and Phase-1 DeepCSV (black). Small improvement is seen for Phase-2 retrained cMVAv2 over that in Phase 0/1. However, DeepCSV performs best in b vs. c branch, while cMVAv2 performs best in b vs. udsg branch.

10.4.2 Validation of b-tagging performance

After training, the performance of the newly trained network is validated on an independent, large set of jets. The newly-trained network's performance on the independent QCD sample is shown in the figures 26 and 27.

These ROC (Receiver Operating Characteristic) curves show mistag probabilities as functions of b-tagging efficiencies. The upper-left branch denotes the charm-mistag rate (b vs. c), while the lower-right branch denotes udsg-mistag rate (b vs. udsg). For each branch, a shift of the curve towards bottom-right (higher b-tagging efficiency with lower mistag rate) signifies an improvement in the b-tagging algorithm.

10.5 New cMVAv2 Variants for Phase-2

From figure 27, we see that DeepCSV performs better than cMVAv2 in the b vs. c branch, while cMVAv2 performs better in b vs. udsg branch. The natural attempt at this point would be to design an algorithm to bring out the best of cMVAv2 and DeepCSV in both branches.

10.5.1 Integrating DeepCSV within cMVAv2

One of cMVAv2's inputs, CSVv2 (CSVAVR and CSVIVF), is a shallow neural network based algorithm that makes use of track and secondary vertex information to generate a b-tagging discriminator output. On the other hand, a newer algorithm, DeepCSV, uses similar information but implements deep neural networks for the training. As a result, DeepCSV has a significantly better performance than CSVv2. So, it is expected that usage of DeepCSV within cMVAv2, replacing or in addition to CSVv2, would improve the performance of cMVAv2.

Hence, we propose two new variants of cMVAv2, as explained in table 11. The first new variant (D) replaces CSVv2 with DeepCSV and the second new variant (C+D) makes use DeepCSV besides CSVv2.

The training settings for the two new variants are kept the same within XGBoost, the only difference being the set of input variables being used.

10.6 Performance of new variants of cMVAv2

The performance of the new variants of cMVAv2 and their comparison with that of the classical cMVAv2 and DeepCSV for Phase-2 QCD samples are shown in figure 28.

Input CSV Name	Variables used (cMVAv2 trained with...)	Comments
CSVv2 (C)	Phase-2 CSVAVR, Phase-2 CSVIVF, Phase-2 JP, JBP, SoftMu, SoftEl	Phase 0/1 combination
DeepCSV (D)	DeepCSV, Phase-2 JP, JBP, SoftMu, SoftEl	New experimental combinations for Phase-2
CSVv2 and DeepCSV (C+D)	Phase-2 CSVAVR, Phase-2 CSVIVF, DeepCSV, Phase-2 JP, JBP, SoftMu, SoftEl	

Table 11: Table showing three different variants of cMVAv2 that we propose for Phase-2. The first column shows the CSV algorithm used within the corresponding variant of cMVAv2. The "C" row shows the classical variant of cMVAv2 that is in use in Phase 0/1. The "D" and "C+D" rows describe the new variants that we test in this project.

- **b vs. c branch:** As already seen in figure 27, DeepCSV performs better than classical cMVAv2. But from figure 28, we find that the two new variants of cMVAv2 perform at least as good as DeepCSV in this branch.
- **b vs. udsg branch:** The classical cMVAv2 performs better than DeepCSV in this branch, but the two new variants of cMVAv2 gain even more in performance and perform better than the classical cMVAv2.

So, the two new variants of cMVAv2 have the best performance in both b vs. c and b vs. udsg branches, the "C+D" variant having a marginally better performance than the "D" variant.

10.7 Other studies with Phase-2 cMVAv2 variants

10.7.1 Correlations among the CSV algorithms

Since we are using 3 different CSV algorithms (CSVAVR, CSVIVF, DeepCSV – all make use of the same raw information) in the same training, it is interesting to find out the correlation between them. The normalized correlation matrix between the three variables across jets is found to be:

$$\begin{array}{r} \\ \text{CSVAVR} \\ \text{CSVIVF} \\ \text{DeepCSV} \end{array} \begin{array}{c} \begin{array}{ccc} \text{CSVAVR} & \text{CSVIVF} & \text{DeepCSV} \end{array} \\ \begin{pmatrix} 0.32110 & 0.32176 & 0.31746 \\ 0.32176 & 0.32321 & 0.31854 \\ 0.31746 & 0.31854 & 0.35569 \end{pmatrix} \end{array}$$

The correlation plots in figure 29 show a high degree of correlation between the three CSV algorithms.

