

# Search for Dark Matter with Graph Neural Networks at CMS

Master Thesis

Jost von den Driesch

At the Department of Physics  
Institute of Experimental Particle Physics

Reviewer:	Prof. Dr. Ulrich Husemann
Second reviewer:	PD Dr. Roger Wolf
Advisor:	Michael Waßmer

Karlsruhe, 13.05.2022



---

This thesis has been accepted by the first reviewer of the master thesis.

**Karlsruhe, 13.05.2022**

.....  
(Prof. Dr. Ulrich Husemann)





---

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

**Karlsruhe, 13.05.2022**

.....  
(Jost von den Driesch)



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Theoretical Background</b>	<b>3</b>
2.1	The Standard Model of Particle Physics . . . . .	3
2.1.1	Electroweak Interaction . . . . .	4
2.1.2	Higgs Mechanism . . . . .	4
2.1.3	Strong Interaction . . . . .	5
2.1.4	Particles . . . . .	5
2.1.5	Limitations of the Standard Model . . . . .	5
2.2	Dark Matter Physics . . . . .	6
2.2.1	Evidence for Dark Matter . . . . .	6
2.2.2	Dark Matter Models . . . . .	7
2.2.3	The WIMP - A Dark Matter Candidate at Hadron Colliders . . . . .	8
2.3	Calculating and Measuring Cross Sections . . . . .	9
<b>3</b>	<b>Experimental Environment</b>	<b>11</b>
3.1	The CMS Experiment . . . . .	11
3.1.1	Kinematic Variables . . . . .	11
3.1.2	Detector Components . . . . .	12
3.1.3	Trigger System . . . . .	13
3.2	Event Simulation . . . . .	14
3.3	Event Reconstruction . . . . .	14
3.3.1	Photons and Charged Leptons . . . . .	14
3.3.2	Quarks and Gluons . . . . .	14
3.3.3	Neutrinos and Dark Matter Particles . . . . .	15
<b>4</b>	<b>Statistical Methods</b>	<b>17</b>
4.1	Fundamental Definitions . . . . .	17
4.1.1	Maximum Likelihood Estimation . . . . .	17
4.1.2	Hypothesis Test . . . . .	18
4.2	Statistical Model at the LHC . . . . .	19
4.2.1	Hypotheses . . . . .	19
4.2.2	Systematic Uncertainties . . . . .	19
4.2.3	Likelihood Function and Test Statistic . . . . .	20
4.2.4	$p$ -Values and Limit Setting Procedure . . . . .	21
<b>5</b>	<b>Neural Networks</b>	<b>23</b>
5.1	Fundamentals . . . . .	23
5.1.1	Artificial Neurons . . . . .	23
5.1.2	Neural Networks . . . . .	24
5.1.3	Training . . . . .	25
5.2	Feedforward Neural Networks . . . . .	25

5.3	Graph Neural Networks . . . . .	26
5.3.1	Graphs . . . . .	26
5.3.2	GNN Process . . . . .	27
<b>6</b>	<b>GraphMET</b>	<b>31</b>
6.1	GraphMET Setup and Training . . . . .	31
6.1.1	Input Features . . . . .	31
6.1.2	General Architecture . . . . .	32
6.1.3	GraphMET Training . . . . .	32
6.2	Comparison Between Different GraphMET Architectures . . . . .	33
6.2.1	Fixed vs. Dynamic GNN . . . . .	34
6.2.2	Graph Convolution Network vs. Edge Convolution Network . . . . .	34
6.2.3	Embedding . . . . .	35
6.2.4	Activation Function . . . . .	36
6.2.5	$\Delta R$ Requirement . . . . .	37
6.2.6	Maximum Numbers of Neighbors Considered . . . . .	37
6.2.7	Number of Graph Layers and Dimension of Feature Vectors . . . . .	38
6.3	Comparison Between GraphMET and Existing MET Estimators . . . . .	40
6.3.1	Previous MET Estimation Methods . . . . .	40
6.3.2	Samples . . . . .	41
6.3.3	Comparison . . . . .	42
<b>7</b>	<b>GraphMET in a Mono-Top Dark Matter Analysis</b>	<b>49</b>
7.1	Mono-Top Signature . . . . .	49
7.2	Event Selection . . . . .	50
7.2.1	Triggers . . . . .	50
7.2.2	Selection Criteria . . . . .	51
7.3	Systematic Uncertainties . . . . .	52
7.4	Validation of Simulation . . . . .	53
7.5	Expected Impact of GraphMET on a Mono-Top Sensitivity Analysis . . . . .	54
7.5.1	Signal Region Plots . . . . .	54
7.5.2	Expected Upper Limits . . . . .	56
<b>8</b>	<b>Conclusion and Outlook</b>	<b>61</b>
	<b>Bibliography</b>	<b>63</b>
	<b>Appendix</b>	<b>73</b>
A	Distributions of GraphMET Input Variables . . . . .	73
B	On the Calculation of Response and Resolution in Data . . . . .	76
C	Unscaled Resolution Plots . . . . .	77
D	Further Control Region and Signal Region Plots . . . . .	78
E	Significance Analysis with Naively Calibrated GraphMET . . . . .	120

## Acronyms

<b>DM</b> Dark Matter .....	1
<b>MET</b> Missing Transverse Energy .....	1
<b>SM</b> Standard Model of particles physics .....	1
<b>CMS</b> Compact Muon Solenoid .....	1
<b>LHC</b> Large Hadron Collider .....	1
<b>WIMP</b> Weakly Interacting Massive Particle .....	1
<b>QCD</b> Quantum Chromodynamics .....	5
<b>EFTs</b> Effective Field Theories .....	7
<b>CMB</b> Cosmic Microwave Background .....	7
<b>LO</b> Leading Order .....	9
<b>NLO</b> Next-to Leading Order .....	9
<b>ECAL</b> Electromagnetic Calorimeter .....	13
<b>HCAL</b> Hadronic Calorimeter .....	13
<b>HF</b> Hadronic Forward Calorimeter .....	13
<b>HLT</b> High-Level Trigger .....	13
<b>PF</b> Particle Flow .....	14
<b>MC</b> Monte Carlo .....	15
<b>JES</b> Jet Energy Scale .....	15
<b>JECs</b> Jet Energy Corrections .....	15
<b>JER</b> Jet Energy Resolution .....	15
<b>ML</b> Maximum Likelihood .....	17
<b>p.d.f.</b> probability density function .....	18
<b>s+b</b> signal+background .....	19
<b>b</b> background-only .....	19
<b>NNs</b> Neural Networks .....	23
<b>GNNs</b> Graph Neural Networks .....	23
<b>GCNConv</b> Graph Convolutional .....	28
<b>EdgeConv</b> Edge Convolutional .....	34
<b>PF MET</b> raw Particle Flow MET .....	40
<b>PUPPI</b> Pile Up Per Particle Identification .....	40
<b>MET-Type-I</b> Type-I corrected PF MET .....	40
<b>GenMET</b> Generated MET .....	42
<b>CR</b> Control Region .....	51
<b>SR</b> Signal Region .....	51



# 1 Introduction

Over the last few centuries, physicists have made great efforts to fundamentally understand nature and have had great success. The key to building models or theories and testing them through experiments is to reduce the complexity of nature. In general, there are two ways to do so. The first is to neglect individual properties of particles and consider larger ensembles of particles that can be described using statistics. This approach is often followed in solid state physics. The second one is to neglect the influences of many particles and focus on aspects that can be described in great detail. This approach has lead to one of the most precise and successful theories in the history of science: the Standard Model of particles physics (SM). It describes three of the four known fundamental forces, the electromagnetic, the weak, and the strong interaction, as well as every fundamental particle that has been observed to date. Only the fourth fundamental interaction, gravity, could not yet be included in the SM. Therefore, even though the SM describes nature very well, it is not complete.

One group of limitations of the SM is found at various cosmological scales. At these scales, there is strong evidence for some sort of matter with a dominant contribution of about 27% to the total energy content of the universe. This matter can interact gravitationally but is not visible and is consequently referred to as Dark Matter (DM). Even though searches for DM have not lead to observations to date, the information gained through these searches has ruled out many possible properties of this kind of matter. One promising type of particle suitable for searches at hadron colliders is called Weakly Interacting Massive Particle (WIMP).

Particles of this type may be produced in proton-proton collisions at the Large Hadron Collider (LHC). Here, bunches of protons are accelerated in opposite directions to proton energies of 6.5 TeV and made to cross each other in four experiments, one of which is the Compact Muon Solenoid (CMS) experiment. However, even if DM particles are produced in some of the resulting proton-proton collisions, they cannot be directly detected by experiments such as CMS. The reason for this is the weakness of the interaction between SM particles, which make up the detector, and possible DM particles. The only hint at a produced DM particle can come from a seemingly disturbed momentum conservation in the direction transverse to the motion of the protons. Being defined as the difference between the expected sum of the transverse momentum of all detected particles, that is zero, and the observed sum, Missing Transverse Energy (MET) provides indirect information on possible DM particles. However, DM is not the only source of MET, as, for instance, neutrinos also leave the detector undetected. Moreover, measurement uncertainties as well as pileup,

i.e. additional collisions that occur during the same bunch crossing, pose challenges for estimating MET. One approach to deal with these challenges is the introduction of weights to the aforementioned transverse momentum sum. By applying weights close to one to particles from the main collision and small weights to particles from pileup, the unwanted contribution to the transverse momentum sum can be mitigated.

The first part of this thesis deals with GraphMET, a new machine learning based approach to determine these weights. As opposed to previous reconstruction methods, GraphMET combines the advantages of a per particle description and a per event description, thus benefiting from the usage of information from the correlation between different particles. For this, GraphMET uses Graph Neural Networks, an advanced type of Neural Network that operates on graph structures.

The second part focuses on a first application case for GraphMET in a physics analysis. Basis for this analysis is a simplified DM model with an additional vector mediator and a stable DM particle. The mediator couples to SM particles as well as DM particles, and allows flavor changing neutral currents, which are strongly suppressed in the SM. Therefore, an overshoot of events with the signature of a top quark and large MET in the final state with respect to SM processes strongly hints at DM being present. The goal of this second part of the thesis is to investigate the impact of using GraphMET on the expected sensitivity of this analysis.

In order to provide the reader with the information necessary to understand the results, chapter 2 and 3 give an overview on the theoretical and experimental background, respectively. The statistical model for the sensitivity study is presented in chapter 4. Chapter 5 provides an explanation of conventional Neural Networks as well as a detailed introduction to Graph Neural Networks.

The following two chapters deal with the development and application of GraphMET. In chapter 6, different setups of GraphMET are investigated and a detailed comparison with previous MET reconstruction methods on large unbiased data sets is given. Then, the sensitivity study of different MET reconstruction methods is explained, performed and interpreted in chapter 7. Finally, a conclusion is drawn and an outlook on future developments as well as further improvements is provided in chapter 8.



## 2 Theoretical Background

The SM is a consistent theory that describes all currently known elementary particles and three of the four known fundamental interactions. Beginning with the first observation of an elementary particle, the electron, in 1896 by Thomson, more and more particles have been found and incorporated into the model or predicted and confirmed by experiment [1]. With the correct prediction of particles such as the top quark [2–5] and the Higgs boson [6–10], the theory has proven to be highly successful. The SM’s key concepts are presented in section 2.1. Although the theory is well established to date, it has a few important weaknesses that have led to new ideas and concepts collectively known as physics beyond the Standard Model. One family of weaknesses and possible solutions, DM, is presented in section 2.2. Section 2.3 deals with the calculation and measurement of cross sections as a way to compare theoretical predictions to experimental results.

### 2.1 The Standard Model of Particle Physics

This section is held rather phenomenological. For further mathematical details and general understanding, the reader is referred to [11–13], on which this section is based.

The SM makes use of the Lagrange formalism, which has already been applied successfully in classical physics and is briefly introduced here for illustrative purposes. In classical physics, the Lagrange function  $\mathcal{L}$  of a physical system of  $N$  particles with generalized coordinates  $q_i$  and momentum  $\dot{q}_i$  is given by

$$\mathcal{L} = T - V, \tag{2.1}$$

with kinetic energy  $T = T(\dot{q}_1, \dot{q}_2, \dots, \dot{q}_N, t)$  and a potential  $V = V(q_1, q_2, \dots, q_N, t)$ . Using this Lagrange function, the equation of motion for particle  $i$  can be derived via

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_i} - \frac{\partial \mathcal{L}}{\partial q_i} = 0, \tag{2.2}$$

which is a result of minimizing the action

$$S(x) = \int \mathcal{L}(q_i(t), \dot{q}_i(t), t) dt. \tag{2.3}$$

This whole concept is generalized from discrete point-like particles to continuous functions called fields. In general, one has to distinguish between different types of fields, in particular

between fields with integer (bosonic) and half-integer (fermionic) spin in units of  $\hbar$ . In order to describe particles as fields, these fields have to be quantized. Quantization ensures that particles can be created or annihilated. Quantized bosonic fields are symmetric under particle exchange, whereas fermionic fields are anti-symmetric under such exchanges.

The Lagrange function has now become a Lagrange density but still contains the information necessary to describe the dynamics of the system. The equations of motion for these fields are the relativistic field equations, e.g. the Klein-Gordon equation for bosonic fields and the Dirac equation for fermionic fields.

The SM Lagrangian is invariant under three types of (local) gauge symmetry operations that are described by the continuous symmetry groups  $U(1)$ ,  $SU(2)$  and  $SU(3)$ . Symmetries are always associated with a quantum number, i.e. a conserved quantity. Fermions can be grouped by their quantum numbers and the number of such groups is not limited within the SM. The number of gauge bosons, on the other hand, is constrained by the dimension of a symmetry group.

### 2.1.1 Electroweak Interaction

The electroweak interaction is invariant under the combined symmetry  $SU(2)_L \times U(1)_Y$ . Here,  $L$  indicates that the  $SU(2)$  symmetry, which has the third component of the weak isospin as a quantum number, applies only to left-handed fermions, and the weak hypercharge  $Y$  is the quantum number of the  $U(1)$  symmetry. With dimensions of  $SU(2)$  and  $U(1)$  of three and one, respectively, the total dimension is four, indicating the existence of four gauge bosons. One of these gauge bosons, the  $B^0$  boson, couples only to particles with non-vanishing hypercharge. The three remaining gauge bosons,  $W^0$ ,  $W^1$  and  $W^2$ , couple only to particles with non-vanishing weak isospin, such as left-handed fermions.

However, the electroweak symmetry is broken by the Higgs mechanism, which will be explained in the following subsection. The physically observed bosons are complex linear combinations of the aforementioned electroweak bosons:

$$\begin{pmatrix} \gamma \\ Z \\ W^+ \\ W^- \end{pmatrix} = \begin{pmatrix} \cos \theta_W & \sin \theta_W & 0 & 0 \\ -\sin \theta_W & \cos \theta_W & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -i\frac{1}{\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} & i\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} B^0 \\ W^0 \\ W^1 \\ W^2 \end{pmatrix}, \quad (2.4)$$

where  $\theta_W = \arccos \frac{m_W}{m_Z}$  is the Weinberg angle and  $m_W$ ,  $m_Z$  are the masses of the physically observable  $W^\pm$  and  $Z$  bosons. The massless Photons are represented by  $\gamma$  and couple only to particles with electric charge, which is a linear combination of the weak isospin and the weak hypercharge. The  $Z$  boson couples only to weakly charged particles, also being a linear combination of weak isospin and weak hypercharge. At last, the  $W^\pm$  bosons couple only to particles with weak isospin, as their  $B^0$ -portion is zero.

### 2.1.2 Higgs Mechanism

Before the Higgs mechanism was introduced, the aforementioned masses of the  $W^\pm$  and  $Z$  bosons could not be consistently described by the Lagrange formalism. Naively introducing additional mass terms in the Lagrange density, for example, would violate the gauge invariance of the theory. Instead, a new complex scalar field  $\phi$  is introduced with the potential

$$V_H = -\mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2, \quad (2.5)$$

with parameters  $\mu$  and  $\lambda$ . The point of symmetry lies at  $\phi = 0$ . The energetic minimum, however, is given by a circle around the symmetry point.

As the minimum is not the symmetry point, the symmetric system is not stable and tends to fall in the unsymmetrical minimum state. This is called spontaneous symmetry breaking and explains the generation of the electroweak gauge boson masses within the SM by construction. The fermion masses must be generated by introducing an additional interaction between the Higgs field and the fermion field, the Yukawa interaction.

### 2.1.3 Strong Interaction

Quantum Chromodynamics (QCD) is currently the best means of describing the strong interaction. The physics of strong interaction is invariant under gauge transformations represented by the group  $SU(3)$ , which has dimension eight. The quantum numbers of this symmetry group are the colors red, green and blue. Since the dimension is eight, there are eight gluons, each carrying the same number of colors and anti-colors. Gluons are massless and do not have any charge but their color. Thus, they interact only with color-charged particles and therefore also among themselves.

Unlike other quantum numbers, no color-charged particle can be observed in nature as free particle. As a consequence of this color confinement, only color-neutral bound states consisting of two or more quarks can be observed, so called hadrons. Color confinement results in effects like hadronization, where single-color particles are shielded by the formation of multiple quarks and gluons, which then form color-neutral hadrons.

### 2.1.4 Particles

The elementary particles within the SM are shown in figure 2.1. Leptons and quarks are fermions and each of them has an antiparticle, not shown here for clarity. The fermions are each divided into weak isospin doublets, referred to as flavor, generation or family. Except for right-handed neutrinos they all carry weak charge and therefore interact with  $Z$  bosons. Transitions between left-handed up-type and down-type quarks or charged lepton and neutrino, respectively, are realized by interactions with  $W$  bosons. Such transitions are also possible for the corresponding right-handed antiparticles, but neither for right-handed particles nor for left-handed antiparticles.

Similar to the  $W^\pm$  and  $Z$  bosons, the mass eigenstates of the quarks do not coincide with the flavor eigenstates. As a result, transitions between quark generations are possible via  $W$  bosons. The transition probabilities are described by the CKM matrix. Quarks and gluons are the only particles with color, so only they interact via the strong interaction.

### 2.1.5 Limitations of the Standard Model

Although the SM describes a large number of processes in great detail, several examples have been found that cannot be described by the SM and provide evidence for physics beyond it.

The strongest proof for that is found in the neutrino sector. According to the SM, neutrinos should have no mass and should not be able to oscillate. Studies of solar [15–18], atmospheric [19], reactor [20, 21] and beam [22–24] neutrinos, however, prove both assumptions to be false.

Another argument for the incompleteness of the SM is its inability to account for gravity. As one of the four fundamental forces, gravity is crucial to the correct description of nature, especially at large scales and near large masses, such as black holes. On small scales and far from large masses, gravity is typically very weak compared to the other fundamental forces. Therefore, most physical phenomena can be described by either general relativity [25, 26], a theory of gravitation, or the SM, but there is no complete description unifying all fundamental forces.

In addition to that, the SM is unable to explain several effects, which could be solved by introducing a new kind of particle, DM, as described in the next section.

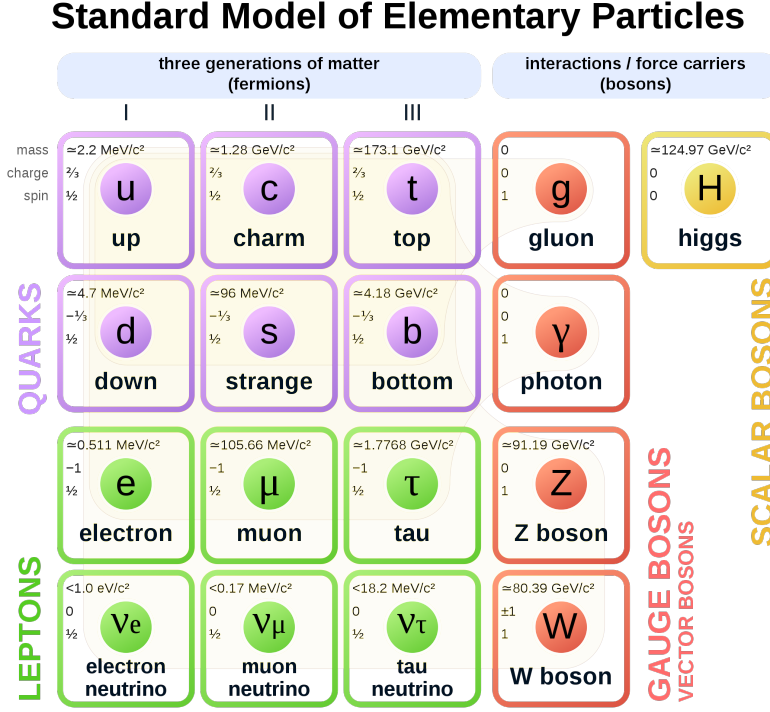


Figure 2.1: Elementary particles of the SM [14].

## 2.2 Dark Matter Physics

This section is based on [27, 28]. The existence of DM is well-motivated by strong evidence at different cosmological scales.

### 2.2.1 Evidence for Dark Matter

The cosmological scales at which evidence is found range from galaxies, i.e. well-separated clusters of stars, to clusters of galaxies and the structure of the universe as a whole.

The first type of evidence can be observed in our neighboring galaxies, e.g. the Triangulum Galaxy. Due to the small distance to earth of about 850 kpc, the Triangulum Galaxy is among the best observed galaxies. Since most of the visible mass is located at the center of galaxies, the mass  $M$  within a sphere of radius  $R$  is expected to be almost constant with respect to the radius for large  $R$ . Therefore, a test mass  $m$  moving around the galaxy center in a circular movement in distance  $R$  is kept in orbit by the gravitational force:

$$F_G = F_{cp} \Leftrightarrow G \frac{mM}{R^2} = \frac{mv^2}{R} \Rightarrow v = \sqrt{\frac{GM}{R}} \propto \frac{1}{\sqrt{R}}, \quad (2.6)$$

with the gravitational constant  $G$ . Observations of the rotational velocity have shown that for large  $R$  applies  $v = \text{const.}$ , which gives rise to additional mass with density  $\rho \propto \frac{1}{R^2}$ . Figure 2.2 shows the expected and measured rotational curve of the Triangulum Galaxy. Measurements of other galaxies show similar behavior [29].

In the study of galaxy clusters, which usually consist of several hundred galaxies, there are several pieces of evidence for DM. Zwicky, for example, discovered inconsistencies in the kinematic distribution of the Coma Cluster when only considering visible matter [31]. Another example is the Bullet Cluster, which consists of two galaxy clusters that have

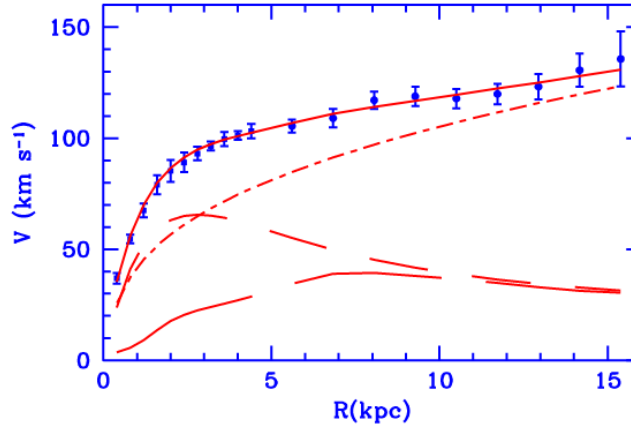


Figure 2.2: Rotation curve of M33, the Triangulum Nebula, as a function of the distance to the galaxy center. While the measured values are shown as blue points, the continuous line shows the best fit model, consisting of three submodels: A thin exponential stellar disk (short dashed line), a gaseous disk (long dashed line), and a dark spherical mass distribution (dashed dotted line) [30].

passed each other [32]. Here, the visible mass is not aligned with the total mass, which can be investigated using gravitational lensing, an effect described by general relativity [25, 33, 34]. The distributions suggest that there is a large fraction of the mass, which was not affected by the passing, i.e. had not interacted by any of the non-weak SM forces.

Cosmological models of structure formation depend heavily on DM as a major energy density component in the universe. In particular the Cosmic Microwave Background (CMB) has been studied to obtain an approximation of the amount of DM in the universe [35, 36]. The CMB is a remnant from about 380 000 years after the Big Bang, where protons and electrons froze out of the thermal equilibrium and formed bound states of atoms. Fluctuations in the CMB, which can be converted into a multipole spectrum, indicate fractions of the total energy density of the universe of about 27% for DM and 5% for visible matter.

Some of these issues could also be solved without the introduction of DM but these methods are mostly ad hoc and can only describe one piece of evidence at a time. DM on the other hand can explain all of the aforementioned evidence. For that, DM models must ensure that there are particles that interact via gravity but not electromagnetically or strongly, as they would have been observed already otherwise.

### 2.2.2 Dark Matter Models

Since DM is not included in the SM, the SM must be extended. In this respect, there are two fundamentally different approaches, effective field theories and complete models. Somewhere in between, there are simplified models that contain elements of both approaches.

#### Effective Field Theories (EFTs)

Often, physics can be described by neglecting or approximating physical theories, which play a role at different lengths and energy scales. This is also the idea behind EFTs [37]. Newtonian dynamics, for example, are a low energy approximation of general relativity. General relativity is a macroscopic physical approximation which neglects quantum theories and vice versa. In general, these theories work well in their defined area.

In particle physics, EFTs are frequently used to approximate the behavior of mediator particles that have a mass larger than the experiment's energy scale. Thanks to Heisenberg's uncertainty principle, particles can "borrow" a certain amount of energy from the vacuum for an amount of time inverse proportional to the amount of energy. This mediator can therefore increase the probability for an interaction even though its mass is larger than the energy of the process. Instead of explicitly integrating the particle into the model, EFTs focus on the mediators effect of an increased probability by approximately describing it as a small change in the coupling constant.

The main benefit of using EFTs for DM is their simplicity and model-independent approach. However, this approach can only properly describe physics below a particular energy scale, where the approximation is still reasonable.

### Complete Models

A fundamentally different approach is to formulate a complete model, such as supersymmetry [38]. This model should be able to describe observations which could not be explained by the previous model. These models typically come with a vast number of free parameters, which need to be constrained experimentally.

The benefit of a complete model is that it not only describes but also gives an explanation for new physics and predictions of new effects. On the other hand, these models tend to have so many free parameters that they can be quite unspecific and challenging to constrain.

### Simplified Models

These models have properties of both of the aforementioned approaches containing at least one mediator and one stable DM particle. As a result, they have both flaws and advantages of complete models and EFTs.

Compared to EFTs, the mediator is not included into the coupling. Therefore, simplified models make more assumptions on the mediator than EFTs. However, this comes with the benefit of not being restricted to mediator masses much larger than the energy of the process. Compared to complete models, simplified models do not have to be consistent. On the one hand, this is a disadvantage, because it means that the physical context is not fully understood. On the other hand, the number of free parameters is much smaller, so the model is easier to constrain and better suited to the problem than complete models.

### 2.2.3 The WIMP - A Dark Matter Candidate at Hadron Colliders

There are many ways to extend the SM in order to describe DM. However, observations restrict the properties of possible DM candidates. A general overview on DM is given in [13]. The most promising DM candidate for the search at hadron colliders is the WIMP. As all DM candidates, WIMPs have zero electric charge and therefore do not interact with photons. Various studies, summarized e.g. in [39], have shown that the cross section for interaction with visible matter is close to or lower than typical cross sections of the weak interaction. DM should therefore not interact via the strong interaction, otherwise the cross section would be much higher, unless the DM candidate is too heavy to be produced in experiments. Additionally, the candidate should be stable on cosmological scales, otherwise it would not explain the cosmological evidence.

With a mass range of  $\mathcal{O}(10 \text{ GeV})$  up to  $\mathcal{O}(1 \text{ TeV})$ , WIMPs are in principle in the measurable scope for the CMS experiment, as previous studies have shown, e.g. [40, 41]. Yet, certain conditions must be met for a statistically significant observation: There must be an interaction between WIMPs and SM particles strong enough to being measured and the

WIMP's mass must be small enough such that the center-of-mass energy is sufficient for the production. The quantity combining these two criteria is the cross section, which will be described in the next section.