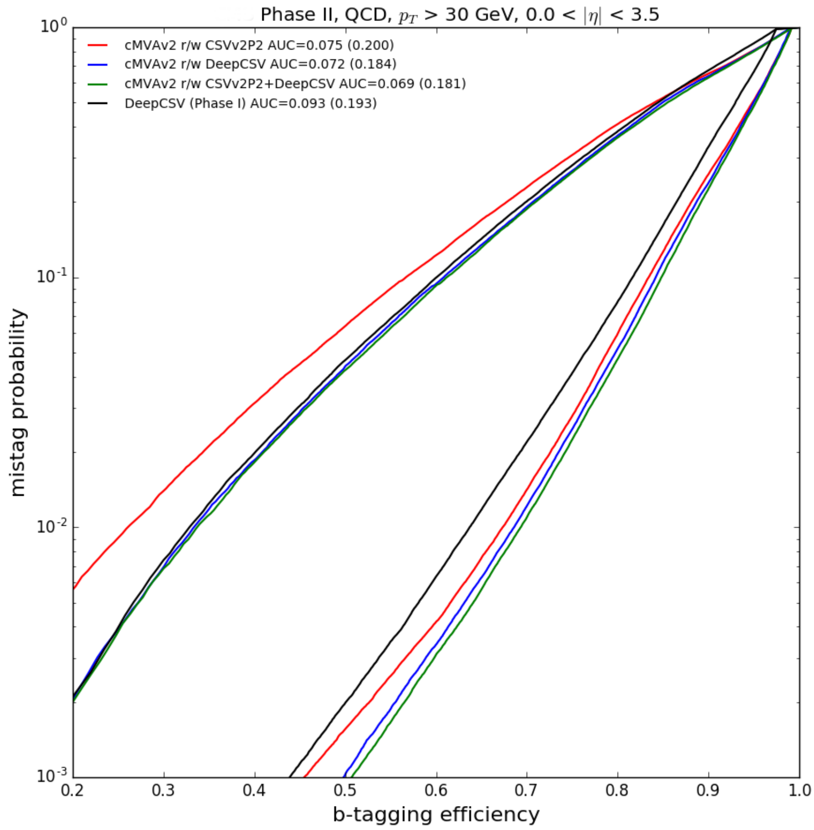


Figure 28: ROC curves for the Phase-2 retrained classical ("C" variant) cMVAv2 (red), "D" variant of Phase-2 cMVAv2 (blue), "C+D" variant of Phase-2 cMVAv2 (green), and Phase-1 DeepCSV (black). "D" and "C+D" variants have the best performance in both the branches, "C+D" being marginally better than "D".

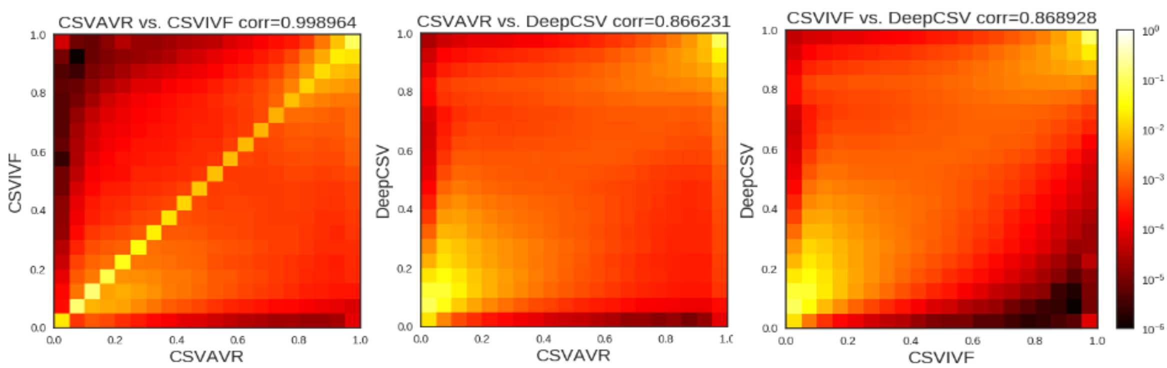


Figure 29: Heat plots showing the correlations between CSVIVF-CSVAVR, DeepCSV-CSVAVR, and DeepCSV-CSVIVF respectively.

10.7.2 Principal Component Analysis of CSV algorithms

Since there is a high degree of correlation among CSVAVR, CSVIVF and DeepCSV, we perform a Principal Component Analysis to extract orthogonal components of the combined CSV information in the three-dimensional CSVAVR-CSVIVF-DeepCSV space. The variations within the orthogonal components can be obtained by diagonalizing the matrix shown in section 10.7.1. The diagonalized matrix is given by

$$\begin{pmatrix} 0.97205 & 0 & 0 \\ 0 & 0.02756 & 0 \\ 0 & 0 & 0.00039 \end{pmatrix}.$$

This matrix shows that about 97.205% of the variation contained within the three CSV variables can be explained with just one variable, while 99.961% can be explained with a combination of two variables. This motivates us to see if the "C+D" variant of Phase-2 cMVAv2 would perform better when trained (and validated) with the PCA-transformed variables of the three CSV values instead of the three CSV variables themselves.

For this, we rerun the XGBoost training step with the other input variables (JP, JBP, SoftMu, SoftEl) and the PCA-transformation of the CSV variables. The comparison of the performances is shown in figure 30. We see a slight degradation of performance in the high-purity region of the b vs. c branch on using PCA-transformed variables to train cMVAv2.

Hence, we discard the usage of Principal Component Analysis in Phase-2 cMVAv2 training.

10.7.3 Null Value Outputs

In some cases, a jet might not have sufficient information that b-tagging algorithms rely on to generate an output discriminator value. In such a case, the algorithm is made to either not produce an output (null) or produce an intermediate default value as output. We check the percentage of null value outputs that different algorithms produce. The results are shown in figure 31.

The plot suggests that Phase-0/1 cMVAv2 always generates an output even when insufficient b-tagging information is available. The percentage of jets for which insufficient b-tagging information is available is 0.13% in the $|\eta|$ range relevant in Phase-1 ($|\eta| < 2.5$). However, for Phase-2, as many as 10.11% of jets in the excess η range ($2.5 < |\eta| < 3.5$) do not have sufficient b-tagging information. With the increased $|\eta|$ range in Phase-2, an overall 1.51% of jets are expected to have information insufficient

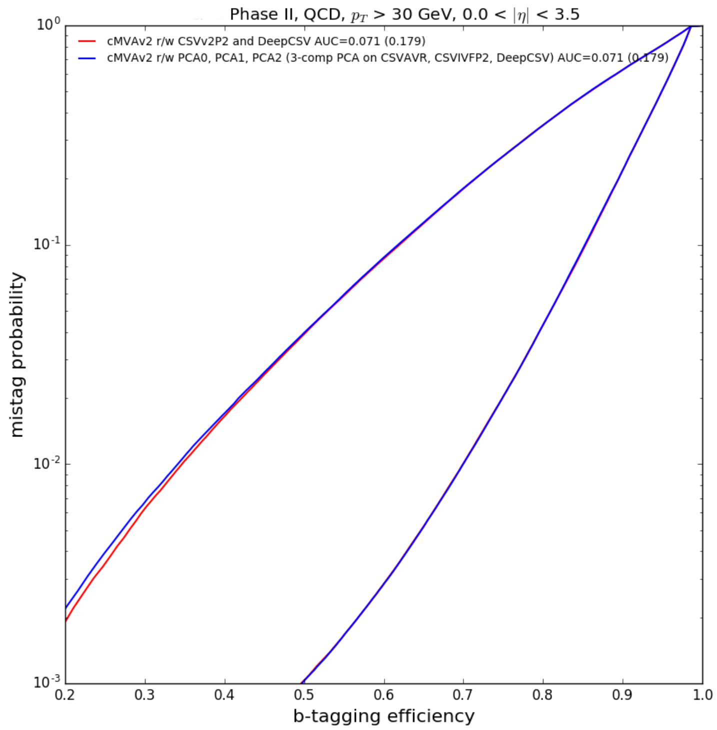


Figure 30: ROC curves for "C+D" variant of Phase-2 cMVAv2 retrained with CSVAVR-CSVIVF-DeepCSV (red) and with PCA-transformed components of CSVAVR-CSVIVF-DeepCSV (blue).