Since the interaction of DM with SM particles is at most weak, the vast majority of WIMPs will leave the detector without interaction. Therefore, WIMPs cannot be observed directly but only indirectly via conservation of momentum. Further explanations can be found in sections 3.3.3 and 7.1.

## 2.3 Calculating and Measuring Cross Sections

A cross section contains information on the interaction strength as well as the kinematics of a process. In collider physics, for example, the cross section of an interaction of two particles with four-momenta  $p_1, p_2$  and masses  $m_1, m_2$  in the initial state and a certain set of particles with their set of four momenta  $P_f$  in the final state can be calculated via

$$d\sigma = \frac{(2\pi)^4 |\mathcal{M}|^2}{4\sqrt{(p_1 \cdot p_2)^2 - m_1^2 m_2^2}} \times d\Phi_n(p_1 + p_2; P_f). \quad (2.7)$$

In this calculation there are two major steps: The matrix element  $\mathcal{M}$  has to be computed and also the differential phase space  $d\Phi$ .

The matrix element describes the probability for all possible transitions from the initial state  $i$  to the final state  $f$  and thus contains information about the interaction strength:

$$\mathcal{M}_{i \rightarrow f} (2\pi)^4 \delta^{(4)}(p_1 + p_2 - \sum_j P_{f,j}) \propto \langle P_f | \frac{1}{i}(S - 1) | p_1, p_2 \rangle, \quad (2.8)$$

where  $\delta$  denotes the delta distribution, which ensures four-momentum conservation,  $\langle . | . \rangle$  is the usual bra-ket notation from quantum physics and  $S$  is the scattering matrix

$$S = \mathbf{T} \exp \left( i \int_{\mathbb{R}^4} \mathcal{L}_I(x) dx \right). \quad (2.9)$$

Here,  $\mathcal{L}_I$  describes the part of the Lagrange function which is responsible for an interaction and  $\mathbf{T}$  is the time-ordering operator. The scattering matrix contains an exponential function, which can be Taylor-expanded. The more series terms are calculated, the more accurate is the calculation. Considering only the first non-vanishing series terms is called Leading Order (LO) calculation. Considering a second non-vanishing term is called Next-to Leading Order (NLO) calculation and so on.

Each term of the series expansion of  $S$  plugged into equation 2.8 is a contribution to the whole matrix element. Using Wick contractions [42], each contribution can be split to a number of processes which can be graphically depicted as Feynman diagrams. These diagrams are used to obtain an intuitive understanding of the rather complex integrals in this calculation.

After having obtained the matrix element, the phase space is calculated and integrated. This integral comprises kinematic information on the initial and final state. The phase space integral depends on the difference between the center-of-mass energy in the initial state and the mass of the particles in the final state. The larger this difference is, the larger the number of ways to split the remaining energy among the final state particles, and thus the larger the phase space. As the phase space is part of the cross section, the probability of interaction increases for larger phase spaces. If, for example, the mass of a particle in the final state exceeds the center-of-mass energy, the phase space integral is zero.

In order to test theoretical models, calculated cross sections are compared to experimentally observed ones. Experimentally, the cross section of a certain process can be obtained from

$$\sigma = \frac{N}{\int L dt}, \quad (2.10)$$

where  $N$  is the number of measured processes and  $\int L dt$  is the integrated luminosity, a measure of the total amount of data taken by an experiment. Since cross sections of different processes are addable and the exclusive measurement of one process is usually not possible, experiments measure several cross sections  $\sigma_i$  at once, thus:

$$\sum_i \sigma_i = \sum_i \frac{N_i}{\int L dt} = \frac{N_{\text{tot}}}{\int L dt}, \quad (2.11)$$

where  $N_i$  denotes the experimentally inaccessible number of observations of a specific process and  $N_{\text{tot}}$  is the experimentally accessible total number of observed processes. If all cross sections  $\sigma_i$  except for one,  $\sigma_0$ , are known, then the unknown cross section is experimentally accessible via

$$\sigma_0 = \frac{N_{\text{tot}}}{\int L dt} - \sum_i \sigma_i. \quad (2.12)$$

Even though this explanation is simplified, the general concept holds for any experiment that searches for new particles, so in particular the experiment used in this thesis.



## 3 Experimental Environment

The LHC [43] accelerates protons in two opposing beam lines to kinetic energies of 6.5 TeV. At four points within the LHC, these proton beams cross each other leading to interactions. If the impact parameter, i.e. the distance between two interacting protons if they would not interact, is small enough, the annihilating kinetic energy is sufficient for the production of heavy particles. These new particles are investigated by four experiments, one of which is the CMS experiment [44].

Section 3.1 describes the basic principle of the CMS experiment, whereas sections 3.2 and 3.3 give more details on the event simulation and the reconstruction of different types of particles.

### 3.1 The CMS Experiment

The CMS experiment is a cylindrical multi layer detector, which aims at identifying the final-state particles and reconstructing their properties. An overview of the detector is depicted in figure 3.1.

#### 3.1.1 Kinematic Variables

With the  $z$  axis along the beam line and symmetry axis of CMS, a Cartesian coordinate system is fixed and a set of kinematic quantities of a particle can be defined:

- $p_x, p_y, p_z$ : Momentum in Cartesian coordinates
- $p_T = \sqrt{p_x^2 + p_y^2}$ : Transverse momentum (also called transverse energy)
- $p = \sqrt{p_x^2 + p_y^2 + p_z^2}$ : Absolute momentum
- $\phi = \arctan\left(\frac{p_y}{p_x}\right)$ : Azimuthal angle of the momentum
- $\eta = \frac{1}{2} \ln\left(\frac{p+p_z}{p-p_z}\right)$ : Pseudorapidity, contains information about the polar angle between the particle's momentum and the  $z$  axis

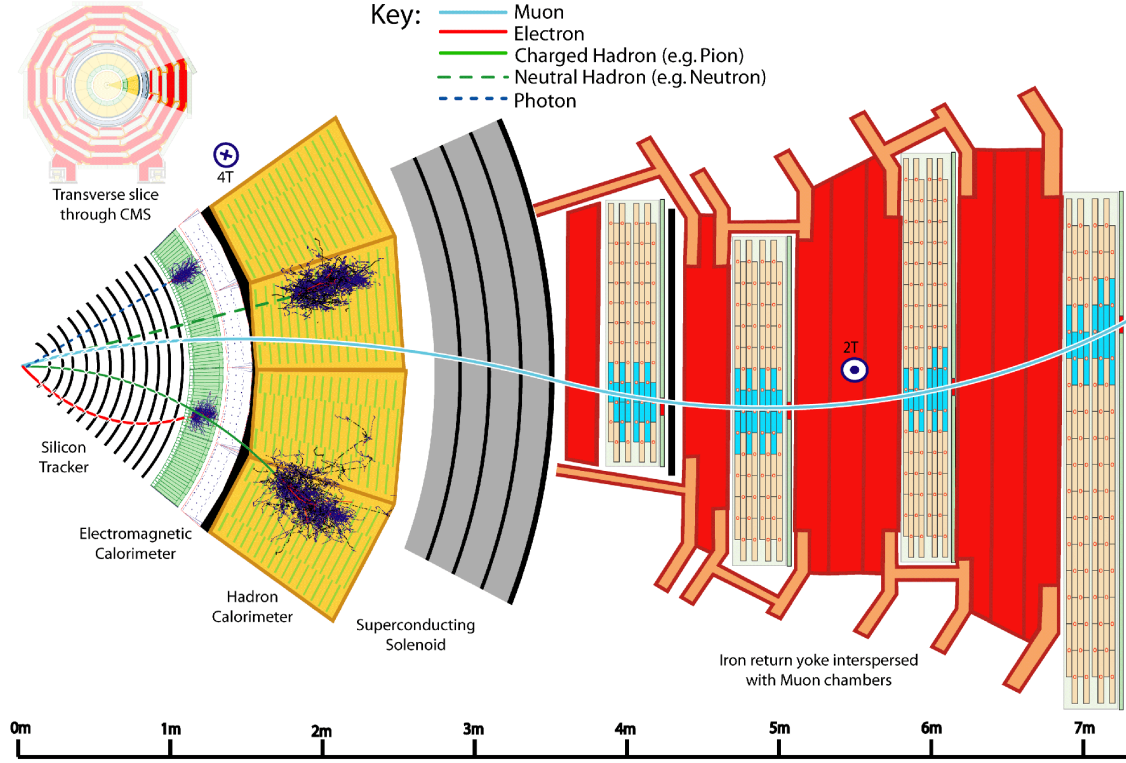


Figure 3.1: Illustration of the CMS detector and its components [45]. The red line describes a positively charged electron, which can be recognized from the curvature. The muon and the charged hadron are negatively charged, and the dashed lines describe neutral particles with a low probability of leaving a track in the silicon tracker.

### 3.1.2 Detector Components

One eponymous component is the superconducting solenoid, which creates a magnetic field of 3.8 T aligned to the beam direction [46]. It helps to reconstruct the particle's charge and momentum, as charged particles are deflected on a spiral track by the resulting Lorentz force  $F_L = qv_T B$ , with charge  $q$ , velocity transverse to the magnetic field  $v_T$  and magnetic field strength  $B$ . A particle's charge can be obtained by measuring the direction of deflection. Given the radius  $r$  of the spiral, one can determine the particle's transverse momentum  $p_T$ , as the centripetal force needed for the circular movement is applied by the Lorentz force:

$$F_{cp} = F_L \Leftrightarrow p_T \frac{v_T}{r} = qv_T B \Leftrightarrow p_T = qBr. \quad (3.1)$$

The particle tracks are measured precisely in the innermost layer of CMS, which consists of silicon pixel and strip detectors. These detectors are made of semiconducting material and use ionization to detect passing particles. Key to the measurement is a p-n junction with a voltage large enough to pull all of the free charge carriers out of the main volume of the sensor, called bulk. When a charged particle moves through this depletion zone, it loses a part of its energy to the material, creating pairs of electrons and holes, which then are moved to the poles and detected by electronics. By combining the information of pixels and strips in different layers, it is possible to reconstruct tracks. If several tracks can be extrapolated to a common origin, this origin is called vertex. A primary vertex is the collision point of two protons [47] and can be reconstructed with a resolution of  $\mathcal{O}(10\mu\text{m})$  if at least  $\mathcal{O}(10)$  tracks are available for the reconstruction [48]. More details can be found in [49].

After the track detector, the particles run into the Electromagnetic Calorimeter (ECAL), which is made of lead-tungstate crystals. These crystals are scintillators, meaning that its molecules get excited by the high-energy photons or charged particles. In the presence of a nucleus, a photon with at least the energy of two electron masses can produce an electron-positron pair. Electrons and positrons experience an attractive or repelling force close to a nucleus and therefore emit bremsstrahlung photons. Repeated occurrence of these two processes results in an electromagnetic cascade, ending if the photon energy is not large enough for the production of the electron-positron pair and the electrons or positrons are caught by a nucleus or annihilate. This cascade should be completely contained in the ECAL to measure the total energy content of the original particle. Therefore, the crystals are about 25 times as long as the radiation length of lead-tungstate. The light emitted by the scintillator material is then detected by photodetectors. As the amount of light emitted by the material is proportional to the energy deposited by the particle, the initial particle's energy can be measured with a resolution of  $\mathcal{O}(1\%)$  as described in [50, 51]

The Hadronic Calorimeter (HCAL) [52] uses the same detection principle as the ECAL. Due to the larger mass of hadrons, they are not affected as much by bremsstrahlung as electrons. Hence, the principle to slow these particles down is to use sampling calorimeters that consist of alternating layers of scintillators for the energy measurement and absorber modules made of brass or steel to increase the interaction probability. The barrel and endcap component of the HCAL range up to a pseudorapidity of  $\eta = 3.0$ . As a part of the HCAL, the Hadronic Forward Calorimeter (HF) provides information up to a pseudorapidity of  $\eta = 5.0$  in order to also measure particles that move in the forward direction close to the beam line.

Outside of the solenoid, the direction of the magnetic field changes. Here, the only remaining measurable particles are muons, which are minimally ionizing particles at typical LHC energies. Muons are then detected in the muon chambers [53], which are gas-filled detectors with an applied electric field, where muons deposit energy by ionizing the gas. Then the free electrons and positive ions drift towards the electrodes and are measured in this process. The efficiency of muon identification ranges from 95% and 99% for tight criteria and even exceeds 99% for loose identification criteria [54].

### 3.1.3 Trigger System

In order to illustrate the idea of triggers, a short estimation on the amount of data is given, based on [55]. The LHC accelerates bunches of about  $10^{11}$  protons to energies of 6.5 TeV. Every 25 ns there is a bunch crossing, resulting in an event frequency of about 40 MHz. With a mean number of 29 interactions per bunch crossing during LHC Run 2 [56], an estimated 1 MB of storage is needed per bunch crossing. Thus, without triggers, the entire 350 000 TB of tape-based storage in CERN's Tier-0 center [57] would be filled after roughly 2.5 h by CMS alone, assuming that such a high data transmission rate could be provided. This simple calculation shows the relevance of some sort of event filter that is supplied by a two-level trigger approach [58].

The Level-1 Trigger reduces the number of events by a factor of 400 by evaluating information from the muon detectors and calorimeters using dedicated hardware and checking certain criteria that should ensure the relevance of the observed event, such as a threshold of a particle's energy [59]. If an event passes one such criterion, the whole detector is read out and the data is forwarded to the High-Level Trigger (HLT) that uses conventional computing hardware and methods to evaluate higher level information and reduce the amount of data by another factor of 1000.

## 3.2 Event Simulation

Following the discussion in section 2.3, the steps in the event simulation can be summarized as follows [60]:

1. Matrix element generation. Here, the basic structure of the event is obtained, i.e. the Feynman diagrams at some order.
2. Simulation of all parton-level processes. This step includes, for example, the simulation of initial and final state radiation and of the underlying event, i.e. the processes of the partons in the proton that are not involved in the main interaction.
3. Simulation of further processes, i.e. hadronization and decays of unstable particles.
4. Simulation of detector-related quantities, such as the detector response.

In this thesis, the first step is done using `MadGraph5_aMC@NLO` [61] or `POWHEG` [62–64]. `PYTHIA8` provides the tools to perform steps 2 and 3. The last step of simulating the detector response behavior is done using `GEANT4` [65].

## 3.3 Event Reconstruction

In order to identify particles, the Particle Flow (PF) algorithm [66] is applied. It combines the information of the different components of CMS such that a global event description is provided. The result of the PF algorithm is a set of PF candidates that are estimates on the particles measured by the detector. Usually, the detector measures multiple collisions at once, where only one collision is investigated and the remnants from the others, pileup, should be mitigated as much as possible. The basic concept of identification for the different sorts of particles is described in the following subsections.

### 3.3.1 Photons and Charged Leptons

Electrons are the lightest charged leptons, so they do not decay but slow down while emitting bremsstrahlung and colliding with particles in the ECAL and finally get absorbed into the material. Their characteristics are therefore a curved track and a signal in the ECAL without a signal in the outer detector layers.

Muons at typical CMS energies are minimally ionizing particles, which is why they only deposit a small fraction of their energy in the detector. They are detected by their curved track throughout the entire detector, especially in the outer laying muon chambers.

Tauons, however, are much heavier than the other charged leptons and therefore decay rapidly into a tau-neutrino, which is a source of MET, and an off-shell  $W$  boson. This  $W$  boson then decays hadronically in about two thirds of the cases and leptonically in the remaining third.

Photons behave similar to electrons in the ECAL. Unless a photon is converted to an electron-positron pair on the way, it passes directly through the tracker and then leaves a signal in the ECAL.

### 3.3.2 Quarks and Gluons

Hadrons are bound states of quarks and gluons, both of which are also called partons. Since colored states cannot exist as free states, an emerging parton leads to the production of multiple quarks and gluons via the strong interaction in order to shield the single color. These quarks and hadrons restructure into hadrons, called hadronization. In this process, other interaction types can occur and lead to the production of photons or leptons. Momentum conservation ensures that these particles are all going into roughly the same

direction, especially when considering higher momenta of the emerged parton. In order to obtain information on the initial parton which originated from the collision, one has to gather all information of the produced particles. The particles in the final state are clustered with algorithms such as the anti- $k_T$  algorithm and the resulting cluster is called jet. These algorithms usually make use of the direction information of the possible jet constituents.

The anti- $k_T$  jet clustering algorithm [67] is an iterative algorithm, which combines in every iteration two particles  $i$  and  $j$  until none are left that satisfy a certain criterion. The condition for combining two particles into one is that the distance measure between them is minimal. This measure is defined as

$$d_{ij} \equiv \min(k_{Ti}^{-2}, k_{Tj}^{-2}) \frac{\Delta_{ij}^2}{R^2}, \quad (3.2)$$

where  $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$  is the distance in the transverse plane,  $R$  the radius parameter and  $k_{Ti}$  the transverse momentum of particle  $i$ . The additional criterion mentioned before is  $d_{ij} < k_{Ti}^{-2}$ . As long as this criterion is fulfilled, particles are combined and if it is not fulfilled anymore, the entity  $i$  is called a jet and not considered in the following iterations anymore. This criterion ensures that only particles roughly within the radius  $R$  are combined into a jet. If there are two jets within a radius of  $2R$  the distance measure ensures that the particles in the overlapping area tend to be assigned to the area with higher density of particles, resulting in a lower  $\Delta_{ij}$  and higher transverse momentum, leading to a lower  $\min(k_{Ti}^{-2}, k_{Tj}^{-2})$ .

Since jets consist of multiple constituents, it is important to calibrate them using Jet Energy Corrections (JECs) [68–71]. The first step is to remove pileup, remnants from other, usually low-energy, proton-proton collisions. After that, further Monte Carlo (MC) based as well as data driven methods are applied for response corrections and optionally flavor corrections, such that the differences in Jet Energy Scale (JES) between data and MC simulations are minimized. In order to improve the agreement between data and simulation in the Jet Energy Resolution (JER), another set of corrections is applied.

Jets from different quark flavors behave differently. This different behavior can be used to tag jets [72]. Bottom quarks and their jets mark a most prominent case. The main difference to other quarks is the stability of bottom quarks, which by kinematics should decay rapidly as their mass is much larger than the one of the charm or up quark. However, the corresponding CKM matrix elements are very small, leading to longer lifetimes. As the particles move close to the speed of light, the longer lifetime often results in a significantly shifted vertex, which can be used to distinguish bottom quark jets from other jets.

### 3.3.3 Neutrinos and Dark Matter Particles

Neutrinos undergo only weak interactions. This makes it impossible to measure their momentum directly, which is also the case for possible DM particles. Nevertheless, it is possible to obtain information on these particles' momenta by constructing MET. This quantity makes use of momentum conservation in the transverse plane, where the total initial and final state momenta have to be zero due to the protons colliding along the orthogonal  $z$ -axis. By summing up the transverse momenta  $\vec{p}_{T,k}$  of all  $N$  reconstructed particles and changing the sign, an estimator is obtained for the transverse momentum that is missing, for example because of particles that undergo only weak interactions:

$$\vec{\cancel{E}}_T = - \sum_{k=1}^N \vec{p}_{T,k}. \quad (3.3)$$

However, effects such as the limited accuracy of the particle reconstruction and pileup motivate to weight this sum:

$$\vec{E}_T = - \sum_{k=1}^N w_k \vec{p}_{T,k}. \quad (3.4)$$

The different methods of weight assignment are discussed in section 6.3.1.

## 4 Statistical Methods

Statistical methods provide the tools for comparing theoretical predictions to experimental studies. In general, it is assumed that there is a fundamental truth. This truth can never be directly observed but only estimated. The reason for that are experimental uncertainties, which can never be eliminated totally due to quantum mechanical effects, such as the observer effect or the uncertainty principle. However, there are usually experimental effects on much larger scales than these quantum effects that are for example imperfections of materials or reconstruction inefficiencies. These imperfections obscure the experimenter's view on the fundamental truth. All measurements must therefore be equipped with uncertainties, giving rise to estimated fluctuations of the measurement result.

There are two types of uncertainties: statistical and systematic uncertainties. The former reflect the precision of a measurement and can be reduced by increasing the number of measurements. The latter represent the accuracy of a measurement. Systematic uncertainties can be reduced by understanding the nature of effects which lead to inaccuracies, and taking them into account or eliminating them in a new measurement.

Section 4.1 is based on [13] and will explain fundamental concepts and define the quantities necessary to understand the statistical methods used in this thesis, which are then described in section 4.2, mainly based on [13, 73].

### 4.1 Fundamental Definitions

A hypothesis test compares a theoretical prediction to an observation. A theory may depend on a set of parameters  $\theta$ , which constrain the underlying model to a unique prediction. Therefore, the first subsection of this chapter handles one way of estimating such parameters, before the second subsection explains the basics of hypothesis tests.

#### 4.1.1 Maximum Likelihood Estimation

Fundamental to the Maximum Likelihood (ML) estimation is the concept of conditional probabilities  $P(A|B) = \frac{P(A \cap B)}{P(B)}$ , which describe the probability to observe  $A$  given a condition  $B$ . Conditional probabilities can be transformed into one another by applying Bayes' theorem

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(A)} \frac{P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}. \quad (4.1)$$

A special case of a conditional probability is the likelihood function  $L$ . It restricts  $A = \mathbf{x}$  to a set of data  $\mathbf{x}$  and  $B = \boldsymbol{\theta}$  to a set of parameters  $\boldsymbol{\theta}$ :

$$L = P(\mathbf{x}|\boldsymbol{\theta}). \quad (4.2)$$

By definition, the likelihood is maximal if the parameters are such that the probability to observe  $\mathbf{x}$  is maximal. In other words, the parameters that best fit the data can be found by maximizing the likelihood. Often, it is convenient to maximize the logarithm of the likelihood, which is equivalent since the logarithmic function is strictly increasing with the argument and probabilities are always positive except for probabilities of zero. Such probabilities are not interesting for the maximization anyway, ensuring that the logarithm is well-defined. For  $N$  independent and identically distributed measurements, the likelihood can be factorized

$$L = \prod_{i=1}^N P(\mathbf{x}_i|\boldsymbol{\theta}), \quad (4.3)$$

where  $\mathbf{x}_i$  is the outcome of the  $i$ -th measurement.

### 4.1.2 Hypothesis Test

Let  $H_0$  be the null hypothesis, and  $H_1$  an alternative hypothesis, disjunctive to the null hypothesis and  $\alpha \in (0, 1)$  the significance level (type I error). It is the aim of the hypothesis test to challenge the null hypothesis.

For this, a test statistic  $\lambda$  is introduced, which reduces the dimension of the problem to one. Furthermore, the probability density function (p.d.f.)  $f(\lambda|H)$  contains information on how likely it is to observe a range of values of the test statistic, if the hypothesis at hand is true. Conveniently, it is normed to one:

$$1 = \int_{\min \lambda}^{\max \lambda} f(\lambda|H) d\lambda. \quad (4.4)$$

The critical value of a hypothesis test is the value of the test statistic setting the limit for the rejection of the null hypothesis. It therefore depends on the significance level  $\alpha$ . For a right sided hypothesis test, which means that values of the test statistic bigger than the critical value  $\lambda^{\text{crit.}}$  lead to a rejection of the null hypothesis, the following equation holds:

$$\alpha = \int_{\lambda^{\text{crit.}}}^{\max \lambda} f(\lambda|H_0) d\lambda. \quad (4.5)$$

The  $p$ -value of an observed value of the test statistic  $\lambda^{\text{obs}}$  in a right-sided hypothesis test is defined as

$$p = \int_{\lambda^{\text{obs}}}^{\max \lambda} f(\lambda|H_0) d\lambda. \quad (4.6)$$

It describes the probability to randomly obtain a result of the test statistic at least as large as the observed one, given the null hypothesis. If  $p < \alpha$ , the null hypothesis is rejected as it is very unlikely to obtain such a result by chance.

Probabilities can be rewritten as a significance in numbers of the standard deviations  $\sigma$  of a normed Gaussian distribution  $f_g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$  with mean  $\mu$ . An  $N\sigma$  excess from the null hypothesis means that the  $p$ -value is smaller than

$$p = 1 - \int_{\mu-N\sigma}^{\mu+N\sigma} f_g(x) dx. \quad (4.7)$$



A significance of  $5\sigma$ , for example, corresponds to the probability  $p_{5\sigma} \approx 3 \times 10^{-7}$ .

If  $p \geq \alpha$ , the null hypothesis is not rejected. The probability for that having happened even though the alternative hypothesis is true is given by the type II error

$$\beta = \int_{\min \lambda}^{\lambda_c} f(\lambda|H_1) d\lambda. \quad (4.8)$$

A test statistic should be chosen such that for a given  $\alpha$ , the type II error is as small as possible. The smaller  $\beta$  for a test and a given  $\alpha$ , the more powerful is the test. If a test statistic has small values of  $\alpha$  and  $\beta$ , it means that the test statistic allows for a good discrimination between the hypotheses. Neyman and Pearson showed that given simple hypotheses  $H_0$  and  $H_1$ , the test statistic given by the likelihood ratio

$$\lambda = \frac{L(\mathbf{x}|H_0)}{L(\mathbf{x}|H_1)} \quad (4.9)$$

is most powerful [74]. A simple hypothesis means that it does not consist of different hypotheses, e.g.  $H = 0$  is a simple hypothesis but  $H \neq 0$  is not.

## 4.2 Statistical Model at the LHC

Collider physics repeats the same experiment multiple times, ideally under same conditions. The basic idea is to count events and compare these numbers to theoretical predictions. However, the theoretical predictions are biased and so are the measurements. If a certain process occurs (signal), the counting rate is different from the expectation of the model without the process (background-only). The signal+background (s+b) hypothesis is therefore tested against the background-only (b) hypothesis.

### 4.2.1 Hypotheses

In the search for new physics, the SM prediction b is usually taken as the null hypothesis, because it is the goal to reject it. The criterion for rejecting the SM is a significance level of  $\alpha \approx 3 \times 10^{-7}$ , corresponding to a  $5\sigma$  deviation from the SM.

However, since the process investigated in this thesis has a very small cross section, it is not expected to be significant and therefore, the aim is to set upper limits on the signal strength. This means that certain regions of new physics will be excluded based on the experimental observations. Thus, the null hypothesis is the s+b hypothesis and the alternative hypothesis is the b hypothesis. It is a convention to reject the new-physics hypotheses already with a significance of  $\alpha = 0.05$ .

Since the signal strength of new processes is not known, the total number of events can be described as

$$\mu s(\boldsymbol{\theta}) + b(\boldsymbol{\theta}), \quad (4.10)$$

where  $\mu$  is called signal strength modifier, a quantity that allows for fit deviations from the signal model prediction  $s$ ,  $b$  is the prediction of the background model, and  $\boldsymbol{\theta}$  are the nuisance parameters, which account for systematic uncertainties.

### 4.2.2 Systematic Uncertainties

With the help of auxiliary measurements, the nuisance parameters can be further constrained. This is done by introducing an additional factor  $p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$ , which is the probability to obtain the nuisance parameter values  $\tilde{\boldsymbol{\theta}}$ , known prior to the measurement, given  $\boldsymbol{\theta}$  in a measurement. In general, there are two types of systematic uncertainties. Uncertainties of the first

type are rate uncertainties and affect the yield of distributions. The factor is given by a log-normal distribution

$$p(\tilde{\boldsymbol{\theta}}_k|\boldsymbol{\theta}_k) = \frac{1}{\sqrt{2\pi} \ln(\kappa)} \exp\left(-\frac{\ln^2(\boldsymbol{\theta}_k/\tilde{\boldsymbol{\theta}}_k)}{2 \ln^2(\kappa)}\right) \frac{1}{\boldsymbol{\theta}_k}, \quad (4.11)$$

where  $\kappa$  resembles the distribution's width and  $k$  is the index of a specific uncertainty. For large values of  $\kappa$ , a varied normalization is more compatible to the prior knowledge of this uncertainty.

The second type of uncertainties affects the shape of the distributions and is modeled with

$$p(\tilde{\boldsymbol{\theta}}_k|\boldsymbol{\theta}_k) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^2}{2}\right), \quad (4.12)$$

which is a Gaussian distribution with standard deviation  $\sigma = 1$ .

The shapes are approximated for every value of  $\theta$  within one standard deviation around  $\tilde{\boldsymbol{\theta}}_k$ . This is done by evaluating the prediction for one standard deviation up and down and interpolating between predictions for the remaining values  $0 < |\boldsymbol{\theta}_k| < 1$  with a spline function and with a linear function for  $|\boldsymbol{\theta}_k| > 1$ . The spline is defined as

$$\alpha(\theta) = \frac{1}{2} \left( (\delta^+ - \delta^-) \theta + \frac{1}{8} (\delta^+ \delta^-) (3\theta^6 - 10\theta^4 + 15\theta^2) \right), \quad (4.13)$$

with  $\delta^\pm$  defined as the difference between the  $\pm 1\sigma$  variations and the nominal distribution in each bin.

#### 4.2.3 Likelihood Function and Test Statistic

For a Poisson distributed binned quantity, which is modeled as in equation 4.10, the likelihood function is

$$L(\text{data}|\mu, \boldsymbol{\theta}) = \left[ \prod_{i=1}^N \frac{(\mu s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}))^{n_i}}{n_i!} \exp(-\mu s_i(\boldsymbol{\theta}) - b_i(\boldsymbol{\theta})) \right] \left[ \prod_{k=1}^{\dim(\boldsymbol{\theta})} p(\tilde{\boldsymbol{\theta}}_k|\boldsymbol{\theta}_k) \right], \quad (4.14)$$

where the lower indices of  $\boldsymbol{\theta}$  and  $\tilde{\boldsymbol{\theta}}$  in latin letters describe the elements of the nuisance vector and  $\dim(\boldsymbol{\theta})$  is the number of nuisance parameters. The data is either given by observation or as in this thesis by pseudodata, while  $n_i$  is the number of entries in bin  $i$  and  $N$  is the number of bins.

The profile likelihood ratio can be used as a powerful test statistic:

$$\tilde{q}_\mu = -2 \ln \frac{L(\text{data}|\mu, \hat{\boldsymbol{\theta}}_\mu)}{L(\text{data}|\hat{\mu}, \hat{\boldsymbol{\theta}}_{\hat{\mu}})}, \quad 0 \leq \hat{\mu} \leq \mu, \quad (4.15)$$

where  $\hat{\mu}$  and  $\hat{\boldsymbol{\theta}}_{\hat{\mu}}$  are the global ML estimators and  $\hat{\boldsymbol{\theta}}_\mu$  is the set of parameters which maximizes the likelihood given a previously fixed value of  $\mu$ , which represents the signal hypothesis. If  $\hat{\mu} > \mu$ , there would be no contradiction with the signal hypothesis, therefore only values between  $\hat{\mu} = \mu$  and the background hypothesis  $\hat{\mu} = 0$  are considered. Since the denominator is the highest possible value of the likelihood, the test statistic's values are at least zero. The main advantage of using this profile likelihood ratio is that Wilks' theorem is applicable [75]. This implies that the test statistic p.d.f. can be given by an asymptotic analytical formula, which saves computing resources.

The observed value  $\tilde{q}_\mu^{\text{obs}}$  is then calculated by maximizing equation 4.14 with respect to the fixed  $\mu$ . In this process, the nuisance parameters  $\hat{\boldsymbol{\theta}}_\mu^{\text{obs}}$  are estimated using equation 4.14.

These as well as the nuisance parameters of the null hypothesis  $\hat{\theta}_0^{\text{obs}}$  are needed for the p.d.f. construction in a later step. The value  $\tilde{q}_\mu^{\text{obs}}$  depends on the true value of the signal strength modifier  $\mu^t$ , which is not known but hidden in the data in a real analysis. There are two cases:

- $\mu \geq \mu^t$ : As  $\hat{\mu}$  is optimized to best match the data, it should be close to  $\mu^t$ . The smaller the difference between  $\mu$  and  $\mu^t$ , the more likely it is to obtain the observation given the s+b hypothesis and therefore the smaller  $\tilde{q}_\mu$ .
- $\mu < \mu^t$ : Again,  $\hat{\mu}$  should be close to  $\mu^t$ . As  $\mu$  is the upper limit for  $\hat{\mu}$ , it should be close to this value, regardless of the choice of  $\mu$ . Therefore, the values of the test statistic should always be close to zero in this case.

Therefore, values of  $\mu$  which exceed the true value of  $\mu^t$  are indicated by high observed values of the test statistic, which makes the test statistic suitable for excluding such values.

#### 4.2.4 $p$ -Values and Limit Setting Procedure

In order to construct  $p$ -values, p.d.f.s of the test statistic  $f(\tilde{q}_\mu|\mu, \hat{\theta}_\mu^{\text{obs}})$  and  $f(\tilde{q}_\mu|0, \hat{\theta}_0^{\text{obs}})$  are calculated for the s+b and b hypothesis, respectively. In principle, this can be done using toy pseudodata generated with MC methods. Since these simulations are challenging in terms of computing resources, it is more convenient to use approximative methods like the one described in [76], for which the application of Wilks' theorem is needed.

The shapes of these p.d.f.s are discussed qualitatively in the following.

Even though a certain value of  $\mu^{\text{sim}} = \begin{cases} \mu & , \text{ s+b} \\ 0 & , \text{ b} \end{cases}$  is used for the toy simulation, there might still be fluctuations of the number of background or signal events generated in a toy pseudo experiment leading to a  $\hat{\mu}$  different from  $\mu^{\text{sim}}$  in the global maximum likelihood. The test statistic is now calculated as

$$\tilde{q}_\mu = -2 \ln \frac{L(\text{toys}_{\text{s+b,b}}|\mu, \hat{\theta}_\mu)}{L(\text{toys}_{\text{s+b,b}}|\hat{\mu}, \hat{\theta}_{\hat{\mu}})}, \quad 0 \leq \hat{\mu} \leq \mu, \quad (4.16)$$

where  $\text{toys}_{\text{s+b,b}}$  denotes the generated toy pseudo experiments for the s+b or b hypothesis. Then:

- $f(\tilde{q}_\mu|\mu, \hat{\theta}_\mu^{\text{obs}})$  is the generated p.d.f. for getting a certain  $\tilde{q}_\mu$  under the s+b hypothesis  $\mu^{\text{sim}} = \mu$ . The globally fitted  $\hat{\mu}$  should be around  $\mu^{\text{sim}}$ , as this was used in the data set generation. The ML value of the numerator of equation 4.16 should therefore be in the same order of magnitude as the denominator, leading to a distribution shifted towards  $\tilde{q}_\mu = 0$ . For other values of  $\mu$  (and  $\mu^{\text{sim}} = \mu$ ), the p.d.f. should not change dramatically, as  $\hat{\mu}$  always is around the same value as  $\mu$ .
- $f(\tilde{q}_\mu|0, \hat{\theta}_0^{\text{obs}})$  is the generated p.d.f. for getting a certain  $\tilde{q}_\mu$  under the b hypothesis  $\mu^{\text{sim}} = 0$ . The globally fitted  $\hat{\mu}$  should be around 0, as this was used in the data set generation. Now, the ML value of the numerator of equation 4.16 differs from the one in the denominator, because the toys do not contain any signal and could only be described by a down-fluctuation of the s+b hypothesis. The p.d.f. will therefore be shifted towards higher values of  $\tilde{q}_\mu$ . The larger  $\mu$ , the larger the shift.

Summarizing these remarks and the ones on the observed value of the test statistic, by increasing  $\mu$  from zero on, the p.d.f. of the s+b hypothesis does not change dramatically. The p.d.f. of the b hypothesis on the other hand is shifted towards higher values of the test statistic with increasing  $\mu$ . As soon as  $\mu$  exceeds the true (but unknown) value of

the signal strength, the observed value of the test statistic starts increasing starting from around zero. Therefore, by construction, the  $p$ -value

$$p_\mu = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\text{obs}} | \text{s+b}) = \int_{\tilde{q}_\mu^{\text{obs}}}^{\infty} f(\tilde{q}_\mu | \mu, \hat{\theta}_\mu^{\text{obs}}) d\tilde{q}_\mu \quad (4.17)$$

decreases as soon as  $\mu$  exceeds the true value. This is an important finding, as it stresses the ability of the test statistic to rule out regions of the signal strength modifier, which exceed the true value.

However, it is possible that the null hypothesis is true but there are downward fluctuations of the background or signal, which mislead into rejecting the s+b hypothesis. In this case, the null hypothesis could be falsely rejected, which is why there is the rather conservative  $\text{CL}_s$  approach [77]. The idea is to also include the behavior of the b hypothesis. For that, another  $p$ -value is defined for the b hypothesis:

$$p_b = 1 - P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\text{obs}} | \text{b}) = 1 - \int_{\tilde{q}_\mu^{\text{obs}}}^{\infty} f(\tilde{q}_\mu | 0, \hat{\theta}_0^{\text{obs}}) d\tilde{q}_\mu. \quad (4.18)$$

$p_b$  refers to the probability to observe a value of the test statistic which is not more extreme than the observed value, given the b hypothesis. This is now used to obtain a confidence level for the signal hypothesis:

$$\text{CL}_s(\mu) \equiv \frac{p_\mu}{1 - p_b}. \quad (4.19)$$

An exclusion of the signal hypothesis is then given, if  $\text{CL}_s \leq 0.05$ . In this case, it is at least 20 times more likely to measure a value of the test statistic at least as extreme as the observed value, given the b hypothesis, than measuring a corresponding value under the s+b hypothesis.

If for example the ground truth is that there is a signal  $\mu^t$  but there is a downward fluctuation of the background leading to a smaller  $\hat{\mu}$ , the observed test statistic value may be high enough to exclude the signal at  $\mu^t$  if only rejecting based on  $p_\mu$ . The  $\text{CL}_s$  method accounts for this situation by considering the probability of such downward fluctuations, which is hidden within  $1 - p_b$ . Nevertheless, this conservative approach comes at the cost of making it hard to set strong limits.

The last step in the limit setting procedure is to vary  $\mu$  up to a value, where  $\text{CL}_s(\mu) = 0.05$ . All values above this threshold are excluded with at least a 95%  $\text{CL}_s$  confidence level.

Instead of using real observations and calculating real limits, expected limits can be obtained by using simulated data and treating them as observations. These expected limits indicate how sensitive the analysis can become based on the influence of systematic effects. As the information of the simulated data is known, the expected analysis not only provides a tool for estimating the sensitivity before doing the analysis, but can also be used as a consistency check, whether the value of  $\mu$  used in the simulation can be reproduced by the fits.

In order to obtain an estimate of the expected sensitivity, asymptotic test statistic distributions are analyzed, resulting in an asymptotic expected upper limit for the analysis. For this, asymptotic formulas are used to construct the p.d.f. and obtain  $\mu$ . Then, using an Asimov data set<sup>1</sup>, uncertainties on this  $\mu$  are approximated. For more details see [76].

---

<sup>1</sup>An Asimov data set is constructed such that its analysis reproduces exactly the true values for all parameters.

## 5 Neural Networks

The theory of Neural Networks (NNs) was developed decades ago, taking its first steps in the work of McCulloch and Pitts [78]. However, it is only in recent years that the capacity and computational power of modern computers have become large enough to use NNs in various fields.

After an introduction to the general concept of NNs in section 5.1, sections 5.2 and 5.3 explain two special types of NNs, feedforward networks and Graph Neural Networks (GNNs), which are used in GraphMET.

### 5.1 Fundamentals

NNs have a most popular archetype, the human brain. On an abstract level, the human body measures signals and sends these as electric impulses via nerves to the brain. Then the brain processes these impulses and thereby transforms the input signal into higher level information. A prime example is the process of reading: The eyes measure an optical signal and turn it into electric signals, which are then processed by the brain and thereby get a meaning.

This concept is also applied when using NNs. NNs are models which transform one kind of data into another. Usually they have a vast number of parameters (e.g. in GraphMET more than 6000) that are adapted in a process called training. The high number of parameters leads to quite complex models, whose decisions can be hard to understand. However, this complexity is also a reason why NNs are interesting for applications in particle physics: Different kinds of information can be combined and used even though the analytic connection between the input and output is not known.

In order to understand the principle of NNs, one has to take a closer look at their structure, which is also inspired by the brain.

#### 5.1.1 Artificial Neurons

The smallest unit of NNs is an (artificial) neuron. A single neuron is shown in figure 5.1.

The  $d$  scalar input variables  $x_1, \dots, x_d$  are combined by the transmission function

$$\Sigma_{b,w_1,\dots,w_d}(x_1, \dots, x_d), \tag{5.1}$$

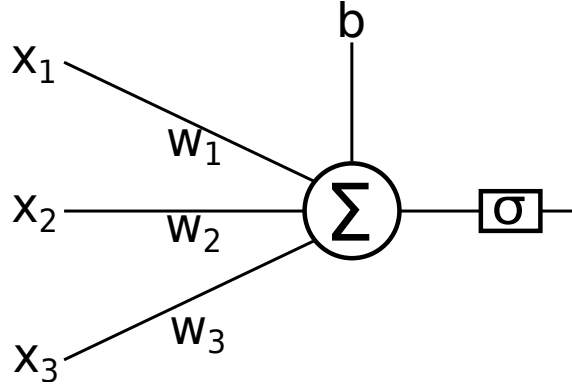


Figure 5.1: Neuron with three input variables  $x_1, x_2, x_3$ , three weights  $w_1, w_2, w_3$ , one bias  $b$  and an activation function  $\sigma$ .

where  $w_1, \dots, w_d$  are called weights and  $b$  is a bias. Usually the transmission function just takes the weighted sum of the input variables:

$$\Sigma_{b, w_1, \dots, w_d}(x_1, \dots, x_d) = \sum_{k=1}^d w_k x_k + b. \quad (5.2)$$

For a more handy notation, the input variables are written in a vector  $\vec{x} = \begin{pmatrix} x_1 \\ \dots \\ x_d \end{pmatrix}$  and so

are the weights  $\vec{w} = \begin{pmatrix} w_1 \\ \dots \\ w_d \end{pmatrix}$ . Now, the transmission function can be written as a scalar product:

$$\Sigma_{b, \vec{w}}(\vec{x}) = \vec{w} \cdot \vec{x} + b. \quad (5.3)$$

The value of the transmission function is used as argument of the activation function  $\sigma$ , which calculates the neuron's output. In general, there are no restrictions for  $\sigma$ , but usually, it is chosen to be non-linear because linear combinations of linear functions are again linear in the argument. Therefore, NNs with linear activation functions cannot properly describe non-linear behavior.

### 5.1.2 Neural Networks

NNs are combinations of several neurons, where the output of a neuron can again be input of itself or other neurons. As described before, NNs transform one kind of data into another. The transformation function

$$\Phi_{\mathcal{W}} : \mathbb{R}^m \rightarrow \mathbb{R}^n, \vec{x} \mapsto \vec{y} = \Phi_{\mathcal{W}}(\vec{x}) \quad (5.4)$$

depends on the type of Neural Network. For now, it shall not be specified further than being non-linear and transforming an  $m$  dimensional set of input variables into an  $n$  dimensional set of output variables, called prediction. The weights  $\mathcal{W}$  are determined during training. After the training is finished, the weights are considered as fixed parameters for the application of the Neural Network.

### 5.1.3 Training

During training, the weights are adjusted to optimize the prediction using methods of supervised learning. For that, a data set with known true output values  $\vec{y}_t$  is needed. This data set is then gradually processed by the network and the output value is compared to the true value. A handy way of doing so is using a loss function  $L_{\mathcal{W}}(\vec{x})$ , which typically is the norm of the difference between prediction  $\Phi_{\mathcal{W}}(\vec{x})$  and true value  $\vec{y}_t$ :

$$L_{\mathcal{W}}(\vec{x}, \vec{y}_t) = \|\Phi_{\mathcal{W}}(\vec{x}) - \vec{y}_t\|_l, \quad (5.5)$$

where  $\|\cdot\|_l$  is a mathematical norm, i.e. for  $l = 2$  the typical Euclidean norm  $\|\vec{y}\|_2 = \sqrt{\vec{y} \cdot \vec{y}}$ .

It should be mentioned that the number of weights increases vastly with the Neural Network's size. Therefore, the loss function is usually optimized using numerical methods, e.g. gradient descent [79].

One frequently used gradient descent method is the backpropagation algorithm [80]. In this case, the gradient of the loss function with respect to the weights is calculated. For this purpose, the loss function is interpreted as a function of the weights with fixed input variables as parameters. The weights are then changed into the direction of the steepest descent. This algorithm starts with optimizing the weights of the output neurons and continues up to the input neurons, hence the name.

There are two main conditions that should be avoided during the training process: overfitting and being "trapped" in a local minimum.

The first one describes a situation in which NNs get adapted to the data rather than its underlying nature. When dealing with new data, an overfitted model is therefore more likely to have a higher value of the loss function than a correct model. This can happen, because a Neural Network's number of degrees of freedom usually exceed the number of the problem's degrees of freedom. Overfitting can be observed if the training data set is divided into two parts. One of them is then used for training, the other one for validation. If the weights are now adapted too much to the training data, the loss, calculated with the validation set, increases and exceeds the loss in the training. A common prevention method is to install random errors such as dead Neurons. This method is called dropout and is controlled via the dropout rate. Other ways to deal with this condition are to use a bigger and less biased training set or to stop the training process as soon as overfitting occurs.

The second type of condition means that the loss function is in a local minimum which is not the global minimum. If the gradient with respect to the weights is calculated in the close neighborhood of the current weights, they will not be adjusted anymore as there is no smaller loss function value nearby. In order to avoid this issue, one can vary the radius of the neighborhood or the learning rate, which determines how much the weights are changed per iteration.

## 5.2 Feedforward Neural Networks

A feedforward network is an aggregation of multiple parallel neurons in  $K$  consecutive layers. The first layer is called input layer, its input is given from outside, and the last layer is named output layer. If  $K > 2$ , the layers between the input and output layer are called hidden layers. The output is then passed to the next layer. The last layer's output is the Neural Network's prediction. Thus, this type of network passes information to the next layers in forward direction without exception, hence the name of this network type.

Let now  $d$  be the dimension of the input and  $d'$  the number of neurons in a certain layer. For neuron number  $n \in \{1, \dots, d'\}$  of this layer, the input vector  $\vec{x} \in \mathbb{R}^d$  gets scalar multiplied with the neuron's weight vector  $\vec{w}_n \in \mathbb{R}^d$ . An example can be seen in figure 5.2.

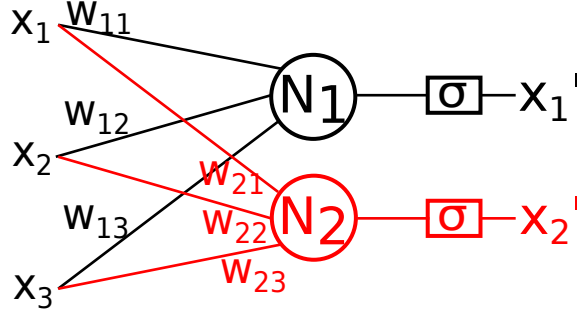


Figure 5.2: Feedforward network with  $d = 3$  input variables,  $d' = 2$  neurons in  $K = 1$  layer.

In order to describe this in one line, one can introduce the weight matrix  $W$  which has the weight vectors of the layer's neurons as columns. The transposed output vector is then given as the matrix product of the transposed input vector and the weight matrix:

$$\begin{aligned} \vec{x}'^T &= (x'_1 \quad \dots \quad x'_{d'}) = \vec{x}^T W = (x_1 \quad \dots \quad x_d) (\vec{w}_1 \quad \dots \quad \vec{w}_{d'}) \\ &= (\vec{x} \cdot \vec{w}_1 \quad \dots \quad \vec{x} \cdot \vec{w}_{d'}) \in \mathbb{R}^{d'}. \end{aligned} \quad (5.6)$$

The output of the  $d'$  different neurons can now be found in the entries of  $\vec{x}'$ .

The number of neurons in the output layer should equal the dimension of the problem. For binary decisions, e.g. 'does this event contain a neutrino?', the output layer has only one neuron. However, the more general case is a multivariate decision, e.g. 'does this event contain no neutrino, one neutrino or more neutrinos?', where every subquestion is answered by a dedicated output layer neuron. These classification tasks can be generalized even more to infinity many continuous classes. This approach is called approximation and is used in this thesis. A question to this approximation approach could be 'what is the probability for a particle to be from pileup?'.

## 5.3 Graph Neural Networks

GNNs use the structure of graphs to propagate information not only in the forward direction but also in a parallel manner. This is an import difference to usual feedforward networks that pass information only in the forward direction.

### 5.3.1 Graphs

A graph  $G = (V, E)$  is a pair, where  $V$  is a set and  $E$  is a set of two-sets. This mathematical definition can be illustrated by a set of nodes or vertices  $V$  with relations  $E$  depicted as edges. Because  $E$  is made of two-sets, each edge can only relate two vertices. The number of edges per vertex is not limited whatsoever. Edges can be directed, which means that the order of the elements of  $E$  plays is not negligible. In this case the whole graph is called directed.

If two vertices share a common edge, they are called adjacent. The adjacency matrix  $A \in \mathbb{R}^{V \times V}$  describes the structure of a graph. For every edge between two vertices  $i$  and  $j$ , there is a one in the adjacency matrix at position  $ij$  while entries representing two unconnected vertices contain a zero.

Graphs are very generic structures. This is why there are lots of examples for graphs, ranging from social graphs connecting people with each other to molecules or Feynman diagrams. More details on graphs can be found e.g. in [81].



### 5.3.2 GNN Process

This subsection is based on [82, 83] and presents a general and rather abstract approach to GNNs, supported by an application-oriented explanation in the following subsection.

In the context of GNNs, it is convenient to extend the definition of graphs as follows: A graph is a tuple  $G = (V, E, S, R, \mathbf{u})$ , where  $V$  and  $E$  are lists of vectors, describing feature information of all the vertices  $\mathbf{v}_i$  and edges  $\mathbf{e}_j$ .  $S$  and  $R$  are lists of indices describing the position of the edges and  $\mathbf{u}$  is a vector that describes global features. The dimension of  $V$  is the number of vertices and the dimension of  $E$  is the number of edges. If there is an undirected edge between nodes  $i$  and  $j$  as well as a directed edge from node  $j$  to node  $k$ , the senders and receivers list would be

$$S = (i, j, j), \quad R = (j, i, k), \quad (5.7)$$

where equal indices of  $S$  and  $R$  describe an edge.

The goal of GNNs is, for each node in the graph, to learn not only from its own input but also from information hidden within its neighborhood. For this, GNNs use NNs that operate on a graph structure. During this learning process, the structure of the graph should not change. For this reason,  $S$  and  $R$  have to remain constant as well as the dimensions of  $V$  and  $E$ . The elements of  $E$  and  $V$  as well as the global features  $\mathbf{u}$ , however, may and should change during the learning process. The updates depend on the objects in the neighborhood:

$$\mathbf{e}'_l = \phi^e(\mathbf{e}_l, \mathbf{v}_{R_l}, \mathbf{v}_{S_l}, \mathbf{u}), \quad \bar{\mathbf{e}}'_m = \rho^{e \rightarrow v}(E'_m), \quad (5.8)$$

$$\mathbf{v}'_m = \phi^v(\mathbf{v}_m, \bar{\mathbf{e}}'_m, \mathbf{u}), \quad \bar{\mathbf{e}}' = \rho^{e \rightarrow u}(E'), \quad (5.9)$$

$$\mathbf{u}' = \phi^u(\bar{\mathbf{e}}', \bar{\mathbf{v}}', \mathbf{u}), \quad \bar{\mathbf{v}}' = \rho^{v \rightarrow u}(V'). \quad (5.10)$$

The update functions  $\phi$  are usually given by trainable NNs and compute the new feature vectors. They have a fixed size input and output. In order for the nodes and edges to learn something about their neighborhood, the information from the variable sized neighborhood needs to be combined into a fixed size by the aggregation functions  $\rho$ . The functions  $\rho$  have to be permutation invariant as the new vector should not depend on the numbering of the nodes but only on the graph structure.

In the first line of equation 5.10, a new edge feature vector is computed by a Neural Network depending on the previous feature vector of this vertex, the feature vectors of the vertices connected by the edge and the global feature vector. For each node, the feature vectors of the set  $E'_i$  of all incoming edges to this vertex are aggregated. In the next line of equation 5.10, the new vertex feature vector is computed by a Neural Network depending on its own previous feature vector, its aggregated edge vector and the global features. All edge and vertex feature vectors are then aggregated separately and act as input to the calculation of the global feature vector, which also depends on its previous value.

Depending on the choice of  $\phi$  and  $\rho$ , one can create different kinds of GNNs and thereby address different problems. Another important aspect is the adjacency matrix as it describes which nodes are connected and thus in each other's neighborhood. In general, GNNs consist of multiple layers, where each layer can be designed individually, if the dimensions of input and output vectors are considered.

In this thesis, two different types of graphs are used for MET estimation: A fixed graph on the one hand does not change its adjacency matrix throughout the different graph layers. This requires a global set of quantities describing node features that are the same for every layer. The Euclidean distance between the global input variables fulfills this criterion.

On the other hand there are dynamic graphs. These can have different adjacency matrices for different graph layers. In this case, the structure of the graph changes between at least two layers. To obtain a dynamic adjacency matrix, one has to construct the adjacency matrix using variables which change between at least two layers. The Euclidean distance between the input variables of the specific layer fulfills this criterion for example.

For a better understanding, one type of Graph Neural Network, which is often used in this thesis, will now be described in detail.

### Graph Convolutional Networks

The following is a summary of [84], supplemented by some explanations. This type of Graph Neural Network is obtained by placing  $N$  similar layers of neurons parallel to each other. Similar means, that the number of input variables  $d$  and neurons  $d'$  should be the same for all parallel layers. One layer of neurons is then denoted as one node, such that  $N$  is also the number of nodes.

The input values can now be described as a matrix  $X$  which has the input vectors of the different nodes as rows. This input matrix is then of dimension  $N \times d$ .

Let now  $W$  be the weight matrix from section 5.2. Then the output matrix  $X' \in \mathbb{R}^{N \times d'}$  of the  $N$  layers can be calculated in one line, that is:

$$X' = \begin{pmatrix} \vec{x}_1^T \\ \dots \\ \vec{x}_N^T \end{pmatrix} = XW = \begin{pmatrix} \vec{x}_1^T \\ \dots \\ \vec{x}_N^T \end{pmatrix} \begin{pmatrix} \vec{w}_1 & \dots & \vec{w}_{d'} \end{pmatrix} \quad (5.11)$$

$$= \begin{pmatrix} \vec{x}_1 \cdot \vec{w}_1 & \dots & \vec{x}_1 \cdot \vec{w}_{d'} \\ \dots & \dots & \dots \\ \vec{x}_N \cdot \vec{w}_1 & \dots & \vec{x}_N \cdot \vec{w}_{d'} \end{pmatrix}. \quad (5.12)$$

The output vector of one node can be found in the corresponding row and has dimension  $d'$  as each node has  $d'$  neurons.

Up to now, the output vector of one node solely depends on its input vector and the weights. In order to introduce mutual influence between the nodes, the adjacency matrix  $A \in \mathbb{R}^{N \times N}$  is applied. In the case of a Graph Convolutional (GCNConv) layer, the adjacency matrix is weighted by a matrix  $\tilde{D}$  to account for the different number of edges between the nodes. If there were no weights, the nodes with more edges would have a higher output than the nodes with fewer edges by construction. The weighted adjacency matrix is

$$A' = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}, \quad \text{with} \quad (5.13)$$

$$\tilde{A} = A + I_N \quad \text{and} \quad \tilde{D}_{ii} = \sum_j \tilde{A}_{ij}, \quad (5.14)$$

where  $A$  is the adjacency matrix without self-connections and  $I_N$  is the  $N$  dimensional identity matrix. The more edges a node has, the lower the weight of these edges. It is important to note that the weights in this case are not trainable, but are assigned strictly according to a rule that depends only on the number of connected nodes.