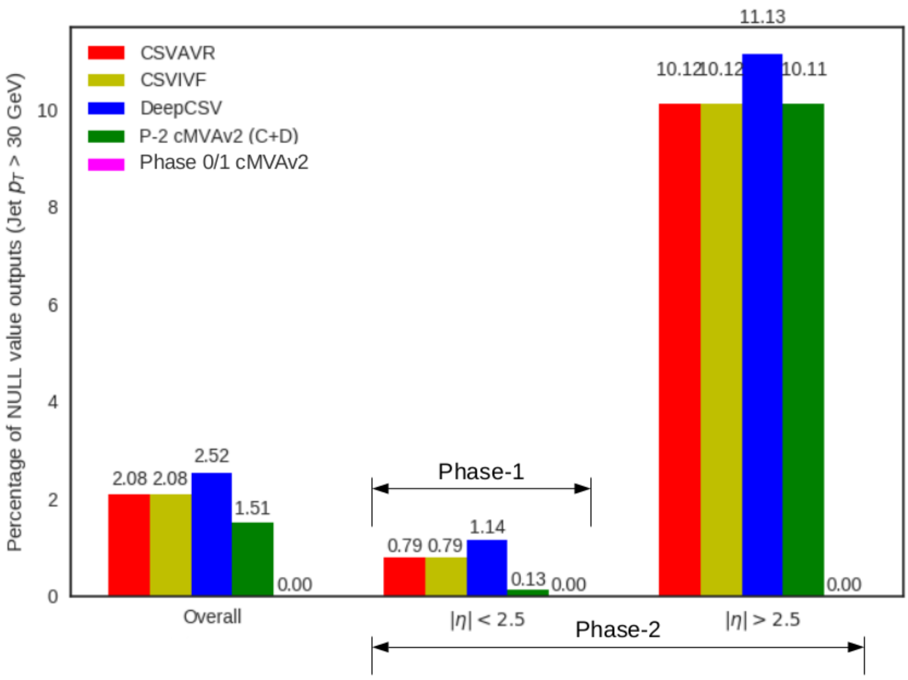


Figure 31: Bar plots showing the percentage of jets for which a null value is returned by the corresponding algorithm, at different η ranges. $|\eta| < 2.5$ region corresponds to the η range relevant in Phase-1, $|\eta| > 2.5$ region shows the additional η range applicable in Phase-2.

for a reliable b-tag discriminator output. Hence, we propose to introduce null values in Phase-2 cMVAv2.

10.8 Conclusion

Retraining the classical variant of cMVAv2 for Phase-2 gives modest improvements over the Phase-0 training of cMVAv2 when validated upon Phase-2 samples. DeepCSV continues to perform better than cMVAv2 in the b vs. c branch. However, integrating DeepCSV output within cMVAv2 gives two new variants of cMVAv2 that have better performance than DeepCSV in both branches.

Currently, DeepCSV and (classical) cMVAv2 are usually recommended for use in analyses because of their superior performance compared to all other b-tagging algorithms. However, integrating DeepCSV within cMVAv2 gives rise to a more robust b-tagger which performs better than both classical cMVAv2 and DeepCSV in both b vs. c and b vs. udsg discrimination. Hence, future analyses at the CMS could consider using the two new variants of cMVAv2 proposed in this paper for improved b-tagging performance.

References

[1] G. Bertone, D. Hooper, J. Silk, Particle Dark Matter: Evidence, Candidates and Constraints, *Phys. Rept.* **405**, 279-390 (2005).

[2] G. Steigman, M.S. Turner, Cosmological Constraints on the Properties of Weakly Interacting Massive Particles, *Nucl. Phys. B* **253**, 375 (1985).

[3] D. Abercrombie et al., Dark Matter Benchmark Models for Early LHC Run-2 Searches: Report of the ATLAS/CMS Dark Matter Forum, arXiv:1507.00966 (2015).

[4] ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, *Journal of Instrumentation* **3**, S08003 (2008).

[5] CMS Collaboration, The CMS experiment at the CERN LHC, *Journal of Instrumentation* **3**, S08004 (2008).

[6] ALICE Collaboration, The ALICE experiment at the CERN LHC, *Journal of Instrumentation* **3**, S08002 (2008).

[7] LHCb Collaboration, The LHCb Detector at the LHC, *Journal of Instrumentation* **3**, S08005 (2008).

[8] TOTEM Collaboration, The TOTEM experiment at the CERN large hadron collider, *Journal of Instrumentation* **3**, S08007 (2008).

[9] LHCf Collaboration, The LHCf detector at the CERN Large Hadron Collider, *Journal of Instrumentation* **3**, S08006 (2008).

[10] B. Acharya et al., The Physics Programme Of The MoEDAL Experiment At The LHC, *Int. J. Mod. Phys. A* **29**, 1430050 (2014).

[11] M. Cacciari, G. P. Salam, G. Soyez, The anti-kt jet clustering algorithm, *J. HEP* **0804** 063 (2008).

[12] S. Agostinelli et al., GEANT4: A Simulation toolkit, *Nucl. Inst. Meth. A* **506**, 250 (2003).

[13] J. Allison et al., Geant4 developments and applications, *IEEE Transactions on Nuclear Science* **53**, 270 – 278 (2006).

[14] J. Alwall et al., MadGraph 5 : Going Beyond, arXiv:1106.0522.

[15] T. Sjöstrand et al., An Introduction to PYTHIA 8.2, *Comput. Phys. Commun.* **191**, 159 (2015).

[16] S. Alioli, P. Nason, C. Oleari, E. Re, A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX, *J. High Energ. Phys.* **06**, 043 (2010).

[17] L. Moneta et al., The RooStats Project, PoS(ACAT2010)057, arXiv:1009.1003 (2010).

[18] CMS Collaboration, CMS Higgs Combine Tool, https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideHiggsAnalysisCombinedLimit (2013).

[19] CMS Collaboration, Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET, *CMS Physics Analysis Summary* CMS-PAS-PFT-09-001 (2009).

[20] CMS Collaboration, Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS detector, *CMS Physics Analysis Summary* CMS-PAS-PFT-10-001 (2010).

[21] Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, arXiv:1603.02754 [cs.LG] (2016).