Considering an activation function  $\sigma$ , the total update of one layer is given by

$$X' = \sigma(A'XW). \quad (5.15)$$

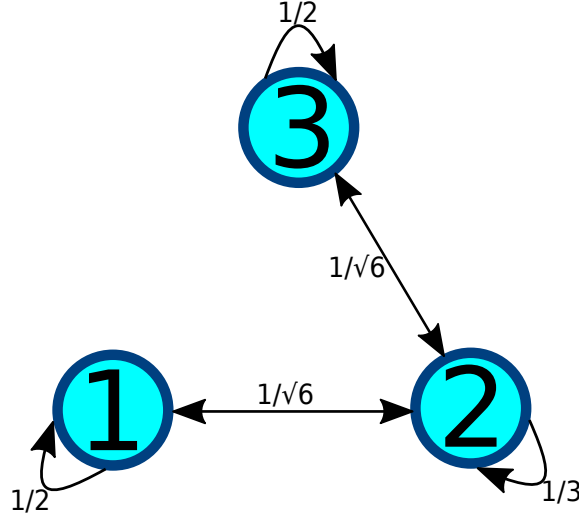


Figure 5.3: Representation of a Graph Neural Network with  $N = 3$  neurons without connection between node one and three. Therefore, the entries  $A_{13}$  and  $A_{31}$  of the adjacency matrix are zero. The weights are obtained by calculating  $\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$ . They are small if a node has many edges, which can be seen in the self connection of node two, which has three connections and therefore a weight of  $\frac{1}{3}$  compared to the other two nodes both having two connections and a self-weight of  $\frac{1}{2}$ .

#### Application: Graph Convolutional Network with Three Neurons

Figure 5.3 shows a simple Graph Neural Network with the following adjacency and weighted adjacency matrix:

$$\tilde{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad A' = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{3} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{6}} & \frac{1}{2} \end{pmatrix}. \quad (5.16)$$

Given a  $d$  dimensional input vector for each node, combined in the input matrix  $X \in \mathbb{R}^{3 \times d}$ , and the  $d'$  dimensional weight vectors of the  $d'$  neurons in each node, combined in the weight matrix  $W \in \mathbb{R}^{d \times d'}$ , the resulting output matrix  $X' \in \mathbb{R}^{3 \times d'}$  is given by

$$X' = \begin{pmatrix} \vec{x}_1'^T \\ \vec{x}_2'^T \\ \vec{x}_3'^T \end{pmatrix} = \sigma(A' X W) \quad (5.17)$$

$$= \sigma \left( \begin{pmatrix} (\frac{1}{2}\vec{x}_1 + \frac{1}{\sqrt{6}}\vec{x}_2) \cdot \vec{w}_1 & \dots & (\frac{1}{2}\vec{x}_1 + \frac{1}{\sqrt{6}}\vec{x}_2) \cdot \vec{w}_{d'} \\ (\frac{1}{\sqrt{6}}\vec{x}_1 + \frac{1}{3}\vec{x}_2 + \frac{1}{\sqrt{6}}\vec{x}_3) \cdot \vec{w}_1 & \dots & (\frac{1}{\sqrt{6}}\vec{x}_1 + \frac{1}{3}\vec{x}_2 + \frac{1}{\sqrt{6}}\vec{x}_3) \cdot \vec{w}_{d'} \\ (\frac{1}{\sqrt{6}}\vec{x}_2 + \frac{1}{2}\vec{x}_3) \cdot \vec{w}_1 & \dots & (\frac{1}{\sqrt{6}}\vec{x}_2 + \frac{1}{2}\vec{x}_3) \cdot \vec{w}_{d'} \end{pmatrix} \right). \quad (5.18)$$

This example shows the difference between usual NNs and GNNs: The output vectors of each node now contain not only their own input vectors but also the input vectors of the nodes which they are connected to. For example the first row of equation 5.18 shows the output vector of node one. The new vector consist of the weighted sum of the input vectors of node one and two, scalar multiplied with the weight vectors of the different neurons within the nodes.

Now that an understanding of GNNs has been developed and the tools for working with GNNs are given, the next chapter introduces a more complex use case of GNNs - GraphMET.



## 6 GraphMET

GraphMET (also referred to as DeepMETv2) uses GNNs and aims to improve MET estimation for the CMS experiment. Section 6.1 focuses on the general setup and training of the GraphMET model before different GraphMET architectures are compared in section 6.2. Finally, the best GraphMET architecture is evaluated on a new data set and compared with previous MET estimation methods in section 6.3.

### 6.1 GraphMET Setup and Training

Following the discussion in section 3.3 the general MET estimation formula is

$$\vec{E}_T = - \sum_{k=1}^N w_k \vec{p}_{T,k}, \quad (6.1)$$

where  $N$  is the number of reconstructed particles,  $\vec{p}_{T,k}$  is the transverse momentum of a reconstructed particle  $k$ , and  $w_k$  is the corresponding weight determined by GraphMET.

#### 6.1.1 Input Features

A total of seven continuous and three discrete variables are used as input features. Their distributions can be found in appendix A. The following list is taken from [85, 86].

##### Continuous Variables

- $dz$ : Distance between leading vertex and PF candidate in  $z$  direction (along the beam axis); neutral PF candidates are assigned a value of zero
- $dxy$ : Impact parameter in the  $xy$ -plane (transverse to the beam axis); neutral PF candidates are assigned a value of zero
- $\eta$ : Pseudorapidity
- $m$ : Mass of the PF candidate
- $p_x, p_y$ : Components of the transverse momentum in  $x$  and  $y$  direction; contain information on the azimuthal angle  $\phi$
- $p_T$ : Transverse momentum of the PF candidate

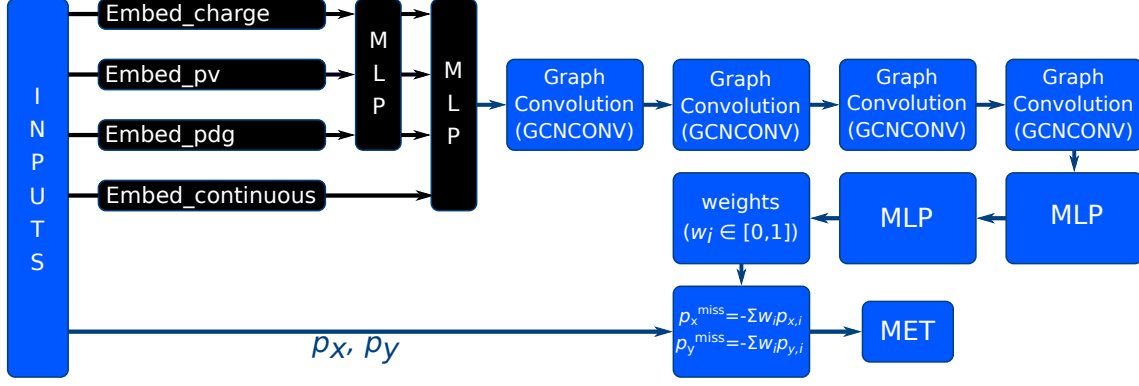


Figure 6.1: Architecture of the GraphMET baseline model. The input variables are embedded with the black part and given into four consecutive graph convolutional layers. The output of the graph layers is then processed by two feedforward networks (multi-layer perceptrons, MLP) and results in the weight, which is used in the MET estimation formula.

### Categorical Variables

- **pdgID**: Particle identification number for PF candidates; can take the values 1 (hadronic PF candidates in the HF), 2 (electromagnetic PF candidates in the HF), 11 (PF electrons), 13 (PF muons), 22 (PF photons), 130 (neutral PF hadrons, e.g.  $K_L^0$ , a CP odd state of the neutral Kaon) and 211 (charged PF hadrons, e.g. pions)
- **electric charge in units of  $e$** : Can take the values  $0, \pm 1$
- **fromPV**: Contains tightness information of the association of the PF candidate to the leading vertex; can take values  $\{0, 1, 2, 3\}$ , where 0 is for PF candidates, which are associated to another primary vertex and 3 is the closest to the leading vertex; neutral PF candidates are assigned a value of 3

#### 6.1.2 General Architecture

In first tests, a certain architecture proved to perform well while still being reasonable in terms of computing resources. This architecture is used as a baseline architecture which other architectures will be compared against in section 6.2.

Each PF candidate's 10 input features are mapped by GraphMET onto a weight, considering also the information of neighboring PF candidates. This information mixing is obtained by applying four graph convolutional layers, where each PF candidate is resembled by a node. Two nodes are connected if the distance in the  $\eta - \phi$  plane between the PF candidates is  $\Delta R \leq 0.4$ . Each node has 32 neurons, meaning that the input information is mapped on a 32-dimensional output vector. After the last graph layer, each PF candidate has its own 32-dimensional feature vector, which is then processed separately by two feedforward networks into a 16- and then one-dimensional output, the weight. The described model is shown as the blue part in figure 6.1. The black part is a technique to improve the handling of discrete values with NNs. It will be explained further in subsection 6.2.3.

#### 6.1.3 GraphMET Training

The different models were trained under similar conditions, i.e. using **PyTorch** [87] with the additional library **PyTorch Geometric** [88] on a 50% share of the GeForce GTX 1080 Ti GPU. The training data set consists of 420 000  $Z \rightarrow \nu\nu$  events with additional jets, produced with **MadGraph5** and **PYTHIA8**. During generation, the events are simulated in

Table 6.1: Investigated GraphMET models. The baseline model is given in bold letters.

degree of freedom	investigated options
graph style	<b>fixed</b> , dynamic
update method	<b>GCNConv</b> , EdgeConv
Embedding	<b>yes</b> /no
activation function	<b>ELU</b> , ReLU, sigmoid, tanh
$\Delta R \leq$	0.2, <b>0.4</b> , 0.8, 1.5
max number of neighbors	0, 100, 250, <b>500</b>
number of graph layers	1, 2, 3, <b>4</b> , 5, 6
dimension of feature vectors	8, 16, <b>32</b> , 64

complementary phase spaces that are given by different ranges of the scalar sum of all jet transverse momenta

$$H_T = \sum_i p_{T,i}^{\text{jet}}. \quad (6.2)$$

For each of the  $H_T$  regions 100 – 200 GeV, 200 – 400 GeV, 400 – 600 GeV, 600 – 800 GeV, 800 – 1200 GeV, 1200 – 2500 GeV and  $\geq 2500$  GeV, a total of 60 000 events are considered in the training. This approach enables GraphMET to learn from each  $H_T$  region with a similar number of events, stabilizing performance over the entire  $H_T$  range.

The total set is randomly split into two samples, one for training with 80% of the events and the other for validation with the remaining 20%. The models are then trained for 30 epochs and the weights of the epoch with the smallest value of the loss function are considered in the comparison. The loss function is defined as

$$L = \frac{1}{B} \sum_b \frac{1}{2} |\vec{\not{E}}_{T,b} - \vec{\not{E}}_{T,b}^t|^2, \quad (6.3)$$

where  $B$  is the batch size and  $b$  refers to a certain event within the batch. The loss function describes the mean value of the squared difference vector between true and estimated MET within a batch of parallel evaluated events.

## 6.2 Comparison Between Different GraphMET Architectures

Starting from the previously introduced baseline model, different parameters can be changed in order to find the best performing network.

Table 6.1 provides an overview of the different options compared in this section. As the number of different GraphMET configurations is 12 288 when considering every combination of parameters, only one parameter change is investigated at a time, resulting in a handy number of 28 comparisons. While the benefit of this approach are saved time and computing resources, it comes at costs of missing possible second order effects, for instance if changing two parameters at once leads to improvements. The baseline model (bold in table 6.1) was chosen due to its good performance during first tests and found out to outperform the other investigated architectures even in more structural tests.

The choice of model architecture has two impacts on the performance. On the one hand, a model should have enough parameters to adapt to the underlying nature, resulting in precise MET estimations. On the other hand, it should not rely on too many parameters as this results in overfitting and large computing times. The loss as defined in equation 6.3 is a good indicator of the estimation quality as it compares estimated to true MET and is small if the estimation is precise and accurate. Regarding computing efficiency, the runtime

Table 6.2: Training results for different update methods.

	dynamic	fixed
best evaluation loss	2106	1548
training time per epoch in s	$8512 \pm 4$	$614 \pm 18$
evaluation time per epoch in s	$2071 \pm 3$	$102 \pm 3$

of training and evaluation as a simple indicator of the usage of computing resources are considered in this comparison.

The comparing subsections are structured as follows: First, the expectation of the quantity’s impact on the performance is explained, if possible. Then the result of the comparison is presented and finally deviations from the expectation are discussed.

### 6.2.1 Fixed vs. Dynamic GNN

Two different graph styles are investigated, graphs with fixed and dynamic structure (see 5.3.2). For the fixed graph, PF candidates with a  $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$  smaller than a specific value are connected. The  $\Delta R$  requirement has proven to be efficient in jet reconstruction and is therefore also used for jet energy corrections. As these are not applied on raw PF candidates yet, GraphMET has to make up for this and learn these corrections. A forced structure with a maximum  $\Delta R$  may lead GraphMET to learn these corrections and therefore improve the result.

The dynamic graph is obtained by calculating for every graph layer the norm of the difference between the input vectors and connecting the  $k = 10$  nearest neighbors. In principle, it is expected that the dynamic graph should yield a smaller loss value as it should be able to learn the requirement for fixed graphs and further improve it. However, it is not guaranteed that the optimum is found as the optimization problem gains complexity. On another note, the runtime of both training and evaluation is expected to increase drastically compared to fixed graphs. The reason for that is the need for recalculation of the adjacency matrix for each graph layer, whereas the matrix remains the same for a fixed graph.

The training of the dynamic version was stopped after only six epochs for the dynamic graph as there was no improvement in loss for four epochs and due to the immense computing times, see table 6.2. The loss function’s value of the version using fixed graphs is considerably lower than the one using dynamic graphs indicating a better MET estimation. The computing times of the dynamic graph are higher by a factor of roughly 14 for training and 20 for evaluation.

The expected improvement in efficiency for the dynamic graphs is not observed. This might be eliminated by considering a higher  $k$ , i.e. more neighbors per node or more training data. As both would increase the training time even more, the dynamic approach is not followed further but may become interesting in the future when larger computing power is available.

### 6.2.2 Graph Convolution Network vs. Edge Convolution Network

Two update methods are investigated. GCNConv networks have already been introduced in subsection 5.3.2. The main difference to Edge Convolutional (EdgeConv) [89] networks is the order of feature aggregation and mapping with NNs. GCNConv first aggregates the connected node’s feature vectors and then maps the aggregated vector onto the output vector of the node using NNs. EdgeConv on the other hand maps the feature vectors of



Table 6.3: Training results for different graph layers.

	EdgeConv	GCNConv
best evaluation loss	1554	1548
training time per epoch in s	$1679.2 \pm 1.4$	$614 \pm 18$
evaluation time per epoch in s	$257.0 \pm 0.2$	$102 \pm 3$

two connected nodes onto an edge feature vector using NNs. In a second step, for each node the edge vectors are aggregated with a symmetric function such as the maximum or sum.

Even though there is a difference by construction between the two algorithms, it is not clear which method is expected to obtain better results in terms of the quality of MET estimation. On the one hand, the upstream pairwise mapping of information might enable the EdgeConv network to learn more precise information about the neighborhood than the GCNConv network, which directly considers the weighted sum of the input vectors within the neighborhood. On the other hand, pairwise mapping could also be disadvantageous, as it prevents learning about the entire neighborhood at once. However, since the networks consist of several layers, the differences become blurred as the aggregated information is passed and processed in both cases. Hence, no major differences are expected regarding the estimation quality. Since a self-connected graph has at least as many edges as nodes, more matrix multiplications are required when using EdgeConv. If the other network parameters are chosen to be the same, both training and evaluation computation times for EdgeConv are expected to be larger than the ones for GCNConv.

The results of this comparison are given in table 6.3. Both loss function values are in the same range, while the computing times are faster by a factor of about 2.5 for GCNConv layers.

As the computing resources needed are smaller in both training and evaluation for GCNConv and there is no improvement in loss for EdgeConv, the latter is not considered in the following.

### 6.2.3 Embedding

When dealing with discrete values, it often makes sense to use embedding in the sense of a continuous representation of a discrete quantity. Especially for quantities that need human interpretation such as the pdgID, embedding is useful as the meta-information about the context is not intrinsic to the value. Embedding addresses this problem by introducing a dummy variable for each possible value of the discrete quantity. This variable is one if and only if the PF candidate has the specific discrete value. For instance the pdgID has seven possible values (only counting the positive ones, as the charge is an input variable itself). The length of the embedding vector is therefore seven and a pdgID value of one is mapped onto a vector with a one in the first entry and six zeros in the other entries. The value 211 is mapped onto the vector with six zeros and a one in the last entry. Thus, the information of the discrete quantity is contained while decreasing the range of values and making sure that the input variables are of similar order of magnitude and hence more comparable for NNs.

This concept is supported by a few feedforward layers, which aggregate the embedded information into a 32-dimensional vector that acts as input vector of the graph layers.

Embedding is expected to improve the performance for the reasons explained above, while not increasing the runtime too much as the number of additional parameters is only about a third of the total numbers of parameters.

Table 6.4: Training results for different processing of input vectors.

	without embedding	with embedding
best evaluation loss	1680	1548
training time per epoch in s	$545.4 \pm 0.7$	$614 \pm 18$
evaluation time per epoch in s	$101.0 \pm 0.1$	$102 \pm 3$

Table 6.5: Training results for different activation functions.

	ELU	ReLU	sigmoid	tanh
best evaluation loss	1548	1555	1590	1574
training time per epoch in s	$614 \pm 18$	$616 \pm 30$	$624.3 \pm 1.0$	$622.2 \pm 1.1$
evaluation time per epoch in s	$102 \pm 3$	$102 \pm 3$	$103.0 \pm 0.2$	$102.5 \pm 0.2$

Table 6.4 shows the comparison between the baseline graph neural network and the same network without embedding. The loss of the network with embedding is about 8 % lower than the one without embedding. The training time is significantly lower without embedding, whereas the evaluation time does not differ remarkably.

This clearly shows that embedding has the expected positive impact on the performance while not increasing the runtime too much. Especially when considering that the training theoretically only has to be done once, whereas evaluation has to be done multiple times, the model with embedding is preferred and used in the further analysis.

#### 6.2.4 Activation Function

Four activation functions are compared. The following definitions are taken from [90]:

$$\text{ELU}(x) = \begin{cases} x & , x > 0 \\ \alpha \exp(x) - 1 & , x \leq 0 \end{cases}, \quad (6.4)$$

$$\text{ReLU}(x) = \max(0, x), \quad (6.5)$$

$$\text{sigmoid}(x) = \frac{1}{1 + \exp(-x)}, \quad (6.6)$$

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}. \quad (6.7)$$

$$(6.8)$$

Sigmoid and Tanh both map  $\mathbb{R}$  onto a small interval. This means that the relative difference between two huge but distinct values is much smaller after the mapping than before. Therefore, it is hard for the gradient descent method to distinguish between very high values or very low values, respectively. Thus, the adaptation of the weights, which is proportional to the gradient, becomes very small resulting in inefficient training.

ELU and ReLU have the advantage of being immune to this vanishing gradient problem at high positive values. The only difference between ELU and ReLU is the behavior for small input values. ELU is differentiable for  $\alpha > 0$ , while ReLU is not differentiable in  $x = 0$ . However, the difference in performance is considered to be small, as the main impact is expected to come from the linear gradient for  $x > 0$ .

Table 6.5 shows the performance of GraphMET with different activation functions. As expected, the loss of ELU and ReLU is significantly lower than the one of Sigmoid and Tanh. The training times are comparable as well and so are the evaluation times. As ELU has the best results, it will be used in the further comparisons.

Table 6.6: Training results for different  $\Delta R$  requirements.

	0.2	<b>0.4</b>	0.8	1.5
best evaluation loss	1609	1548	1573	1707
training time per epoch in s	$399 \pm 1$	$614 \pm 18$	$1574 \pm 1$	$5007 \pm 43$
evaluation time per epoch in s	$67.9 \pm 0.2$	$102 \pm 3$	$244.2 \pm 0.2$	$862.4 \pm 0.4$

### 6.2.5 $\Delta R$ Requirement

The smaller the radius in the  $\eta - \phi$  plane is chosen, the fewer PF candidates are connected in the graph. Thus, the information gain through message passing is potentially smaller for smaller  $\Delta R$ . Yet, a small  $\Delta R$  requirement could also be beneficial, since the information of PF candidates which are close to each other is weighted higher, if the sum in equation 5.18 has fewer summands. Therefore, there is no clear expectation regarding the performance when changing the  $\Delta R$  requirement. A larger value of  $\Delta R$  permits more PF candidates to be connected and therefore allows for the model to gain complexity as more information is passed. Hence, an increased training and evaluation time is expected for larger  $\Delta R$  values. Table 6.6 shows the training results for different  $\Delta R$  values. The optimum of the loss function is at  $\Delta R = 0.4$ . The training time increases with the radius. However, it should be noted that the maximum numbers of considered neighbors is set to 500. Therefore, for  $\Delta R = 0.2$  and  $0.4$ , all the PF candidates in the given range are connected, whereas for a bigger radius, some information can get lost because of this criterion. Yet, since the training and evaluation time increases with the number of neighbors considered, the evaluation time is expected to grow even more with an adjusted maximum number of considered neighbors. For this reason, the value of  $\Delta R \leq 0.4$  is chosen for GraphMET. It was shown that the  $\Delta R$  requirement has a strong impact on both computing resources and performance. Especially the latter one is an important finding and can also be seen in the comparison of the performance with different maximum numbers of considered neighbors.

### 6.2.6 Maximum Numbers of Neighbors Considered

In the case of GCNConv, this quantity defines the number of neighbors considered, which also match the criteria for node connections. This means that if the requirement for two nodes being connected is for example  $\Delta R \leq 0.4$  and the maximum number of neighbors considered is 100, then of all the PF candidates within the radius of  $\Delta R = 0.4$ , a random set of 100 PF candidates are connected to the node at hand.

This requirement has mostly technical reasons but should be investigated nonetheless as changes in the performance may indicate how important the information of surrounding PF candidates is for the MET estimation.

There is no clear expectation to whether higher or lower numbers of nearest neighbors are better. The strongest argument for higher numbers is consistency. If the number of neighbors in the radius is higher than the maximum number, a random set of neighbors is chosen. In this case, there is no physical explanation why the neighbors are connected. In general, a consistent way would be preferred here, e.g. to connect the 100 nearest neighbors within the radius  $\Delta R$ . This could be investigated further.

On the other hand, a small number of nearest neighbors could be beneficial, because this could lead to a more consistent number of neighboring nodes throughout the graph. If one PF candidate is for example isolated and has fewer neighbors by the  $\Delta R$  criterion, the comparability to other nodes would be better if they also have fewer neighbors. As the input vectors of connected nodes get added, the nodes with more neighbors usually have a

Table 6.7: Training results for different maximum numbers of neighbors considered.

	0	100	250	500
best evaluation loss	1978	1548	1544	1548
training time per epoch in s	$198.0 \pm 0.5$	$592.9 \pm 0.6$	$604.7 \pm 0.8$	$614 \pm 18$
evaluation time per epoch in s	$25.7 \pm 0.2$	$95.2 \pm 0.2$	$98.0 \pm 0.1$	$102 \pm 3$

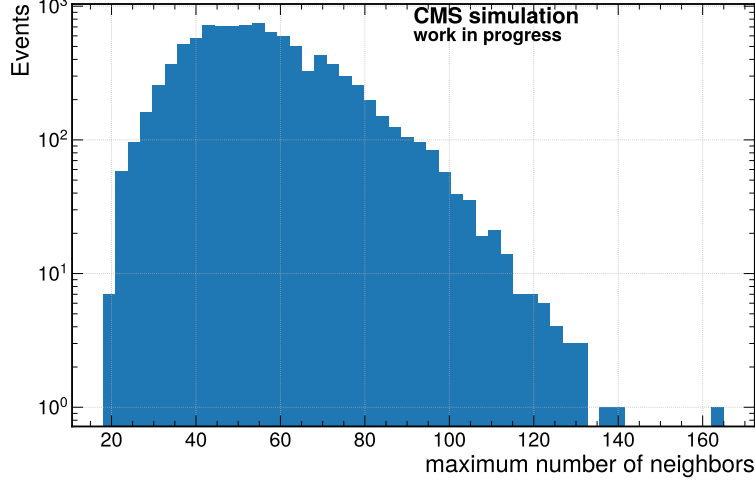


Figure 6.2: Histogram of the maximum number of edges attached to a node in one event for  $\Delta R \leq 0.4$  and 10 000  $Z \rightarrow \nu\nu$  events. 98% of the events have a maximum edge number per node of less than 100.

higher output value by construction than nodes with fewer neighbors. A smaller maximum number of neighbors considered could prevent this issue. Another way to get rid of this problem could be to normalize the sum of the input vectors by the number of connected nodes. But this is also out of the scope of this thesis, as the used methods do not provide this feature.

Table 6.7 clearly shows that the information of the neighboring nodes improves the performance of GraphMET, indicating the positive effect of using GNNs in the context of MET reconstruction. However, this effect seems to be saturated at 100 nodes. A possible explanation is that only few PF candidates have more than these 100 neighbors within the distance of  $\Delta R = 0.4$ . This argument is supported by figure 6.2, which shows the histogram of the maximum number of connections per node within one event. As about 98% of the training events have a maximum number of neighbors of less than 100, no major improvement can be observed by further increasing the maximum number of neighbors considered.

As the difference in loss and computing time is very small, the maximum number of 500 is chosen. This is because other processes might have a higher density of PF candidates and therefore a larger probability exceed the maximum of 250 neighbors. Furthermore, the number of pileup events will be further increased in the High-Luminosity LHC by a factor of more than four on average [56, 91].

### 6.2.7 Number of Graph Layers and Dimension of Feature Vectors

The number of trainable weights in a GCNConv network is driven by the product of the dimensions of input and output feature vector. Therefore, there are two ways to increase

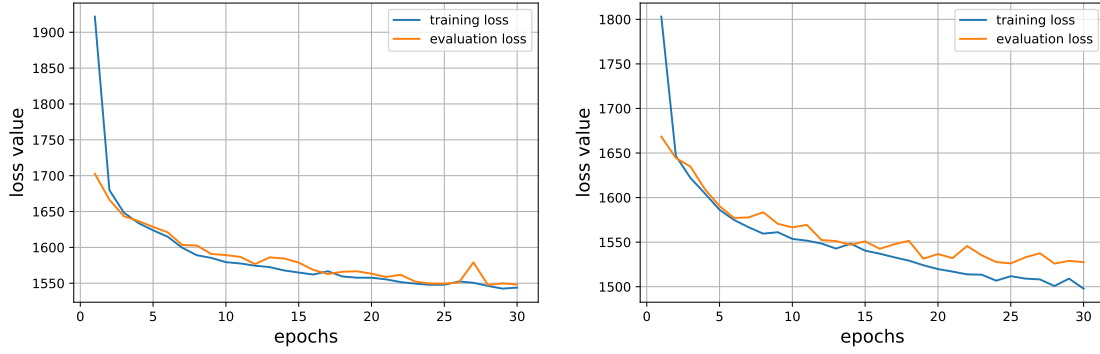


Figure 6.3: Loss function of training and evaluation set on the left for the baseline model and on the right for the model with input vector dimension 64. The baseline model loss functions show no major discrepancies between the training and evaluation set, whereas the evaluation loss of the other model is significantly higher than its training loss.

Table 6.8: Training results for different number of graph layers.

	1	2	3
best evaluation loss	1615	1557	1555
training time per epoch in s	$360.9 \pm 0.5$	$448.4 \pm 0.7$	$538.1 \pm 0.9$
evaluation time per epoch in s	$61.9 \pm 0.2$	$77.4 \pm 0.3$	$88.6 \pm 0.2$

	4	5	6
best evaluation loss	1548	1553	1557
training time per epoch in s	$614 \pm 18$	$712 \pm 4$	$803 \pm 27$
evaluation time per epoch in s	$102 \pm 3$	$115.5 \pm 0.2$	$128.2 \pm 0.3$

the number of weights: adding more dimensions to the in- or output vector and adding more layers. In general, a higher number of weights increases the flexibility of the model. This comes, however, at the cost of potential overtraining issues leading to worse generalization and robustness, and larger computing resources needed.

Table 6.8 shows that the performance of GraphMET with only one graph layer performs significantly worse than a graph with more layers. Adding more layers requires more computing resources as the number of weights increases. There is no significant change in the loss value when comparing the results for two or more layers. However, there seems to be a sweet spot of four layers, where the loss is slightly better than for other configurations. Table 6.9 shows a significant improvement of a higher feature vector dimension. This, however, comes at the cost of an overtrained model, as the comparison of the loss curves in figure 6.3 shows. Since the model is supposed to also work well on other data sets, the one with less overfitting is preferred.

Table 6.9: Training results for different dimension of feature vectors.

	8	16	32	64
best evaluation loss	1769	1595	1548	1525
training time per epoch in s	$393 \pm 3$	$468.0 \pm 1.0$	$614 \pm 18$	$953.2 \pm 1.0$
evaluation time per epoch in s	$67 \pm 1$	$78.0 \pm 0.2$	$102 \pm 3$	$146.8 \pm 0.1$

### 6.3 Comparison Between GraphMET and Existing MET Estimators

Based on the results of the previous section, only the baseline architecture is considered in the following. The idea of this comparison is to check the behavior in different MET regions and check for possible irregularities within one estimator as well as differences between the different MET estimators, which are introduced in subsection 6.3.1. The three different processes considered in this comparison are introduced in subsection 6.3.2, the results are then presented in subsection 6.3.3.

#### 6.3.1 Previous MET Estimation Methods

The PF algorithm identifies all reconstructed final state particles in a given event as described in 3.3 and is basis for all MET estimation methods compared in this section.

##### Particle Flow MET

The raw Particle Flow MET (PF MET) [66] is obtained by calculating the sum

$$\vec{E}_T = - \sum_{k=1}^{N_{\text{PF}}} \vec{p}_{T,k}, \quad (6.9)$$

where  $N_{\text{PF}}$  is the number of PF candidates and no corrections or weights are applied. By considering the JECs for each of the  $N_{\text{PF-jets}}$  PF jets with  $p_T > 10$  GeV, the so-called Type-I corrected PF MET (MET-Type-I) is obtained:

$$\vec{E}_T^{\text{Type-I}} = - \sum_{k=1}^{N_{\text{PF}}} \vec{p}_{T,k} - \sum_{k=1}^{N_{\text{PF-jets}}} (\vec{p}_{T,k}^{\text{jec-corr}} - \vec{p}_{T,k}). \quad (6.10)$$

##### PileUp Per Particle Identification MET

While PF MET corrects jets as a whole, Pile Up Per Particle Identification (PUPPI) MET goes one step further and also corrects the jet constituents for pileup. The following qualitative explanation summarizes the more detailed description in [56, 92].

As MET-Type-I, PUPPI MET takes PF candidates as input but assigns the weights in equation 3.4. In the best case, particles from pileup get a weight of zero, whereas particles from the leading vertex get a weight of one. Since charged particles leave a track in the detector, they can be assigned to the leading vertex or pileup and are therefore given a weight of one or zero, respectively.

For the weight assignment of the neutral particles, an observable is defined using information of all charged pileup particles. This observable is optimized to discriminate between pileup and leading vertex particles. Then, a chi-squared function calculates the difference in the shape of the distribution of this observable between a remaining particle and the pileup shape. If the chi-squared value is small, the particle is likely to be from pileup and gets a

small weight. If the value is big, the particle is likely to be from the leading vertex and therefore gets a weight closer to one. This weight is then used to rescale the particle's four-momentum. Finally, all particles with weights below a certain value are neglected (i.e. assigned a weight of zero), the same goes for particles with a rescaled momentum below a certain value.

Using these weights, PUPPI MET is obtained by calculating the weighted sum:

$$\vec{E}_T = - \sum_{k=1}^N w_k \vec{p}_{T,k}. \quad (6.11)$$

### DeepMET

DeepMET [85] is the predecessor of GraphMET and is therefore similar in many ways. The main difference is that the algorithm works on each PF candidate independently using regular feedforward networks for the weight assignment. In addition to the input features of GraphMET, DeepMET also uses PUPPI weights, making it dependent on another estimator. These features are mapped on one weight  $w_i$  and two biases  $b_{i,x}$ ,  $b_{i,y}$ , where the latter ones may account for possible prejudice in the  $x$  or  $y$  direction. As shown in [85], their difference in MET estimation quality is negligible. Even so, the bias terms are used the MET estimation of DeepMET:

$$\vec{E}_T = - \sum_{k=1}^N \left[ w_k \begin{pmatrix} p_{x,k} \\ p_{y,k} \end{pmatrix} + \begin{pmatrix} b_{x,k} \\ b_{y,k} \end{pmatrix} \right]. \quad (6.12)$$

There are two versions of DeepMET. DeepMETResolutionTune is trained to perform well in resolution tests, whereas DeepMETResponseTune focuses more on getting a precise response. These quantities will be defined in subsection 6.3.3.

#### 6.3.2 Samples

In this comparison, three different MC samples are considered, each representing a different process:  $Z/\gamma^* \rightarrow l\bar{l}$ ,  $W \rightarrow l\nu$ , and  $Z \rightarrow \nu\nu$ . For a fair comparison, neither of them is biased by any additional cuts. In order to understand the difference in the MET distributions between these samples, the main process is neglected at first and will be considered later in the argumentation chain. Apart from the MET of the main process, there are different sources of MET in each sample.

On the one hand, events may contain fake MET that emerges because of the finite precision of measurements. If a particle's transverse momentum is not estimated perfectly, the difference to its true transverse momentum acts as MET and is therefore called fake MET. Another source of fake MET are detected particles from pileup as they do not belong to the event and therefore distort the estimation formula. Since there are lots of particles in one event, there are also many sources of fake MET. However, the particles are approximately uniformly distributed in the transverse direction. As a result, fake MET is also approximately uniformly distributed and the vector sum of all sources of fake MET should be close to zero in most cases.

On the other hand there is MET from undetectable daughter particles, such as neutrinos, produced from weak interactions of jet constituents. MET induced by such daughter particles does not make the MET estimation wrong but it interferes with the interpretation of MET as transverse momentum of the possible undetectable particle from the main event. Even so, undetectable daughter particles with large transverse momentum are not likely to occur since the share of the initial transverse momentum decreases with each decay.

In summary, if there are no systematic mismeasurements, the MET distribution apart from

the main process is expected to have a sharp peak around zero and contain only few events with larger amounts of MET.

On top of that, zero, one or two sources of real MET are added in the different samples:

- $Z/\gamma^* \rightarrow ll(+0 \text{ jets})$ : As discussed in appendix B, this process is often used for MET performance tests if the generator information is not available. Here,  $ll$  denotes an electron-positron or a muon-anti-muon pair. The full sample was evaluated with the different MET estimation methods, resulting in a total number of  $N_{Z/\gamma^* \rightarrow ll} = 86\,443\,196$  processed events. Including no undetectable particles from the main interaction, this process is considered to contain only fake MET and MET from undetectable daughter particles.
- $W \rightarrow l\nu(+2 \text{ jets})$ : This process has one neutrino from the main interaction in the final state. Having no counterpart in the other direction, the transverse momentum of the neutrino should exceed fake MET by far, resulting in a MET distribution different from zero. The sample contains  $N_{W \rightarrow l\nu} = 94\,830\,399$  events.
- $Z \rightarrow \nu\nu (+1 \text{ jet})$ : This process has two neutrinos from the main interaction and therefore also at least two neutrinos passing the detector producing MET. Since the neutrinos are in general not back to back in the reference system of CMS, MET should again exceed fake MET by far, resulting in a MET distribution different from zero. The sample consists of four MET-binned samples with a total number of  $N_{Z \rightarrow \nu\nu} = 69\,583\,383$  events.

Figure 6.4 shows the behavior of Generated MET (GenMET) and the MET estimations for the different samples. The key findings from these plots are the following: First and foremost, GraphMET and DeepMETResolutionTune tend to estimate MET to be smaller than the other reconstruction methods. The peaks in the distributions of GraphMET and PF MET are about 10–20 GeV apart from each other. Moreover, the MET distributions of PF MET, PUPPI MET and DeepMETResponseTune tend to be wider than the corresponding distributions of GraphMET and DeepMETResolutionTune.

Regarding the  $\phi$  distribution of MET, the target of a flat distribution is matched best by GraphMET and PUPPI MET, while the other MET reconstruction methods show a cosine-shaped modulation. An interesting behavior can be observed in the  $Z \rightarrow \nu\nu$  sample, where a dip at  $\phi = 0$  distorts the otherwise reasonable behavior. A possible reason for this dip is a systematic underestimation of the  $y$  or overestimation of the  $x$  component of MET by GraphMET, resulting in a tendency towards small angles. However, this dip is not observed in the other regions making this argument unlikely. Statistical fluctuations are also unlikely as the expected fluctuation in a bin with 700 000 events is  $\sqrt{700\,000} \approx 837$ . As such dips are not observed in other investigated samples, this behavior was not further investigated but should be kept in mind in case of recurrence in further analyses.

### 6.3.3 Comparison

In order to compare the different kinds of reconstruction algorithms, the resolution and response are calculated, which indicate the precision and accuracy, respectively. In general, there are some differences in the handling of simulated and measured events. Since only simulated data sets are used in this comparison, the formalism is introduced only for them. The formalism for real data is given in appendix B. In simulated events, GenMET can be used as a proxy for true MET.



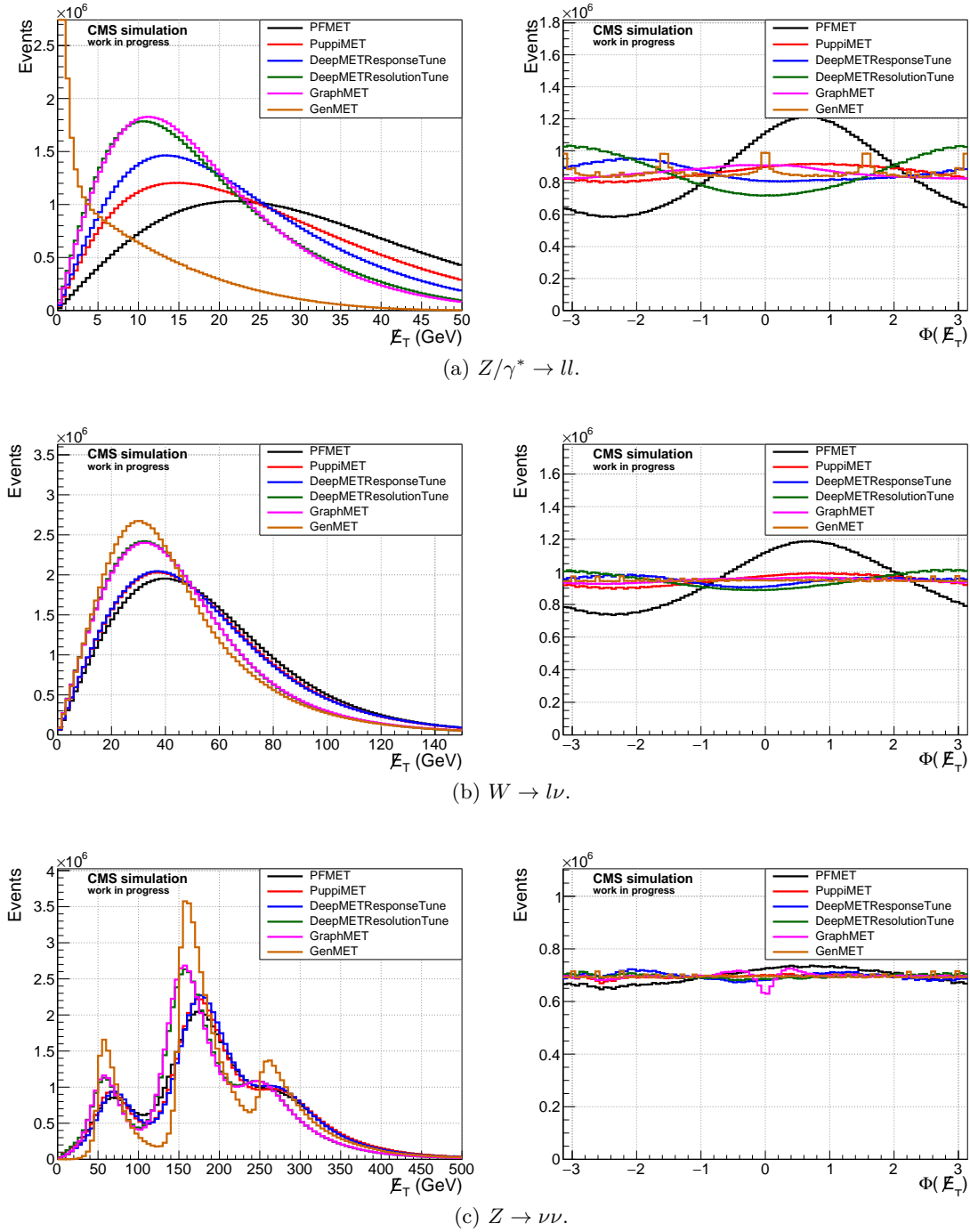


Figure 6.4: Norm (left) and angular (right) distribution of MET. In general, the true shape (GenMET) is matched best by DeepMETResolutionTune and GraphMET. For  $Z/\gamma^* \rightarrow ll$ , the sharp peak around  $\cancel{E}_T = 0$  is consistently overestimated and the estimated angular distributions show modulations, which are especially strong for PF MET and DeepMET. The expected uniform distribution in GenMET is disturbed by four small peaks at  $\phi = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$ , which are a remnant from the generation process. For  $W \rightarrow l\nu$ , the MET distribution is estimated better and the modulation in  $\phi$  is only strong for PF MET. The combined  $Z \rightarrow \nu\nu$  sample shows a slight tendency for underestimation of GraphMET and DeepMETResolutionTune, while the angular distribution shows no anomalies except for a small dip around  $\phi = 0$  for GraphMET.

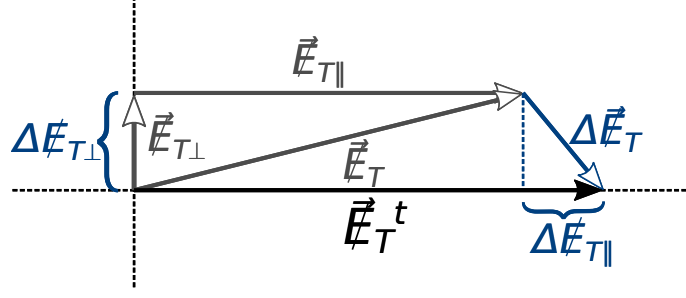


Figure 6.5: MET reconstruction in the transverse plane. The true MET is denoted with a superscript t. The estimated MET does not have a superscript.

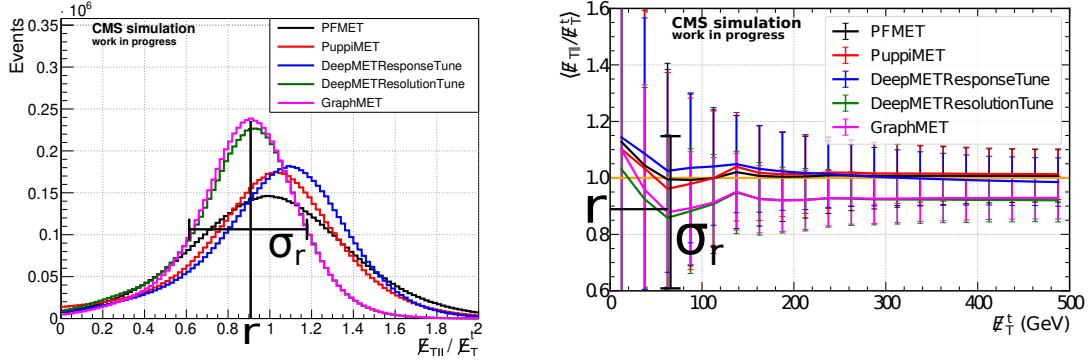


Figure 6.6: Construction example for a response plot. The plot on the left contains the histogram of the response values, where the true MET for all events is between 50 GeV and 75 GeV. The mean  $r$  and the standard deviation  $\sigma_r$  of the distribution are then entered into the corresponding true MET bin of the plot on the right. This plot shows now the response behavior for a wide range of true MET.

## Response

The response corresponds to the value of the estimated MET projected onto the direction of true MET  $\frac{\vec{E}_T \cdot \vec{E}_T^t}{|\vec{E}_T^t|}$ , normed to the amount of true MET  $|\vec{E}_T^t|$ :

$$r \equiv \frac{\vec{E}_T \cdot \vec{E}_T^t}{|\vec{E}_T^t|^2}. \quad (6.13)$$

Figure 6.5 shows the transverse plane with the estimated and true MET. An estimation is correct if and only if  $\vec{E}_T = \vec{E}_T^t$ . In this case, the response is one.

However, MET estimators do not necessarily perform equally well in different MET regions. Therefore, the response is often calculated for different MET bins leading to a distribution as depicted in figure 6.6. The resulting response curves are shown in figure 6.7. A poor response behavior is implied by both a response far from one and a large distribution width. In the first case, MET is consistently under- or overestimated while the second case implies that there is a big variance around the target value, thus often resulting in an over- or underestimated MET. Since the definition of the response has the true MET in its denominator, regions with low MET tend to have a large distribution width. For these bins, a comparison of the response values alone is not meaningful as the differences between the distributions are negligible. For higher MET regions, where the distribution widths decrease, it can be concluded that the tendency of GraphMET and DeepMETResolutionTune to underestimate the parallel MET component is not negligible.

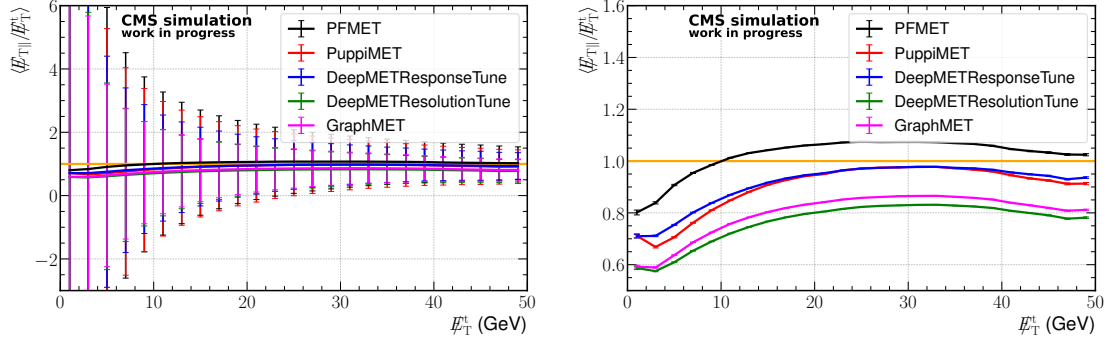
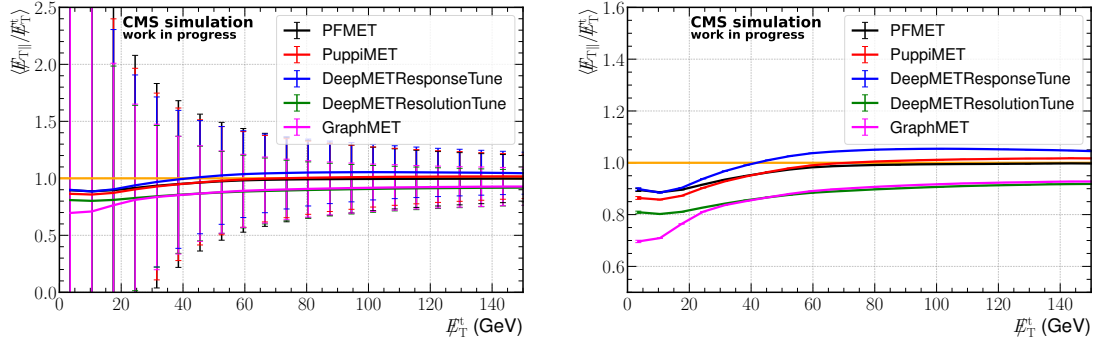
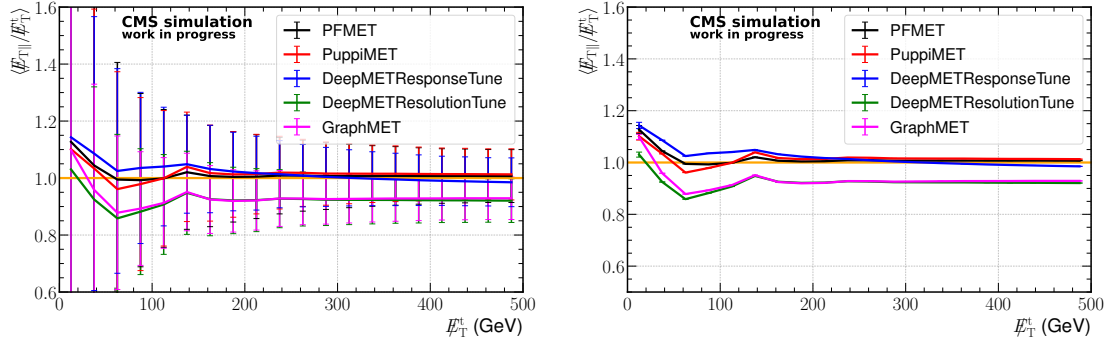
(a)  $Z/\gamma^* \rightarrow ll$ .(b)  $W \rightarrow l\nu$ .(c)  $Z \rightarrow \nu\nu$ .

Figure 6.7: Response plots with the distribution widths as error bars on the left and standard deviation on the mean value as error bars on the right. In general GraphMET and DeepMETResolutionTune show similar behavior with response values of about 90% for regions with large MET, where the other estimators are close to 100%. For low MET regions, the distribution widths are large compared to the mean value and tend to decrease with increasing true MET.

## Resolution

The precision of a MET estimator can be expressed as its resolution. The resolution of the component in direction of the true MET is called parallel resolution. With the definition of the response, the parallel and perpendicular components of the difference between estimated and true MET,  $\Delta\vec{E}_T = \vec{E}_T - \vec{E}_T^t$ , can be obtained by calculating

$$\Delta E_{T,\parallel} \equiv |\vec{E}_T^t| - |r \cdot \vec{E}_T^t|, \quad (6.14)$$

$$\Delta E_{T,\perp} \equiv |-\vec{E}_T + r \cdot \vec{E}_T^t|. \quad (6.15)$$

Figure 6.5 shows the transverse plane with the estimated and true MET as well as the difference vector and its components. The resolutions are then defined as the standard deviation of the distribution of the corresponding component of the difference vector:

$$\sigma_{\Delta E_{T,\parallel}} = \sigma(\Delta E_{T,\parallel}) = \frac{1}{2}(\Delta E_{T,\parallel,0.84} - \Delta E_{T,\parallel,0.16}), \quad (6.16)$$

$$\sigma_{\Delta E_{T,\perp}} = \sigma(\Delta E_{T,\perp}) = \Delta E_{T,\perp,0.68}, \quad (6.17)$$

where instead of the standard deviation the difference between the 84% and 16% quantile is taken for the parallel resolution and the 68% quantile for the perpendicular resolution. The difference between the definitions is due to the parallel resolution taking both positive and negative values whereas the perpendicular resolution is by definition positive.

Again, an estimation is correct if and only if  $\vec{E}_T = \vec{E}_T^t$ . This is the case if and only if the difference vector is zero. In this case, the resolutions are closer to zero than if the MET is not estimated correctly in this event. The aim of the MET estimation is therefore to reduce the resolutions as much as possible while getting a response of one. As was seen in the response comparison, some estimators yield a response less than one. These estimators have, by construction, an advantage in resolution comparisons, because both resolutions are compressed by the factor of underestimation. By scaling the resolutions with the inverse response, this effect is compensated leading to a more fair comparison. This scaling is a sort of calibration and might be included in future versions of GraphMET as it shifts the response to one.

The scaled resolutions are given in figure 6.8. While the response is worst for GraphMET and DeepMETResolutionTune, they perform best in parallel and perpendicular resolution. This means that the part of the estimated MET parallel to true MET is consistently underestimated but with a smaller deviation from its mean in both parallel and perpendicular directions.

It can be summarized that the new methods for MET estimation, DeepMET and GraphMET, improve the overall quality of the reconstruction. Dependent on what an analysis focuses on, a calibration could be applied or not in order to obtain a reconstruction method with focus on the response or resolution, respectively. Even if the response is in the focus, the resolution is still better for the new estimation methods than for the previous ones, PF MET and PUPPI MET. The improvement is seen in different MET regions for different samples and is hence expected to benefit all kinds of analyses. One exemplary sensitivity analysis will mark the closing chapter of this thesis.

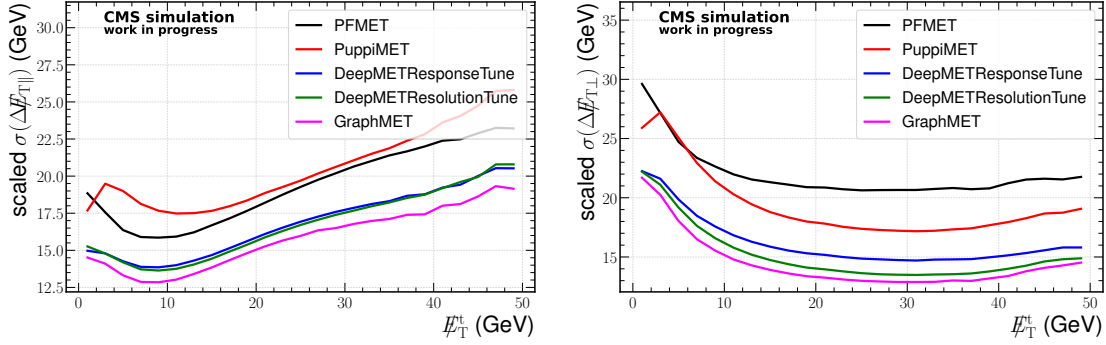
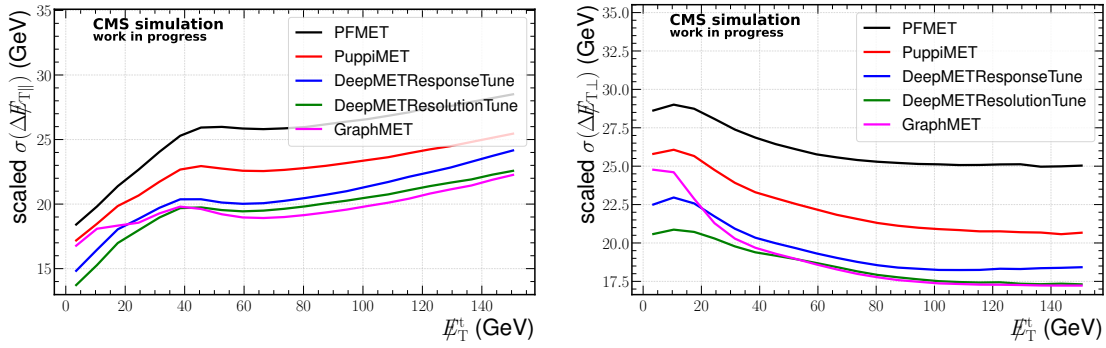
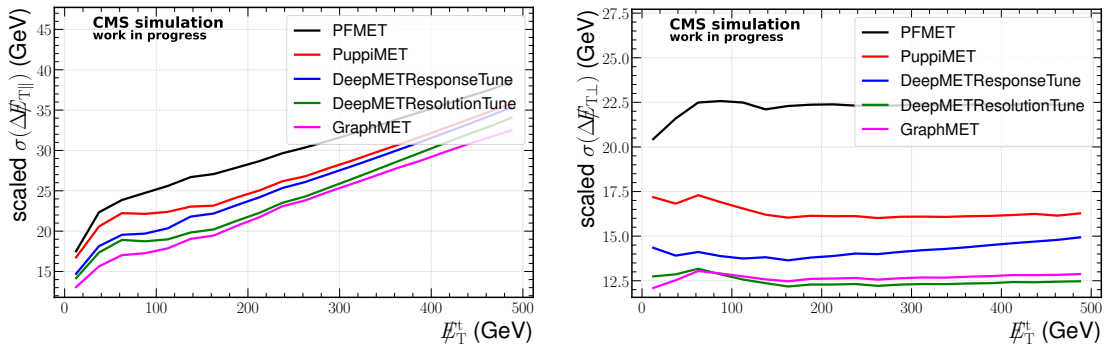
(a)  $Z/\gamma^* \rightarrow ll$ .(b)  $W \rightarrow l\nu$ .(c)  $Z \rightarrow \nu\nu$ .

Figure 6.8: Scaled parallel (left) and perpendicular (right) resolution plots. In general GraphMET and DeepMETResolutionTune show the smallest resolution values in all regions. PF MET has the largest perpendicular resolution values for all regions and also the largest values for the  $W \rightarrow l\nu$  and  $Z \rightarrow \nu\nu$  regions. PUPPI MET mostly has medium resolution values except for the parallel resolution in the  $Z\gamma \rightarrow ll$  region, where it has the largest values aside from the first bin.



## 7 GraphMET in a Mono-Top Dark Matter Analysis

Having shown that using GraphMET might improve the estimation of Missing Transverse Energy (MET) at the CMS experiment, it remains to be investigated how this affects the sensitivity of an analysis. For this, searches for Dark Matter (DM) at hadron colliders, where MET is a key observable, are well suited. The approach in this thesis follows the idea of previous publications [41, 93] as well as current searches for DM together with a single top quark. This mono-top signature is explained in section 7.1. Section 7.2 gives an overview of the selection criteria applied to the phase space detected by the CMS experiment, followed by a description of the systematic uncertainties considered in the further analysis in section 7.3. The simulated data is then validated in section 7.4 before presenting the results of a sensitivity analysis for GraphMET and already existing MET reconstruction methods in section 7.5.

### 7.1 Mono-Top Signature

In general, there are three ways to observe DM. Direct detection works by measuring the recoil caused by DM scattering with SM matter. Indirect detection on the other hand aims at measuring signals originating from annihilating DM particles into SM particles. The third method is the production of DM from SM particles and observation of an excess in events arising from the additional contribution to the observed phase space. The methods are described in more detail for example in [94].

The approach in this thesis belongs to the third group. Using a simplified model for DM production, a prediction with respect to possible model parameter combinations is given and compared to the experiment by means of the LHC test statistic (see section 4.2).

The simplified model in this thesis is based on [95–99]. It introduces a coupling between fermionic DM particles  $\chi$  with mass  $m_\chi$  to SM particles via a mediator  $\phi$  with mass  $m_\phi$  and spin one in units of  $\hbar$ . In addition to the masses, the coupling strengths between the mediator and SM or DM particles are parameters of the model. This model allows for additional processes enhancing the probability to obtain certain final states. One of these is the production of a single top quark with large MET (mono-top) where the top quark decays into a bottom quark and a leptonically decaying  $W$  boson, also called leptonic channel. Figure 7.1 gives example Feynman diagrams producing a mono-top signature.

Within the SM, this process is strongly suppressed for two reasons. The first reason is

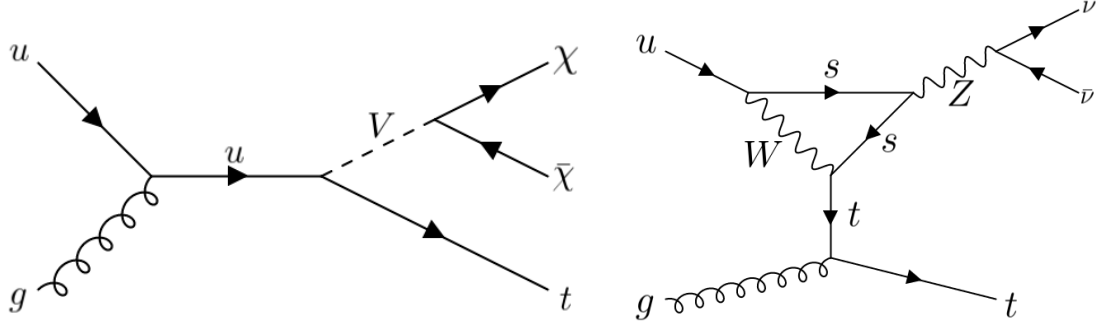


Figure 7.1: Example Feynman diagrams for the production of a mono-top signature in a simplified DM model (left) and with SM processes only (right). The diagrams are replicated from [41] using [100].

the necessary appearance of several weak interactions including quark flavor changes. As only a single top quark is produced, it has to emerge from a weak interaction. Emerging particles other than the top quark must either annihilate again resulting in loop diagrams or convert to neutrinos, which needs additional weak couplings. Moreover, the flavor change from the proton's constituents, up and down quarks, to the top quark requires flavor changing processes, decreasing the cross section even more as the off-diagonal elements of the CKM matrix are well below unity. The second reason is a strong suppression of flavor changing neutral currents, to which the process belongs. This suppression is explained by the GIM mechanism [101] and results from similar processes with amplitudes of different sign interfering destructively. Thus, additional events with a mono-top signature could in principle be detected easily as they hardly occur in the SM. Even though this seems to be clean from a theoretical point of view, it is challenging for the experimental analysis to rule out events with a different theoretical signature but much larger cross section, such as events with a  $W$  quark decaying leptonically and additional jets. This challenge is tackled by applying an event selection.

## 7.2 Event Selection

The aim of an event selection is to find a subset of the total CMS phase space in which the signal-to-background ratio is sufficient for the given purposes. First, several triggers are chosen in order to ensure that only well-constructed events are considered and to decrease the vast amount of data. Second, a more detailed selection is applied, to suppress background without affecting the signal too much.

### 7.2.1 Triggers

In the leptonic mono-top channel, there should be one well-reconstructed lepton. Consequently, at least one of the recommended leptonic HLTs in table 7.1 must have fired. The single electron triggers require either an electron isolated from jets in the  $\eta - \phi$  plane with a transverse momentum of at least 32 GeV, or a non-isolated electron with at least 50 GeV and a PF jet with at least a transverse momentum of 165 GeV, or a non-isolated electron with at least 115 GeV. Alternatively, events are also considered if a photon trigger with a threshold transverse momentum of 200 GeV has fired because photons and electrons behave similarly at large momenta. The single muon triggers require an isolated muon with a transverse momentum of at least 24 GeV, or a non-isolated muon with at transverse momentum threshold of 50 GeV or 100 GeV.



Table 7.1: Leptonic triggers for the analyses, recommended by [102, 103].

lepton	HLT triggerpath
muon	HLT_IsoMu24
	HLT_Mu50
	HLT_OldMu100
	HLT_TkMu100
electron	HLT_Ele32_WPTight_Gsf
	HLT_Ele115_CaloIdVT_GsfTrkIdT
	HLT_Ele50_CaloIdVT_GsfTrkIdT_PFJet165
	HLT_Photon200

### 7.2.2 Selection Criteria

In the mono-top DM analysis at least one jet from the decaying top quark is needed and one additional jet is allowed to account for possible gluon radiation and to construct a  $t\bar{t}$  Control Region (CR). Jets need to have at least a transverse momentum of 30 GeV, while the leading jet with the highest transverse momentum is required to have at least 70 GeV of transverse momentum. In this thesis, only leptonically decaying top quarks are considered. Therefore, exactly one lepton  $\ell$  is required. Both, leading jet and lepton have to be within the central part of the detector with  $|\eta| \leq 2.5$ , where the inner tracker provides excellent information. The MET criterion of  $|\vec{\cancel{E}}_T| \geq 100$  GeV accounts for the DM particles leaving the detector without interaction. This cut depends on the particular reconstruction method under study.

In addition to these low-level criteria, there are two high-level requirements also depending on the investigated MET estimator. At leading order, the mediator and top quark are back-to-back in  $\phi$ . From this, an additional criterion is derived, namely that  $\Delta\phi$  between MET and the jet with the largest transverse momentum is larger than 1.5. This value is chosen rather loose because the decay products of the top quark can split the transverse momentum among themselves. Moreover, in the leptonic decay channel, the neutrino's contribution to MET distorts the picture. Finally, the transverse mass

$$M_T \equiv \sqrt{2p_T(\ell)p_T(\vec{\cancel{E}}_T)\{1 - \cos[\phi(\ell) - \phi(\vec{\cancel{E}}_T)]\}} \quad (7.1)$$

is required to exceed 40 GeV. In a pure  $W$  decay, this distribution has a steep decrease, called Jacobian edge, beyond the  $W$  boson mass, which can be explained best in the rest frame of the  $W$  boson. Here, the daughter particles have same momenta but in opposite direction. Therefore, in the rest frame of the  $W$  boson, the transverse mass, now given by

$$M_{T,\text{rf}} = \sqrt{4p_{T,\text{rf}}(\ell)p_{T,\text{rf}}(\vec{\cancel{E}}_T)} = 2p_{T,\text{rf}}(\ell) \leq 2p_{\text{rf}}(\ell) = m_W, \quad (7.2)$$

cannot exceed the mass of the  $W$  boson. Even though the transverse mass is invariant under boosts along the  $z$  axis by construction, it is smeared by a boost in transverse direction. As a result, the abrupt decrease is also smeared [104].

In addition to these selection criteria, a Signal Region (SR) and two CRs are defined. The SR is the phase space in which the analysis will be most sensitive to the signal. Here, as an additional requirement, exactly one b-tagged jet is demanded. As a result, this phase space is enriched in events with only one b-tag. The CRs are needed for checking the behavior of the MC simulation with data in a phase space orthogonal to the SR in order to avoid human bias [105] and to improve the background estimation in the SR. Demanding two b-tagged jets, the first CR is enriched in  $t\bar{t}$  events and is therefore referred to as  $t\bar{t}$  CR.

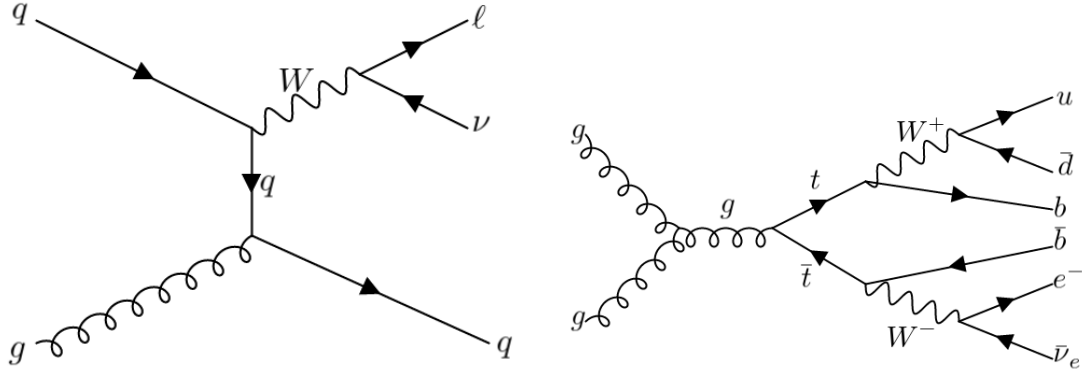


Figure 7.2: Example Feynman diagrams of the main background processes  $W$ +jets on the left, where the  $W$  boson decays leptonically, and  $t\bar{t}$  on the right, where the  $W^+$  boson decays hadronically and the  $W^-$  boson decays leptonically. The diagrams are produced with [100].

The  $W$  CR on the other hand requires exactly zero b-tagged jets and is therefore enriched in  $W$ +jets events.

The signal selection criteria lead to a phase space subset enriched in possible DM signal. However, there are processes which produce signatures so similar that they may survive the aforementioned selection criteria. Such events as well as the strongly suppressed SM processes with exactly the same signature are the background processes of the search for DM in the mono-top channel. Since the PF information needed for the evaluation with GraphMET requires much storage, only the main background processes are considered in this thesis, namely  $W$ +jets and  $t\bar{t}$  events, depicted in figure 7.2.

The first background process can survive the selection criteria if the  $W$  decays leptonically, leading to MET and a lepton in the final state and if, in addition to that, a light flavor jet is mistaken as b-jet. Processes with two top quarks also often survive the selection criteria if one of the b-jets is not identified as such, and if one  $W$  boson decays leptonically and the other one hadronically or both  $W$  bosons decay leptonically and one of the two leptons is not detected.

The strategy of only using the two main backgrounds is motivated by limited available storage space. Nevertheless, neglecting the minor backgrounds is not considered a major disadvantage as this thesis does not focus on producing expected upper limits for the real analysis but on checking and comparing the impact of different MET reconstruction methods on an analysis, where the expected upper limit indicates differences in the sensitivity.

### 7.3 Systematic Uncertainties

In the following, the set of systematic uncertainties considered in the analysis is described. For each uncertainty, a nuisance parameter will be introduced in the fits for the expected asymptotic upper limit.

- **Pileup:** To account for pileup, simulated events are weighted to match the distribution of the number of interactions in the data events [69]. Using the uncertainties to the inelastic proton-proton cross section, uncertainties related to pileup are derived.
- **B-Tag:** B-tagging has a different efficiency in data and simulation and is therefore subject to scale factors [72], for which uncertainties are introduced.

- **Mistagging:** Mistagging also has different rates in data and simulation. The corresponding uncertainties are accounted for.
- **JEC:** This uncertainty group is introduced for JES and JER corrections [69], affecting both shape and rate of distributions. In this thesis, they are propagated to all jet-related quantities except for MET. This deviation from the usual approach is due to similar uncertainties not being available for all MET reconstruction methods. Consequently, the corresponding systematic uncertainties are not taken into account for any MET reconstruction methods for a fair comparison. Even though these jet-related uncertainties are not propagated to MET, they are still considered as nuisances in the fit.  
This approach is not problematic when performing the sensitivity analysis only using statistical uncertainties, as done in subsection 7.5.2. However, there will be an impact on the analysis when using systematic uncertainties, since an additional source of uncertainty may lead to tighter fit results but increases the total uncertainty.
- **PDFs:** This uncertainty comprises information on deviations of parton distribution functions [106, 107].
- **Theoretical uncertainties:** These mainly come from only considering the first orders in the Taylor expansion of the matrix element for signal and background processes and are separated into uncertainties on factorization and renormalization scales.
- **Statistics:** Sparse numbers of simulated events cause uncertainties that are treated as systematics. The incorporation into the fit follows the Barlow-Beeston light approach [108–110].
- **Lepton:** This uncertainty group contains uncertainties in the lepton identification and reconstruction [54, 111]. These directly affect the scale of the lepton distributions and therefore both the scale and the shape of related distributions, such as the transverse mass.
- **Parton shower:** This uncertainty group accounts for uncertainties in the parton shower generation, i.e. initial state and final state radiation [112, 113].
- **Luminosity:** The uncertainty on the integrated luminosity provided by CMS is 2.5% [114] and is treated as a rate uncertainty.

## 7.4 Validation of Simulation

This analysis is pursued using simulated data of LHC Run 2 in the year 2018. Before analyzing the SR, MC simulated background samples are validated with real data in the CRs. The real data samples consist of events from the year 2018, in which an electron, high-energy photon, or a muon could be reconstructed by HLTs. The validation has two main objectives: As a first step, the general behavior of the simulated events is checked against data. Since several different MET estimation methods are analyzed in this thesis, the second task is to compare the distributions of quantities depending on MET.

In order to fulfill these tasks in a compact way, three MET related distributions are shown in figure 7.3 for the commonly used jet energy corrected MET-Type-I and for the new GraphMET. Additional distributions are shown in appendix D. In principle, the distributions show that the background model is sufficient for the given purposes. The slight MC overshoot could be reduced, e.g. by applying higher order QCD corrections to the  $W + \text{jets}$  samples, and the shape differences could be reduced by applying electroweak corrections to the  $W + \text{jets}$  samples.

There are two important differences which can be explained by the behavior found in the tests on unbiased samples in section 6.3. Firstly, the event yield using GraphMET is smaller than the one using MET-Type-I. This can be explained by the GraphMET response being smaller than one. This response behavior corresponds to MET being systematically underestimated. Consequently, the applied MET cut at 100 GeV affects more events when using GraphMET than when using MET-Type-I. This effect could be countered with a smaller MET cut for GraphMET. As this approach is only a largely simplified way of applying response calibrations, it will not be studied in detail in this thesis. The impact of this response effect on the sensitivity of the analysis is unclear at this point. Responsible for this is the unknown signal distribution allowing for three cases based on the amount of signal rejected due to this response behavior compared to the amount of rejected background, respectively. If the ratio of the rejected to the total number of events is similar for both, the signal-to-background ratio remains the same. If, on the other hand, the ratio of neglected background events to the total number of background events is larger (smaller) than the respective quantity for the signal, the signal-to-background ratio will increase (decrease) and thus have a positive (negative) impact on the sensitivity. Secondly, the Jacobian edge is much sharper when using GraphMET, which can be explained by the impact of the improved resolution on the transverse mass distribution. This effect can be clarified by considering an event, which is located at the edge of the distribution in terms of generator information. Since the precision of MET-Type-I is not as good as the one of GraphMET, it will often assign values different from the true MET value. As MET is a contribution to the calculation of the transverse mass, this quantity will have a larger distribution width resulting in a smeared and thus less prominent descend. Judging only by the steeper descend in the transverse mass distribution, a better discrimination power is expected for GraphMET in the SR.

## 7.5 Expected Impact of GraphMET on a Mono-Top Sensitivity Analysis

In the SR, only pseudodata is considered. Pseudodata consists of simulated background events as well as simulated signal events. The signal events are normalized to a cross section of 1 pb for the different mediator and DM mass combinations, which is in the order of magnitude of the theoretical cross section predictions for a DM mass of  $m_\chi = 150$  GeV and mediator masses between roughly 200 GeV and 1000 GeV. In these calculations, the coupling constant for interactions between the mediator and SM quarks is set to 0.25, and to 1.0 for interactions between the mediator and DM candidates, respectively. Increasing these coupling constants would increase the cross sections and vice versa.

### 7.5.1 Signal Region Plots

For reasons of comparability, figure 7.4 shows the same plots as in the previous section but now in the SR with exactly one b-tagged jet and a muon in the final state. In these plots, the background behaves similar to the previously discussed CRs. Again, the event yield is larger for MET-Type-I than for GraphMET, and the Jacobian edge in the transverse mass distribution is more pronounced for GraphMET. This allows in general for a better discriminating power using GraphMET. The multiplication factor applied to the signal normalizes the signal yield to the background yield. Hence, the smaller factor for GraphMET can be translated into a better signal-to-background ratio.

Surprisingly, the pseudodata over background for MET seems to be better for MET-Type-I. A possible explanation is that GraphMET neglects more data than background because of the limited response. However, this seems unlikely as the signal shape tends to have more

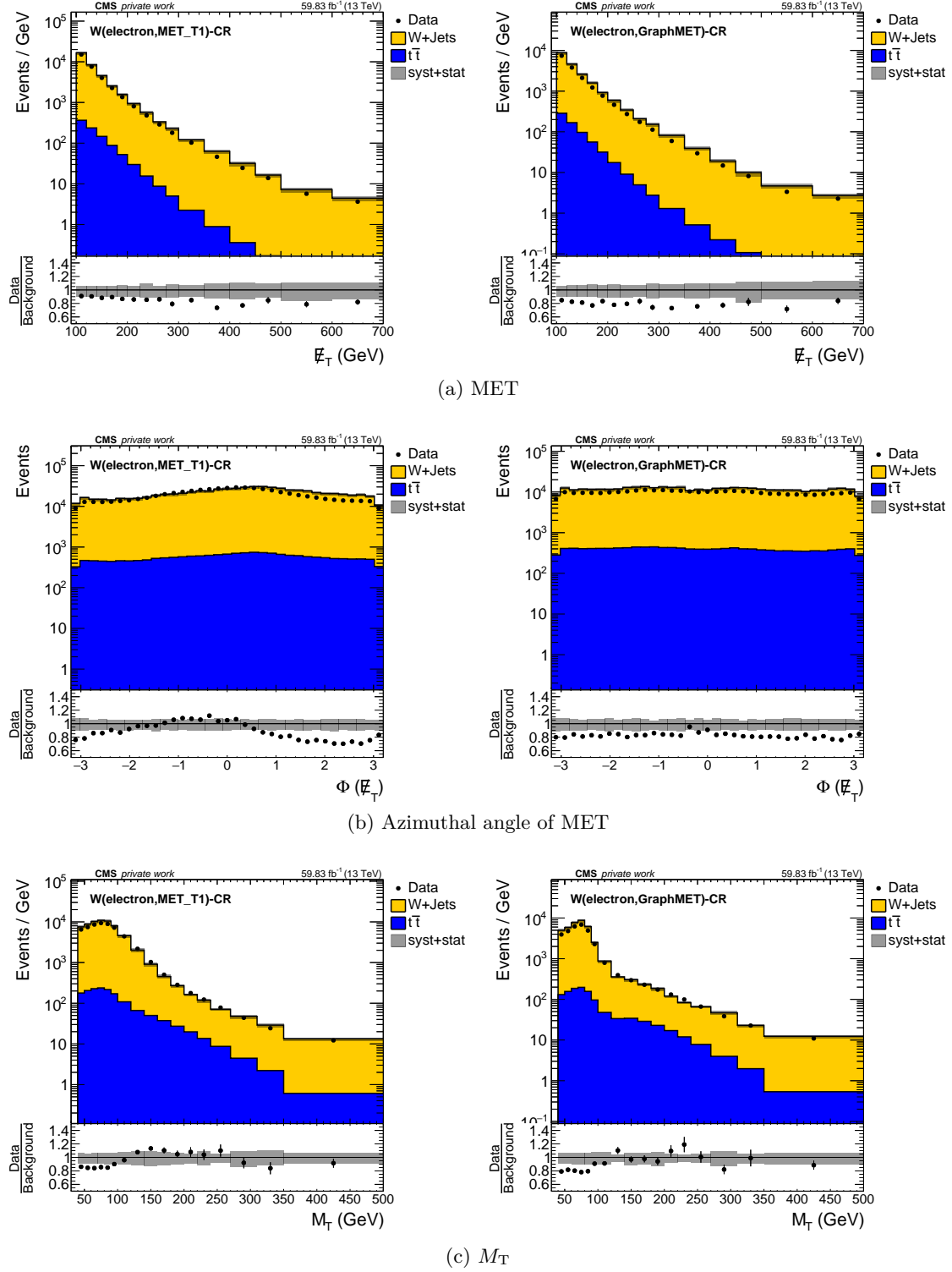


Figure 7.3: MET-dependent distributions in the W CR with a muon in the final state. On the left, MET-Type-I was used for MET reconstruction and GraphMET on the right. In general, the plots show an MC overshoot of roughly 10 – 20 % for both MET reconstruction methods. The shapes of MET look similar but the total yield is smaller for GraphMET. The  $\Phi(\text{MET})$  distribution shows a modulation for MET-Type-I. A major discrepancy can be seen in the  $M_T$  distribution, where the decrease is much steeper using GraphMET. Here, the bin width is variable to increase the statistics in the less populated bins. Thus, the bin content has to be multiplied by the bin width in order to obtain the absolute event number.

entries relative to the background for larger MET. Another possible reason is the worse resolution of MET-Type-I, smearing the signal distribution towards higher MET. Yet, also this seems unlikely as the background is not affected by this behavior even though the peaks in both distributions are less than 40 GeV apart. In general, this could also be due to statistical fluctuations as the difference in the pseudodata to background behavior is well within one standard deviation of the pseudodata points.

The transverse mass distribution shows the expected improvement in pseudodata-to-background when using GraphMET. Even though the mean values of pseudodata over background indicate slight improvements when using GraphMET, this is also within the pseudodata standard deviation.

As the fit will only be performed in one distribution, it remains to be decided whether to use the MET or  $M_T$  distribution. Based on the pseudo data to background ratio, the transverse mass is to be preferred for the fits when using GraphMET as there is a significant deviation from the background in the rightmost bins of the  $M_T$  distribution which cannot be observed in the MET distribution. For MET-Type-I, there is no clear expectation towards the more promising quantity as some bins of the MET distribution also show promising pseudodata over background results. These observations will now be quantified by the calculation of median expected asymptotic upper limits.

### 7.5.2 Expected Upper Limits

The way to obtain sensitivity expectations is followed as described in chapter 4. The expected upper limit  $\mu^{95\%CL_s}$  is the expected smallest value of the signal strength parameter that can be excluded with at least 95%  $CL_s$  confidence level. In case of the  $CL_s$  method, an exclusion means that the probability to obtain a value of the test statistic larger than the observed value given the s+b hypothesis is twenty times smaller than the corresponding probability given the b hypothesis. Intuitively, this means that it is twenty times more likely to find an upward fluctuation of the b hypothesis to match the observation than a downward fluctuation of the s+b hypothesis. The smaller the expected upper limit, the larger the excluded parameter space and the more sensitive is an analysis.

A limit is called stat-only if the systematic uncertainties are disregarded during the fit. This allows for a comparison of different MET reconstruction methods free of bias from systematic uncertainties. The alternative approach of considering systematic uncertainties as nuisances in the fit, on the other hand, has the main advantage of being more realistic.

#### Stat-Only

In order to investigate the effects of the choice of fitting observable on the sensitivity, a stat-only fit is performed using the MET and  $M_T$  distributions in the SR with a muon in the final state. This allows for a direct comparison between the SR plots in figure 7.4 and their impact on the expected upper limit. The results of these stat-only fits are given in figure 7.5.

There are two major differences between the limits obtained with the different distributions. Firstly, in comparison to the other upper limits for a given mediator mass, the GraphMET limits are smaller than the other limits for the transverse mass while the opposite holds for MET. This is in agreement with the results of the previous subsection, where the Jacobian edge in the background was steeper using GraphMET, allowing for better separation between signal and background.

Secondly, the upper limits to the cross section are smaller when using the transverse mass compared to using MET. This confirms the observation of the previous subsection at least for GraphMET, but also MET-Type-I, for which no clear expectations were present has smaller limits when using the transverse mass. Since the transverse mass delivers more

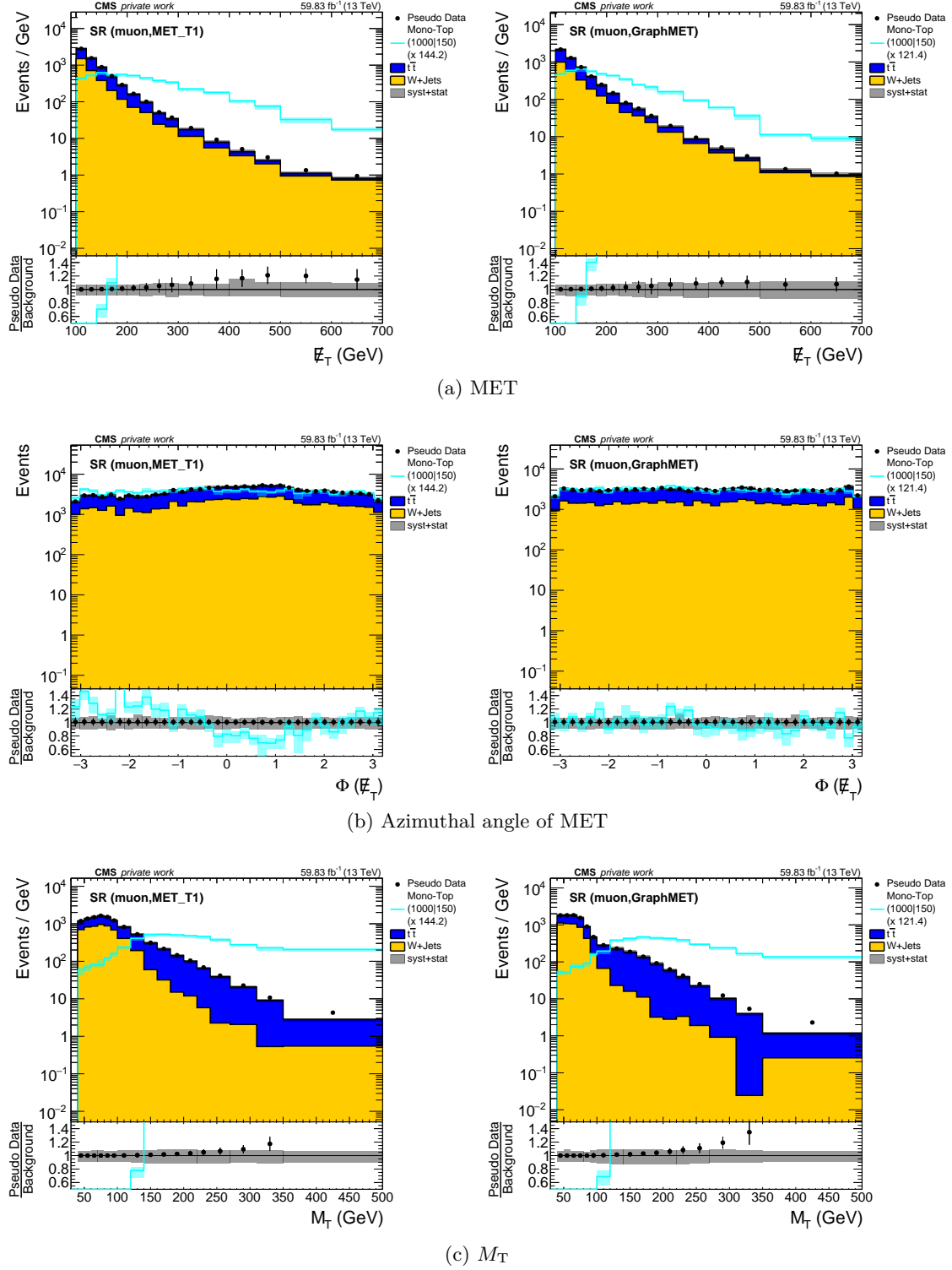


Figure 7.4: MET-dependent distributions in the SR with a muon in the final state. On the left, MET-Type-I was used for MET reconstruction and GraphMET on the right. The distributions contain the main background processes as well as an MC signal sample multiplied with a scaling factor normalizing the signal-to-the background. This factor is not applied to the pseudodata. The event yield is about the same and the MET shape of the background does not show major discrepancies. The pseudodata-over-background is distinctively larger than one when using MET-Type-I and only slightly above one for GraphMET. The angular distribution shows the usual modulation for MET-Type-I, which is not seen for GraphMET. The transverse mass distribution shows a steeper descend for GraphMET. Furthermore, the pseudodata-over-background is larger for GraphMET.

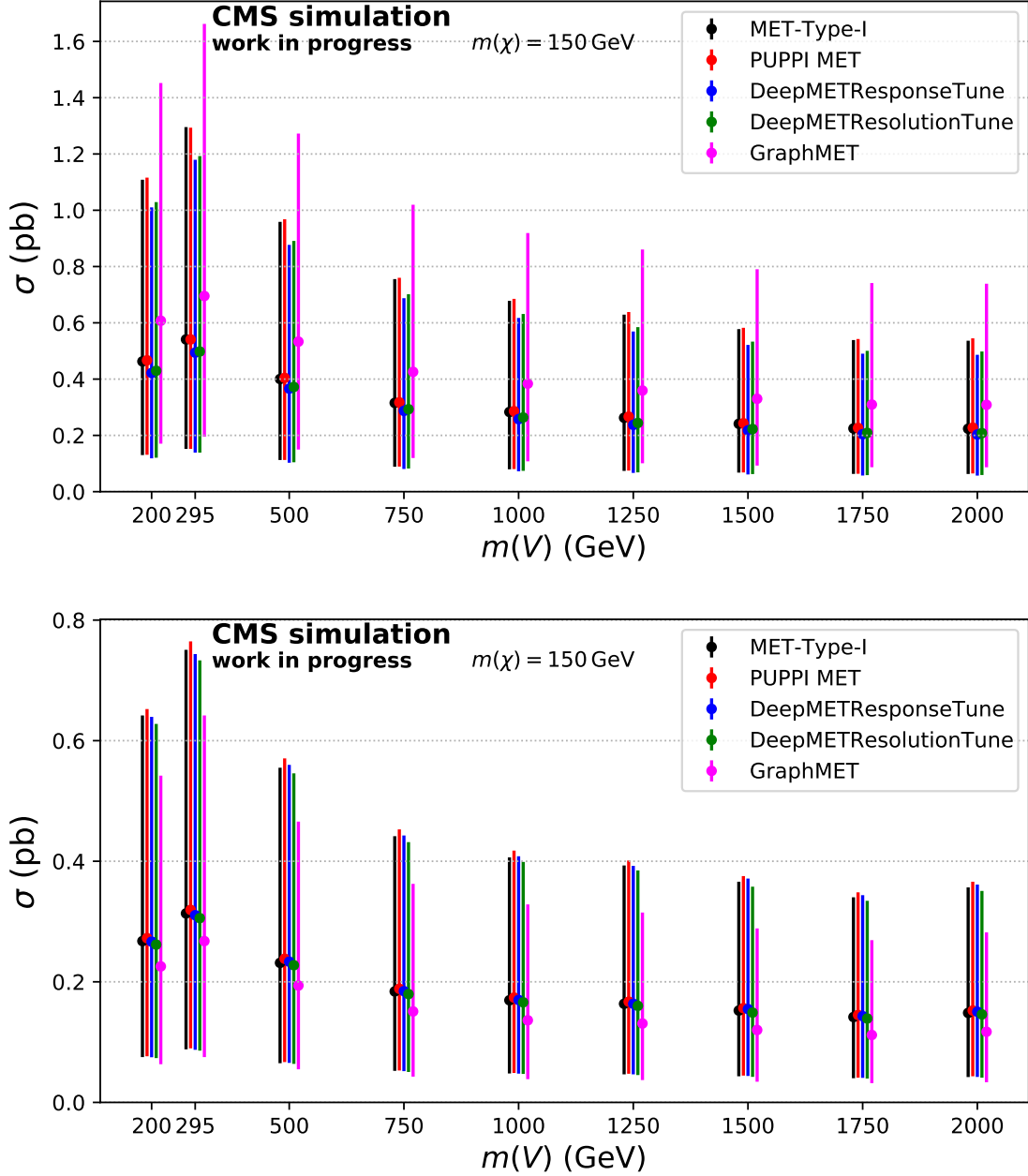


Figure 7.5: Expected asymptotic upper limits on the DM production cross section for different mediator masses and a DM mass of  $m_\chi = 150 \text{ GeV}$  using stat-only fits in the SR with a muon in the final state. The dots are the expected asymptotic upper limits and the error bars indicate their 16% and 84% quantiles. At the top, the expected analysis calculations is performed using the MET distributions and the  $M_T$  distributions were used in the calculations at the bottom. From 200 GeV to 295 GeV, an increase is observed with both quantities. Beyond this mediator mass, the cross section limits decrease and remain almost constant from 1000 GeV on. In general, the limits in the plot at the bottom are tighter. Moreover, GraphMET produces larger limits in the MET fit but smaller limits in the  $M_T$  fit.



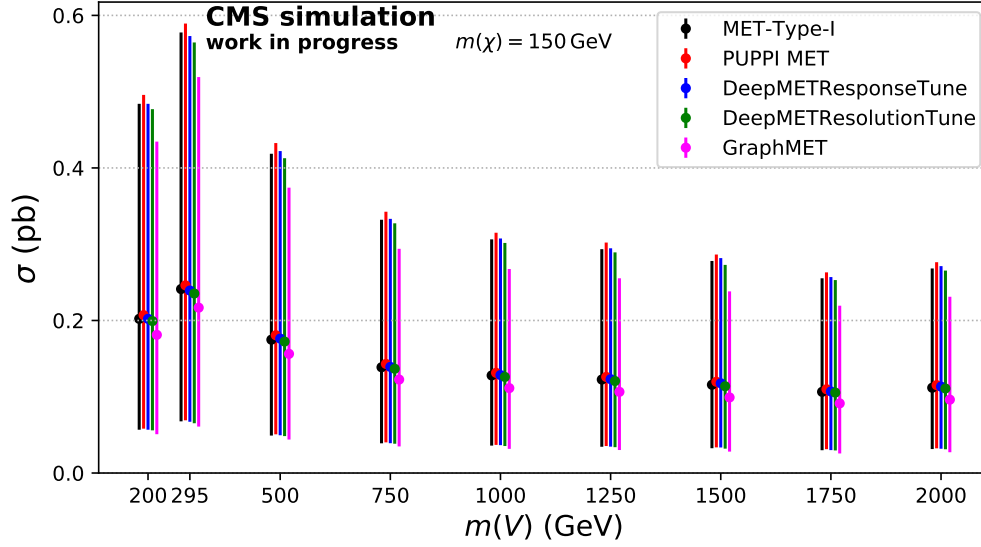


Figure 7.6: Expected asymptotic upper limits on the DM production cross section for different mediator masses and a DM mass of  $m_\chi = 150$  GeV using stat-only fits in the SR. The sub-regions with an electron and muon in the final state are simultaneously fitted. The dots are the expected asymptotic upper limits and the error bars indicate their 16% and 84% quantiles. The expected analysis calculations is performed using the  $M_T$  distributions. From 200 GeV to 295 GeV, an increase is observed. Beyond this mediator mass, the cross section limits decrease and remain almost constant from 1000 GeV on. The limits of the conventional methods are similar with only the GraphMET limit having both, a smaller value as well as smaller deviations.

promising results both, visible in the SR plots and also in the expected upper limits, it will be used from now on in the calculation of combined limits for both signal sub-regions, with a muon or electron in the final state.

The results of this fit are presented in figure 7.6. It is shown that the smallest limits can be produced using GraphMET. The difference between the limits is small compared to the standard deviation but the consistency of the improvement over all mediator masses indicates the stability of this improvement.

### Including Systematics

In this fit, systematic uncertainties are considered as nuisance parameters. With the nuisance parameters allowing for fits closer to reality than with statistical uncertainties alone, the resulting upper limits are compared to the theoretical cross sections in figure 7.7. These theoretical predictions are derived using **MadGraph5** with the implemented mono-top model from [97]. The increase in theoretical cross section for the smallest three mediator masses can be explained by the phase space which is small below 300 GeV since the mass of the two DM particles exceeds the mass of the mediator in these cases. Thus, the mediator is produced off-shell, which suppresses the cross section. The production of mediators with large masses is also suppressed, since the invariant mass of the system of the scattering partons needs to be large as well for the production of heavy mediators. This becomes increasingly rare the higher the mediator masses, hence the decrease in the cross section for larger mediator masses.

The behavior of the upper limits is similar to the ones with statistical uncertainties only.

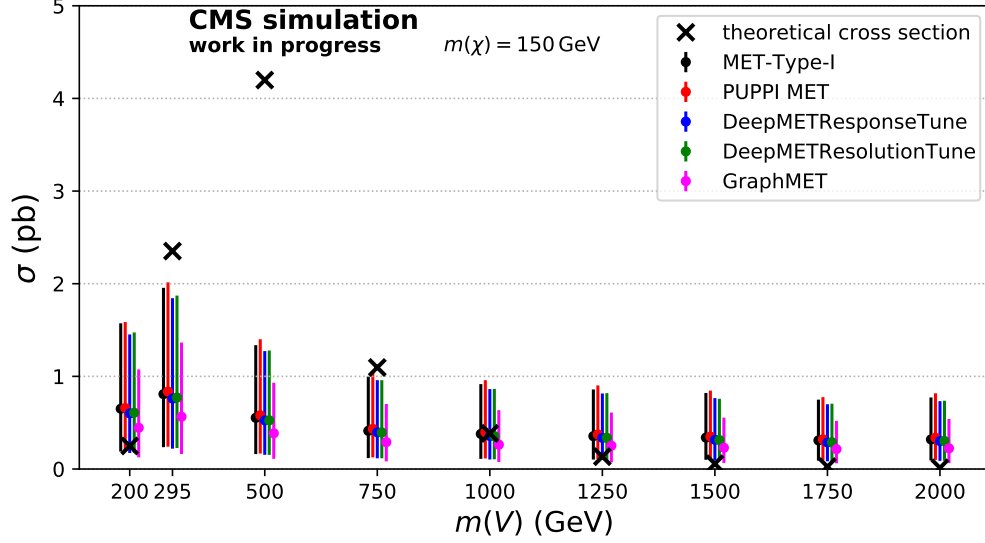


Figure 7.7: Expected asymptotic upper limits on the DM production cross section for different mediator masses and a DM mass of  $m_\chi = 150$  GeV using fits including systematics in the SR. The sub-regions with an electron and muon in the final state are simultaneously fitted. The dots are the expected asymptotic upper limits and the error bars indicate their 16% and 84% quantiles. The crosses indicate the theoretical cross sections from simulations. The expected analysis calculations is performed using the  $M_T$  distributions. From 200 GeV to 295 GeV, an increase is observed in expected limits. Beyond this mediator mass, the cross section limits decrease and remain almost constant from 1000 GeV on. The theoretical cross section increases up to 500 GeV, from where it decreases rapidly. For 295 GeV, 500 GeV and 750 GeV, the theoretical cross section is clearly larger than all expected limits. For 1000 GeV, the theoretical cross section is in the same range as the expected upper limits of the existing MET reconstruction methods but would be expected to exceed the upper limit using GraphMET.

But it can be seen that the improvement in sensitivity would make a difference in this case. As all values above the upper limit are expected to be excluded by the analysis, the cross section of the mediator masses 295, 500, and 750 GeV could be excluded by all three analyses. The cross section for a mediator with mass 1000 GeV, however, could only be expected to be excluded by the analysis using GraphMET.

This demonstrates that GraphMET does not only promise improvements in theory, as shown in the last chapter, but is also expected to make a difference in future analyses.

In order to investigate whether GraphMET's worse response and the induced cut with a possible better signal-to-background ratio are responsible for this improvement in sensitivity, a naive calibration is applied to GraphMET by scaling it with the inverse mean response value. This should shift the response to one while losing slightly in resolution as discussed in section 6.3.3. As the results, which are shown in appendix E, do not indicate a significant loss in sensitivity compared to the uncalibrated GraphMET, there is strong evidence that the observed improvement when using GraphMET is at least partly a result of the improved resolution.

## 8 Conclusion and Outlook

Missing Transverse Energy (MET) is an important quantity for many analyses at the Compact Muon Solenoid (CMS) experiment because it allows an indirect measurement of particles that cannot be detected directly, such as neutrinos or Dark Matter (DM) particles. As the CMS detector measures multiple collisions at once, the key challenge is to mitigate the impact of this pileup on the MET reconstruction. In the past, this has been done by applying Jet Energy Corrections (JECs) to the jets or by assigning weights to reconstructed particles in the MET reconstruction. The latter approach is followed by GraphMET which calculates these weights using Graph Neural Networks (GNNs). This type of Neural Network allows to utilize information not only from one particle at once but also from possible correlations between particles.

This master thesis consists of two parts, the first focusing on GraphMET, a new method for MET reconstruction at CMS, while the second part, a sensitivity analysis searching for DM, provides a first use case for this new reconstruction method.

Several GraphMET architectures were investigated and compared regarding their ability to estimate MET while not demanding unreasonable computing resources. One architecture was found to fit these requirements best. This method was investigated in great detail and compared to previous reconstruction methods on large data sets orthogonal to the training data set. While the response hints at a tendency for GraphMET to underestimate MET, GraphMET was found to perform better than previous MET reconstruction methods in terms of resolution. The response challenge may in principle be faced with calibrations or targeted loss functions in future trainings, leading to a MET reconstruction method performing on a high level in both resolution and response. The architecture is also manageable in terms of required computing resources, making this approach suited for a broader range of users within the CMS group.

With GraphMET showing promising results in dedicated MET performance studies, it is interesting to quantify the impact on an analysis result. Therefore, studies of the expected significance, quantified by the asymptotic median expected upper limit, are performed in the second part of this thesis. In order to observe differences between the different MET reconstruction methods, the analysis should be sensitive to MET or related quantities. Hence, the new reconstruction method was applied to a search for DM, where MET is the only available piece of information on DM. The simplified model introduces a mediator and a stable particle in the DM sector. The mediator allows flavor changing neutral currents

that are strongly suppressed within the SM. By applying several selection criteria, the phase space is reduced to a subset with an optimized signal to background ratio. Here, the two main backgrounds are considered, leading to a background model sufficient for the purposes of comparing expected asymptotic upper limits without paying too much attention to their specific absolute value.

Comparisons in control and signal regions indicate that the improved resolution allows for a sharper Jacobian edge in the  $M_T$  distribution, while the worse response decreases the total event yield. The sharper edge allows better discrimination between signal and background, which can also be observed when comparing the expected asymptotic upper limits. In order to account for the specific systematic uncertainties that are not yet available for GraphMET, these are also not propagated to the other MET estimations. Statistics-only and systematics-included expected asymptotic upper limits consistently show smaller limits when using GraphMET, indicating improved sensitivity when using this MET reconstruction method. However, this improvement is always well within a standard deviation of the limit and is therefore not statistically significant.

Even though some improvements have already been observed in this thesis, there are multiple aspects that may lead to further improvements. Regarding GraphMET, the main challenge is to improve the response while not losing too much resolution. If alternative approaches in the definition of the loss function or architecture do not lead to the desired improvement, calibrations might be another way to obtain an unbiased response. Another possibility is to train one version to perform well in resolution and another version for response, as it has been done with DeepMET before [85]. In this case, different analyses can choose which reconstruction to consider based on their needs for resolution or response. At last, systematic uncertainties directly related to GraphMET should be studied in order to fully understand the behavior of GraphMET and propagate this into analyses accordingly. As soon as these uncertainties are available, it may be interesting to reperform the mono-top analysis in the leptonic channel. Until then, it might be more interesting to focus on analyses that are less statistically challenging or more sensitive to MET resolution and response in order to investigate possibly larger differences between the various MET estimators. Searches for DM in events with a single jet and MET might be well-suited for this, as these do not contain neutrinos with large transverse momentum. Thus, MET provides better access to the transverse momentum of the DM particles in such analyses.

In general, GraphMET showed promising results that suggest further studies. Especially in view of the upcoming data taking eras with increased luminosity and pileup, GraphMET might become a key tool in the search for new physics at the CMS experiment.

# Bibliography

- [1] S. F. Novaes. *Standard Model: An Introduction*. 2000. DOI: 10.48550/ARXIV.HEP-PH/0001283. URL: <https://arxiv.org/abs/hep-ph/0001283>.
- [2] M. Kobayashi and T. Maskawa. “CP-Violation in the Renormalizable Theory of Weak Interaction”. In: *Progress of Theoretical Physics* 49.2 (Feb. 1973), pp. 652–657. ISSN: 0033-068X. DOI: 10.1143/PTP.49.652. eprint: <https://academic.oup.com/ptp/article-pdf/49/2/652/5257692/49-2-652.pdf>. URL: <https://doi.org/10.1143/PTP.49.652>.
- [3] S. Abachi et al. “Observation of the Top Quark”. In: *Phys. Rev. Lett.* 74 (14 Apr. 1995), pp. 2632–2637. DOI: 10.1103/PhysRevLett.74.2632. URL: <https://link.aps.org/doi/10.1103/PhysRevLett.74.2632>.
- [4] F. Abe et al. “Observation of top quark production in  $\bar{p}p$  collisions”. In: *Phys. Rev. Lett.* 74 (1995), pp. 2626–2631. DOI: 10.1103/PhysRevLett.74.2626. arXiv: hep-ex/9503002.
- [5] S. Abachi et al. “Observation of the top quark”. In: *Phys. Rev. Lett.* 74 (1995), pp. 2632–2637. DOI: 10.1103/PhysRevLett.74.2632. arXiv: hep-ex/9503003.
- [6] F. Englert and R. Brout. “Broken Symmetry and the Mass of Gauge Vector Mesons”. In: *Phys. Rev. Lett.* 13 (9 Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321. URL: <https://link.aps.org/doi/10.1103/PhysRevLett.13.321>.
- [7] P. W. Higgs. “Broken Symmetries and the Masses of Gauge Bosons”. In: *Phys. Rev. Lett.* 13 (1964). Ed. by J. C. Taylor, pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.
- [8] P. W. Higgs. “Broken symmetries, massless particles and gauge fields”. In: *Phys. Lett.* 12 (1964), pp. 132–133. DOI: 10.1016/0031-9163(64)91136-9.
- [9] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (2012), pp. 1–29. ISSN: 0370-2693. DOI: <https://doi.org/10.1016/j.physletb.2012.08.020>. URL: <https://www.sciencedirect.com/science/article/pii/S037026931200857X>.
- [10] S. Chatrchyan et al. “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”. In: *Physics Letters B* 716.1 (2012), pp. 30–61. ISSN: 0370-2693. DOI: <https://doi.org/10.1016/j.physletb.2012.08.021>. URL: <https://www.sciencedirect.com/science/article/pii/S0370269312008581>.
- [11] R. Wolf. *The Higgs Boson Discovery at the Large Hadron Collider*. Vol. 264. Springer, 2015. ISBN: 978-3-319-18512-5. DOI: <https://doi.org/10.1007/978-3-319-18512-5>.

- [12] K. J. Barnes. *Group theory for the standard model of particle physics and beyond*. Boca Raton, [2010]. URL: <https://doi.org/10.1201/9781439895207%20%20https://www.taylorfrancis.com/books/oa-mono/10.1201/9781439895207/group-theory-standard-model-particle-physics-beyond-ken-barnes>.
- [13] P. D. Group et al. “Review of Particle Physics”. In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020). 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptaa104. eprint: <https://academic.oup.com/ptep/article-pdf/2020/8/083C01/34673722/ptaa104.pdf>. URL: <https://doi.org/10.1093/ptep/ptaa104>.
- [14] Wikipedia, the free encyclopedia. *Standard Model of Elementary Particles*. [Online; accessed August, 9th, 2021]. 2019. URL: [https://en.wikipedia.org/wiki/File:Standard\\_Model\\_of\\_Elementary\\_Particles.svg](https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg).
- [15] R. Davis Jr., D. S. Harmer, and K. C. Hoffman. “Search for neutrinos from the sun”. In: *Phys. Rev. Lett.* 20 (1968), pp. 1205–1209. DOI: 10.1103/PhysRevLett.20.1205.
- [16] Y. Fukuda et al. “Solar Neutrino Data Covering Solar Cycle 22”. In: *Phys. Rev. Lett.* 77 (9 Aug. 1996), pp. 1683–1686. DOI: 10.1103/PhysRevLett.77.1683. URL: <https://link.aps.org/doi/10.1103/PhysRevLett.77.1683>.
- [17] W. Hampel et al. “GALLEX solar neutrino observations: results for GALLEX IV”. In: *Physics Letters B* 447.1 (1999), pp. 127–133. ISSN: 0370-2693. DOI: [https://doi.org/10.1016/S0370-2693\(98\)01579-2](https://doi.org/10.1016/S0370-2693(98)01579-2). URL: <https://www.sciencedirect.com/science/article/pii/S0370269398015792>.
- [18] J. N. Abdurashitov et al. “Measurement of the solar neutrino capture rate with gallium metal”. In: *Phys. Rev. C* 60 (5 Oct. 1999), p. 055801. DOI: 10.1103/PhysRevC.60.055801. URL: <https://link.aps.org/doi/10.1103/PhysRevC.60.055801>.
- [19] Y. Fukuda et al. “Evidence for Oscillation of Atmospheric Neutrinos”. In: *Physical Review Letters* 81.8 (Aug. 1998), pp. 1562–1567. DOI: 10.1103/physrevlett.81.1562. URL: <https://doi.org/10.1103%2Fphysrevlett.81.1562>.
- [20] M. Apollonio et al. “Limits on neutrino oscillations from the CHOOZ experiment”. In: *Physics Letters B* 466.2-4 (Nov. 1999), pp. 415–430. DOI: 10.1016/s0370-2693(99)01072-2. URL: <https://doi.org/10.1016%2Fs0370-2693%2899%2901072-2>.
- [21] F. P. An et al. “Observation of Electron-Antineutrino Disappearance at Daya Bay”. In: *Phys. Rev. Lett.* 108 (17 Apr. 2012), p. 171803. DOI: 10.1103/PhysRevLett.108.171803. URL: <https://link.aps.org/doi/10.1103/PhysRevLett.108.171803>.
- [22] M. H. Ahn et al. “Measurement of neutrino oscillation by the K2K experiment”. In: *Physical Review D* 74.7 (Oct. 2006). DOI: 10.1103/physrevd.74.072003. URL: <https://doi.org/10.1103%2Fphysrevd.74.072003>.
- [23] P. Adamson et al. “Measurement of Neutrino Oscillations with the MINOS Detectors in the NuMI Beam”. In: *Physical Review Letters* 101.13 (Sept. 2008). DOI: 10.1103/physrevlett.101.131802. URL: <https://doi.org/10.1103%2Fphysrevlett.101.131802>.
- [24] S. Abe et al. “Precision Measurement of Neutrino Oscillation Parameters with KamLAND”. In: *Physical Review Letters* 100.22 (June 2008). DOI: 10.1103/physrevlett.100.221803. URL: <https://doi.org/10.1103%2Fphysrevlett.100.221803>.

- [25] A. Einstein. “Die Grundlage der allgemeinen Relativitätstheorie”. In: *Annalen der Physik* 354.7 (1916), pp. 769–822. DOI: <https://doi.org/10.1002/andp.19163540702>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.19163540702>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19163540702>.
- [26] S. Weinberg. *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*. New York, NY: Wiley, 1972. URL: <https://cds.cern.ch/record/100595>.
- [27] T. Hebbeker et al. *General Model Independent Searches for Physics Beyond the Standard Model*. SpringerBriefs in Physics. Springer, Aug. 2020. ISBN: 978-3-030-53782-1, 978-3-030-53783-8. DOI: 10.1007/978-3-030-53783-8.
- [28] J. Silk et al. *Particle Dark Matter: Observations, Models and Searches*. Ed. by G. Bertone. Cambridge: Cambridge Univ. Press, 2010. ISBN: 978-1-107-65392-4. DOI: 10.1017/CB09780511770739.
- [29] V. C. Rubin, J. Ford W. K., and N. Thonnard. “Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 (R=4kpc) to UGC 2885 (R=122kpc).” In: *The Astrophysical Journal* 238 (June 1980), pp. 471–487. DOI: 10.1086/158003.
- [30] E. Corbelli and P. Salucci. “The extended rotation curve and the dark matter halo of M33”. In: *Monthly Notices of the Royal Astronomical Society* 311.2 (Jan. 2000), pp. 441–447. DOI: 10.1046/j.1365-8711.2000.03075.x. URL: <https://doi.org/10.1046%2Fj.1365-8711.2000.03075.x>.
- [31] F. Zwicky. “Republication of: The redshift of extragalactic nebulae”. In: *General Relativity and Gravitation* 41.1 (Jan. 2009), pp. 207–224. DOI: 10.1007/s10714-008-0707-4.
- [32] D. Clowe et al. “A Direct Empirical Proof of the Existence of Dark Matter”. In: *The Astrophysical Journal* 648.2 (Aug. 2006), pp. L109–L113. DOI: 10.1086/508162. URL: <https://doi.org/10.1086%2F508162>.
- [33] O. Chwolson. “Über eine mögliche Form fiktiver Doppelsterne”. In: *Astronomische Nachrichten* 221.20 (1924), pp. 329–330. DOI: <https://doi.org/10.1002/asna.19242212003>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asna.19242212003>.
- [34] A. Einstein. “Lens-Like Action of a Star by the Deviation of Light in the Gravitational Field”. In: *Science* 84.2188 (1936), pp. 506–507. DOI: 10.1126/science.84.2188.506. eprint: <https://www.science.org/doi/pdf/10.1126/science.84.2188.506>. URL: <https://www.science.org/doi/abs/10.1126/science.84.2188.506>.
- [35] E. Komatsu et al. “Five-Year Wilkinson Microwave Anisotropy Probe Observations: Cosmological Interpretation”. In: *The Astrophysical Journal Supplement Series* 180.2 (Feb. 2009), pp. 330–376. DOI: 10.1088/0067-0049/180/2/330. URL: <https://doi.org/10.1088%2F0067-0049%2F180%2F2%2F330>.
- [36] N. Aghanim et al. “Planck 2018 results”. In: *Astronomy & Astrophysics* 641 (Sept. 2020), A6. DOI: 10.1051/0004-6361/201833910. URL: <https://doi.org/10.1051%2F0004-6361%2F201833910>.
- [37] R. Penco. *An Introduction to Effective Field Theories*. 2020. DOI: 10.48550/ARXIV.2006.16285. URL: <https://arxiv.org/abs/2006.16285>.
- [38] S. P. MARTIN. “A Supersymmetry Primer”. In: *Perspectives on Supersymmetry*. World Scientific, July 1998, pp. 1–98. DOI: 10.1142/9789812839657\_0001. URL: [https://doi.org/10.1142%2F9789812839657\\_0001](https://doi.org/10.1142%2F9789812839657_0001).

- [39] L. Roszkowski, E. M. Sessolo, and S. Trojanowski. “WIMP dark matter candidates and searches—current status and future prospects”. In: *Reports on Progress in Physics* 81.6 (May 2018), p. 066201. DOI: 10.1088/1361-6633/aab913. URL: <https://doi.org/10.1088/1361-6633/aab913>.
- [40] *Search for monotop in the muon channel in proton-proton collisions at  $\sqrt{s} = 8$  TeV*. Tech. rep. Geneva: CERN, 2016. URL: <https://cds.cern.ch/record/2143260>.
- [41] M. Waßmer. “Search for the production of Dark Matter in hadronic mono-top signatures with the CMS experiment”. PhD thesis. Karlsruhe Institute of Technology (KIT), 2021. DOI: 10.5445/IR/1000129768.
- [42] G. C. Wick. “The Evaluation of the Collision Matrix”. In: *Phys. Rev.* 80 (2 Oct. 1950), pp. 268–272. DOI: 10.1103/PhysRev.80.268. URL: <https://link.aps.org/doi/10.1103/PhysRev.80.268>.
- [43] “LHC Machine”. In: *JINST* 3 (2008). Ed. by L. Evans and P. Bryant, S08001. DOI: 10.1088/1748-0221/3/08/S08001.
- [44] S. Chatrchyan et al. “The CMS Experiment at the CERN LHC”. In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.
- [45] *CMS Detector*. [http://www.physikblog.eu/wp-content/uploads/2008/09/cms\\_aufbau.png](http://www.physikblog.eu/wp-content/uploads/2008/09/cms_aufbau.png). Accessed: 2022-05-03.
- [46] *The CMS magnet project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: <https://cds.cern.ch/record/331056>.
- [47] R. Frühwirth and A. Strandlie. *Pattern Recognition, Tracking and Vertex Reconstruction in Particle Detectors*. Jan. 2021. ISBN: 978-3-030-65770-3. DOI: 10.1007/978-3-030-65771-0.
- [48] T. C. Collaboration. “Description and performance of track and primary-vertex reconstruction with the CMS tracker”. In: *Journal of Instrumentation* 9.10 (Oct. 2014), P10009–P10009. DOI: 10.1088/1748-0221/9/10/p10009. URL: <https://doi.org/10.1088/1748-0221/9/10/p10009>.
- [49] F. M. Hartmann. *Evolution of Silicon Sensor Technology in Particle Physics*. Cham, 2017. URL: <https://doi.org/10.1007/978-3-319-64436-3>.
- [50] A. Benaglia. *The CMS ECAL performance with examples*. Tech. rep. Geneva: CERN, Nov. 2013. DOI: 10.1088/1748-0221/9/02/C02008. URL: <https://cds.cern.ch/record/1632384>.
- [51] “Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at  $\sqrt{s} = 7$  TeV”. In: *Journal of Instrumentation* 8.09 (Sept. 2013), P09009–P09009. DOI: 10.1088/1748-0221/8/09/p09009. URL: <https://doi.org/10.1088/1748-0221/8/09/p09009>.
- [52] *The CMS hadron calorimeter project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: <http://cds.cern.ch/record/357153>.
- [53] J. G. Layter. *The CMS muon project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: <http://cds.cern.ch/record/343814>.
- [54] A. Sirunyan et al. “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at  $\sqrt{s} = 13$  TeV”. In: *Journal of Instrumentation* 13.06 (June 2018), P06015–P06015. DOI: 10.1088/1748-0221/13/06/p06015. URL: <https://doi.org/10.1088/1748-0221/13/06/p06015>.



- [55] S. Cittolin, A. Rácz, and P. Spiccas. *CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger*. CMS trigger and data-acquisition project. Technical design report. CMS. Geneva: CERN, 2002. URL: <http://cds.cern.ch/record/578006>.
- [56] A. Sirunyan et al. “Pileup mitigation at CMS in 13 TeV data”. In: *Journal of Instrumentation* 15.09 (Sept. 2020), P09018–P09018. DOI: 10.1088/1748-0221/15/09/p09018. URL: <https://doi.org/10.1088/1748-0221/15/09/p09018>.
- [57] O. Baerring. “Preparing CERN Tier-0 data centres for LHC Run 3”. In: *EPJ Web Conf.* 245 (2020), 07057. 6 p. DOI: 10.1051/epjconf/202024507057. URL: <https://cds.cern.ch/record/2753511>.
- [58] V. Khachatryan and A. S. et al. “The CMS trigger system”. In: *Journal of Instrumentation* 12.01 (Jan. 2017), P01020–P01020. DOI: 10.1088/1748-0221/12/01/p01020. URL: <https://doi.org/10.1088/1748-0221/12/01/p01020>.
- [59] A. Sirunyan et al. “Performance of the CMS Level-1 trigger in proton-proton collisions at  $\sqrt{s} = 13$  TeV”. In: *Journal of Instrumentation* 15.10 (Oct. 2020), P10017–P10017. DOI: 10.1088/1748-0221/15/10/p10017. URL: <https://doi.org/10.1088/1748-0221/15/10/p10017>.
- [60] T. Sjöstrand, S. Mrenna, and P. Skands. “A brief introduction to PYTHIA 8.1”. In: *Computer Physics Communications* 178.11 (June 2008), pp. 852–867. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2008.01.036. URL: <http://dx.doi.org/10.1016/j.cpc.2008.01.036>.
- [61] J. Alwall et al. “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”. In: *Journal of High Energy Physics* 2014.7 (July 2014). ISSN: 1029-8479. DOI: 10.1007/jhep07(2014)079. URL: [http://dx.doi.org/10.1007/JHEP07\(2014\)079](http://dx.doi.org/10.1007/JHEP07(2014)079).
- [62] P. Nason. “A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms”. In: *Journal of High Energy Physics* 2004.11 (Nov. 2004), pp. 040–040. DOI: 10.1088/1126-6708/2004/11/040. URL: <https://doi.org/10.1088/1126-6708/2004/11/040>.
- [63] S. Frixione, P. Nason, and C. Oleari. “Matching NLO QCD computations with parton shower simulations: the POWHEG method”. In: *Journal of High Energy Physics* 2007.11 (Nov. 2007), pp. 070–070. DOI: 10.1088/1126-6708/2007/11/070. URL: <https://doi.org/10.1088/1126-6708/2007/11/070>.
- [64] S. Alioli et al. “A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX”. In: *Journal of High Energy Physics* 2010.6 (June 2010). ISSN: 1029-8479. DOI: 10.1007/jhep06(2010)043. URL: [http://dx.doi.org/10.1007/JHEP06\(2010\)043](http://dx.doi.org/10.1007/JHEP06(2010)043).
- [65] S. Agostinelli et al. “Geant4—a simulation toolkit”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: [https://doi.org/10.1016/S0168-9002\(03\)01368-8](https://doi.org/10.1016/S0168-9002(03)01368-8). URL: <https://www.sciencedirect.com/science/article/pii/S0168900203013688>.
- [66] A. Sirunyan et al. “Particle-flow reconstruction and global event description with the CMS detector”. In: *Journal of Instrumentation* 12.10 (Oct. 2017), P10003–P10003. DOI: 10.1088/1748-0221/12/10/p10003. URL: <https://doi.org/10.1088/1748-0221/12/10/p10003>.

- [67] M. Cacciari, G. P. Salam, and G. Soyez. “The anti-ktjet clustering algorithm”. In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 063–063. ISSN: 1029-8479. DOI: 10.1088/1126-6708/2008/04/063. URL: <http://dx.doi.org/10.1088/1126-6708/2008/04/063>.
- [68] *Jet energy scale and resolution performance with 13 TeV data collected by CMS in 2016-2018*. Apr. 2020. URL: <https://cds.cern.ch/record/2715872>.
- [69] V. Khachatryan et al. “Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV”. In: *Journal of Instrumentation* 12.02 (Feb. 2017), P02014–P02014. DOI: 10.1088/1748-0221/12/02/p02014. URL: <https://doi.org/10.1088/1748-0221/12/02/p02014>.
- [70] *Introduction to Jet Energy Corrections at CMS*. <https://twiki.cern.ch/twiki/bin/view/CMS/IntroToJEC>. Accessed: 2022-04-19.
- [71] The CMS collaboration. “Determination of jet energy calibration and transverse momentum resolution in CMS”. In: *Journal of Instrumentation* 6.11 (Nov. 2011), P11002–P11002. DOI: 10.1088/1748-0221/6/11/p11002. URL: <https://doi.org/10.1088/1748-0221/6/11/p11002>.
- [72] A. Sirunyan et al. “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV”. In: *Journal of Instrumentation* 13.05 (May 2018), P05011–P05011. DOI: 10.1088/1748-0221/13/05/p05011. URL: <https://doi.org/10.1088/1748-0221/13/05/p05011>.
- [73] *Procedure for the LHC Higgs boson search combination in Summer 2011*. Tech. rep. Geneva: CERN, Aug. 2011. URL: <https://cds.cern.ch/record/1379837>.
- [74] J. Neyman and E. S. Pearson. “On the Problem of the Most Efficient Tests of Statistical Hypotheses”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231 (1933), pp. 289–337. ISSN: 02643952. URL: <http://www.jstor.org/stable/91247> (visited on 05/04/2022).
- [75] S. S. Wilks. “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses”. In: *The Annals of Mathematical Statistics* 9.1 (1938), pp. 60–62. DOI: 10.1214/aoms/1177732360. URL: <https://doi.org/10.1214/aoms/1177732360>.
- [76] G. Cowan et al. “Asymptotic formulae for likelihood-based tests of new physics”. In: *The European Physical Journal C* 71.2 (Feb. 2011). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-011-1554-0. URL: <http://dx.doi.org/10.1140/epjc/s10052-011-1554-0>.
- [77] A. L. Read. “Presentation of search results: the CLs technique”. In: *Journal of Physics G: Nuclear and Particle Physics* 28.10 (Sept. 2002), pp. 2693–2704. DOI: 10.1088/0954-3899/28/10/313. URL: <https://doi.org/10.1088/0954-3899/28/10/313>.
- [78] W. S. McCulloch and W. Pitts. *A logical calculus of the ideas immanent in nervous activity*. 1943. URL: <https://doi.org/10.1007/2FBF02478259>.
- [79] S. Ruder. *An overview of gradient descent optimization algorithms*. 2016. DOI: 10.48550/ARXIV.1609.04747. URL: <https://arxiv.org/abs/1609.04747>.
- [80] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (Oct. 1986), pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0. URL: <https://doi.org/10.1038/323533a0>.

- [81] A. E. Bickle. *Fundamentals of graph theory*. Pure and applied undergraduate texts ; 43The Sally series. Literaturverzeichnis: Seite 313-328. Providence, Rhode Island: American Mathematical Society, [2020]. ISBN: 97814704534289781470455491. URL: <https://www.gbv.de/dms/tib-ub-hannover/1684125456.pdf>.
- [82] J. Shlomi, P. Battaglia, and J.-R. Vlimant. “Graph neural networks in particle physics”. In: *Machine Learning: Science and Technology* 2.2 (Jan. 2021), p. 021001. ISSN: 2632-2153. DOI: 10.1088/2632-2153/abfb9a. URL: <http://dx.doi.org/10.1088/2632-2153/abfb9a>.
- [83] J. Zhou et al. “Graph neural networks: A review of methods and applications”. In: *AI Open* 1 (2020), pp. 57–81. ISSN: 2666-6510. DOI: <https://doi.org/10.1016/j.aiopen.2021.01.001>. URL: <https://www.sciencedirect.com/science/article/pii/S2666651021000012>.
- [84] T. N. Kipf and M. Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. 2017. arXiv: 1609.02907 [cs.LG].
- [85] Y. Feng. “A New Deep-Neural-Network-Based Missing Transverse Momentum Estimator, and its Application to W Recoil”. PhD thesis. University of Maryland, 2020. URL: <https://cds.cern.ch/record/2744871/files/CERN%20THESIS%202020%20194.pdf>.
- [86] *MiniAOD Analysis Documentation*. <https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookMiniAOD2015>. Accessed: 2022-05-05.
- [87] A. Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [88] M. Fey and J. E. Lenssen. “Fast Graph Representation Learning with PyTorch Geometric”. In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*. 2019.
- [89] Y. Wang et al. *Dynamic Graph CNN for Learning on Point Clouds*. 2018. DOI: 10.48550/ARXIV.1801.07829. URL: <https://arxiv.org/abs/1801.07829>.
- [90] C. Nwankpa et al. *Activation Functions: Comparison of trends in Practice and Research for Deep Learning*. 2018. arXiv: 1811.03378 [cs.LG].
- [91] *Expected pileup values at the HL-LHC*. Tech. rep. Geneva: CERN, Sept. 2013. URL: <https://cds.cern.ch/record/1604492>.
- [92] D. Bertolini et al. “Pileup per particle identification”. In: *Journal of High Energy Physics* 2014.10 (Oct. 2014). ISSN: 1029-8479. DOI: 10.1007/jhep10(2014)059. URL: [http://dx.doi.org/10.1007/JHEP10\(2014\)059](http://dx.doi.org/10.1007/JHEP10(2014)059).
- [93] *Search for Dark Matter produced in association with bottom quarks*. Tech. rep. Geneva: CERN, 2016. URL: <https://cds.cern.ch/record/2138506>.
- [94] M. Klasen, M. Pohl, and G. Sigl. “Indirect and direct search for dark matter”. In: *Progress in Particle and Nuclear Physics* 85 (Nov. 2015), pp. 1–32. DOI: 10.1016/j.pnpnp.2015.07.001. URL: <https://doi.org/10.1016%2Fj.pnpnp.2015.07.001>.
- [95] D. Abercrombie et al. “Dark Matter benchmark models for early LHC Run-2 Searches: Report of the ATLAS/CMS Dark Matter Forum”. In: *Physics of the Dark Universe* 27 (Jan. 2020), p. 100371. DOI: 10.1016/j.dark.2019.100371. URL: <https://doi.org/10.1016%2Fj.dark.2019.100371>.

- [96] E. Morgante. *Simplified Dark Matter Models*. 2018. DOI: 10.48550/ARXIV.1804.01245. URL: <https://arxiv.org/abs/1804.01245>.
- [97] J. Andrea, B. Fuks, and F. Maltoni. “Monotops at the LHC”. In: *Physical Review D* 84.7 (Oct. 2011). DOI: 10.1103/physrevd.84.074025. URL: <https://doi.org/10.1103/physrevd.84.074025>.
- [98] J.-L. Agram et al. “Monotop phenomenology at the Large Hadron Collider”. In: *Physical Review D* 89.1 (Jan. 2014). DOI: 10.1103/physrevd.89.014028. URL: <https://doi.org/10.1103/physrevd.89.014028>.
- [99] I. Boucheneb et al. “Revisiting monotop production at the LHC”. In: *Journal of High Energy Physics* 2015.1 (Jan. 2015). DOI: 10.1007/jhep01(2015)017. URL: [https://doi.org/10.1007/jhep01\(2015\)017](https://doi.org/10.1007/jhep01(2015)017).
- [100] *Feynman diagram maker*. <https://www.aidansean.com/feynman/>. Accessed: 2022-04-06.
- [101] S. L. Glashow, J. Iliopoulos, and L. Maiani. “Weak Interactions with Lepton-Hadron Symmetry”. In: *Phys. Rev. D* 2 (7 Oct. 1970), pp. 1285–1292. DOI: 10.1103/PhysRevD.2.1285. URL: <https://link.aps.org/doi/10.1103/PhysRevD.2.1285>.
- [102] *CMS recommended muon high level triggers*. <https://twiki.cern.ch/twiki/bin/viewauth/CMS/MuonHLT2018>. Accessed: 2022-04-20.
- [103] *CMS recommended electron high level triggers*. <https://twiki.cern.ch/twiki/bin/view/CMS/EgHLTRunIISummary>. Accessed: 2022-04-20.
- [104] D. R. Tovey. “Transformation properties of the transverse mass under transverse Lorentz boosts at hadron colliders”. In: *The European Physical Journal C* 79.4 (Apr. 2019), p. 294. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-019-6813-5. URL: <https://doi.org/10.1140/epjc/s10052-019-6813-5>.
- [105] R. MacCoun and S. Perlmutter. “Blind analysis: Hide results to seek the truth”. In: *Nature* 526 (Oct. 2015), pp. 187–189. DOI: 10.1038/526187a.
- [106] W. T. Giele, S. A. Keller, and D. A. Kosower. *Parton Distribution Function Uncertainties*. 2001. DOI: 10.48550/ARXIV.HEP-PH/0104052. URL: <https://arxiv.org/abs/hep-ph/0104052>.
- [107] C. G. Willis et al. “New method for reducing parton distribution function uncertainties in the high-mass Drell-Yan spectrum”. In: *Phys. Rev. D* 99 (5 Mar. 2019), p. 054004. DOI: 10.1103/PhysRevD.99.054004. URL: <https://link.aps.org/doi/10.1103/PhysRevD.99.054004>.
- [108] R. Barlow and C. Beeston. “Fitting using finite Monte Carlo samples”. In: *Computer Physics Communications* 77.2 (1993), pp. 219–228. ISSN: 0010-4655. DOI: [https://doi.org/10.1016/0010-4655\(93\)90005-W](https://doi.org/10.1016/0010-4655(93)90005-W). URL: <https://www.sciencedirect.com/science/article/pii/001046559390005W>.
- [109] J. Conway. “Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra”. In: (Mar. 2011). Comments: Presented at PHYSTAT 2011, CERN, Geneva, Switzerland, January 2011, to be published in a CERN Yellow Report, 115–120. 6 p. DOI: 10.5170/CERN-2011-006.115. arXiv: 1103.0354. URL: <https://cds.cern.ch/record/1333496>.
- [110] *Higgs-Analysis-CombinedLimit: Automatic statistical uncertainties*. <https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/part2/bin-wise-stats/>. Accessed: 2022-05-05.

- [111] A. Sirunyan, A. Tumasyan, and et al. “Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC”. In: *Journal of Instrumentation* 16.05 (May 2021), P05014. DOI: 10.1088/1748-0221/16/05/p05014. URL: <https://doi.org/10.1088/1748-0221/16/05/p05014>.
- [112] J. Bellm et al. “Parton-shower uncertainties with Herwig 7: benchmarks at leading order”. In: *The European Physical Journal C* 76.12 (Dec. 2016). DOI: 10.1140/epjc/s10052-016-4506-x. URL: <https://doi.org/10.1140/epjc/s10052-016-4506-x>.
- [113] S. Mrenna and P. Skands. “Automated parton-shower variations in pythia 8”. In: *Physical Review D* 94.7 (Oct. 2016). DOI: 10.1103/physrevd.94.074005. URL: <https://doi.org/10.1103/physrevd.94.074005>.
- [114] *CMS luminosity measurement for the 2018 data-taking period at  $\sqrt{s} = 13$  TeV*. Tech. rep. Geneva: CERN, 2019. URL: <https://cds.cern.ch/record/2676164>.



# Appendix

## A Distributions of GraphMET Input Variables

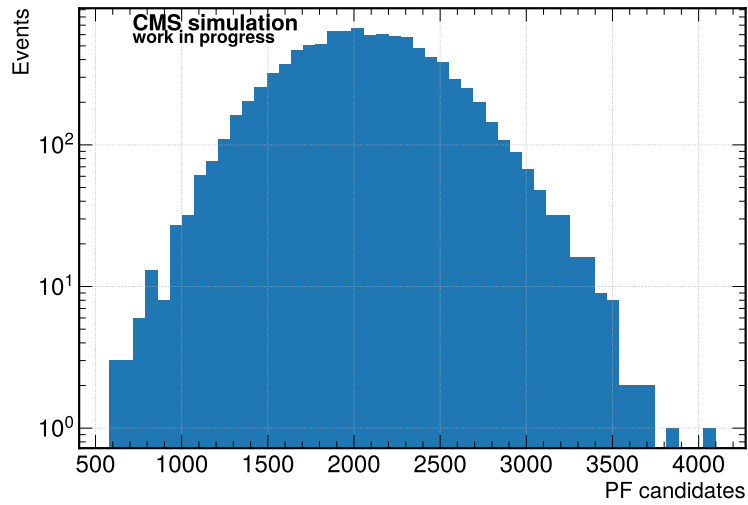
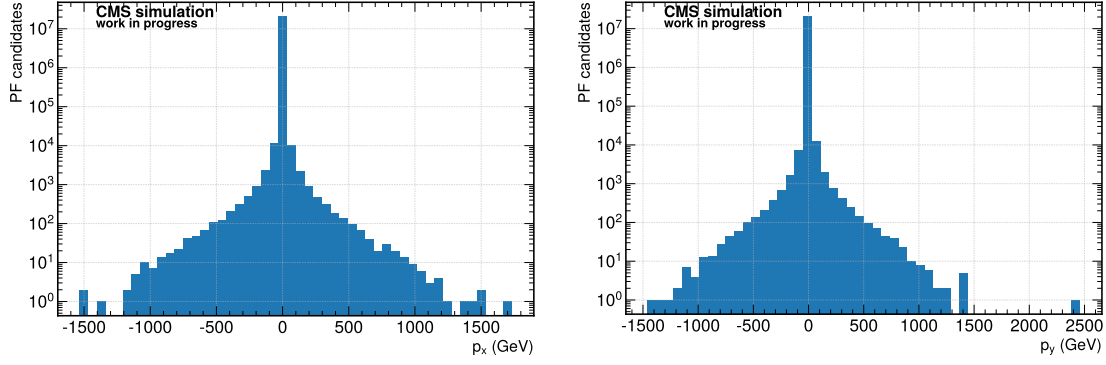
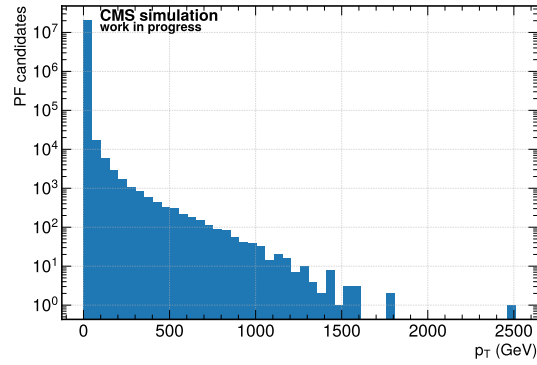


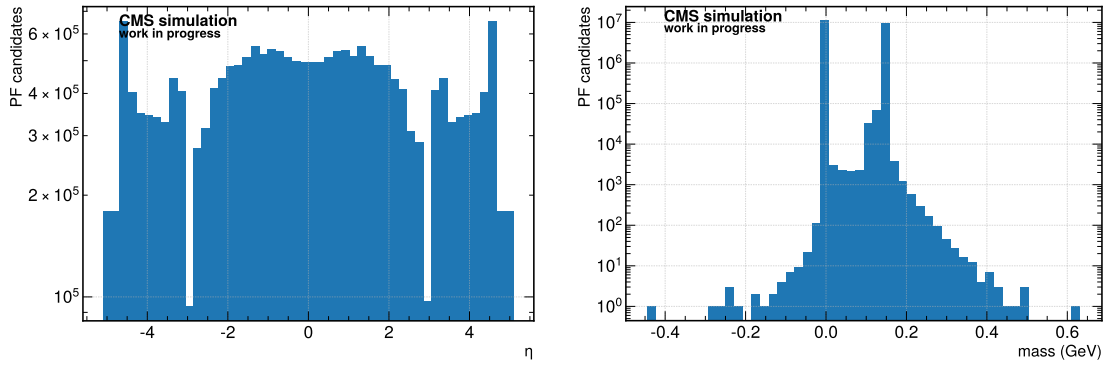
Figure A.1: Distribution of the number of PF candidates per event in the training data set for 10 000 events. With only few exceptions, all events contain between 1000 and 3000 PF candidates. This is not an input variable to GraphMET but the number of PF degoods the number of vertices of an event’s graph.



(a) Components of the transverse momentum.



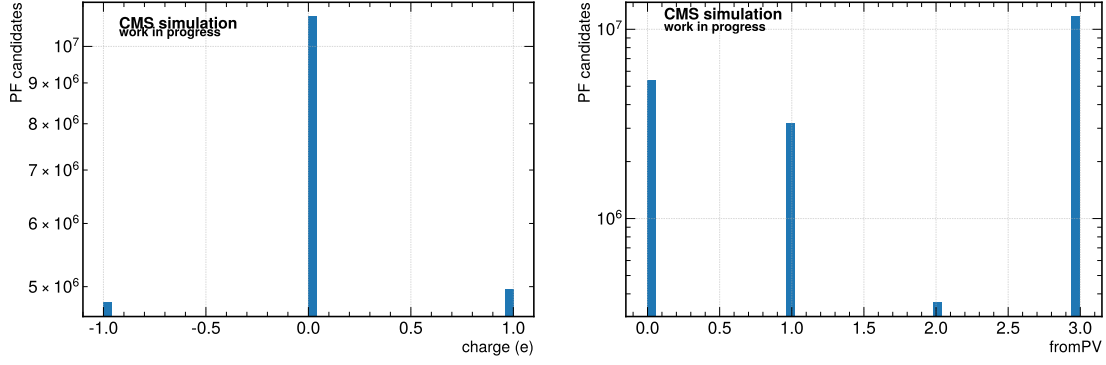
(b) Transverse momentum distribution.



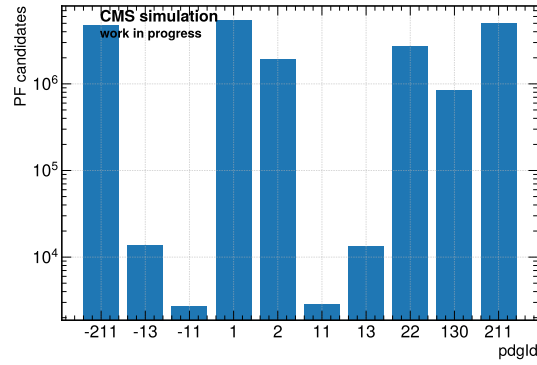
(c) Pseudorapidity (left) and mass (right).

Figure A.2: Kinematic distributions of the PF candidates in the training data set for 10 000 events. Most PF candidates have little transverse momentum with single outliers containing up to 1800 GeV. The pseudorapidity distribution shows a dip at roughly  $\eta = 3$ , where the transition between the end caps and the HF of the detector is located. The mass distribution shows three crowded bins: The bin at 0 GeV contains photons and electrons, the bin at 106 MeV contains muons, and the bin at 135 MeV pions. There are a few PF candidates with negative mass, which is a result of reconstruction errors.

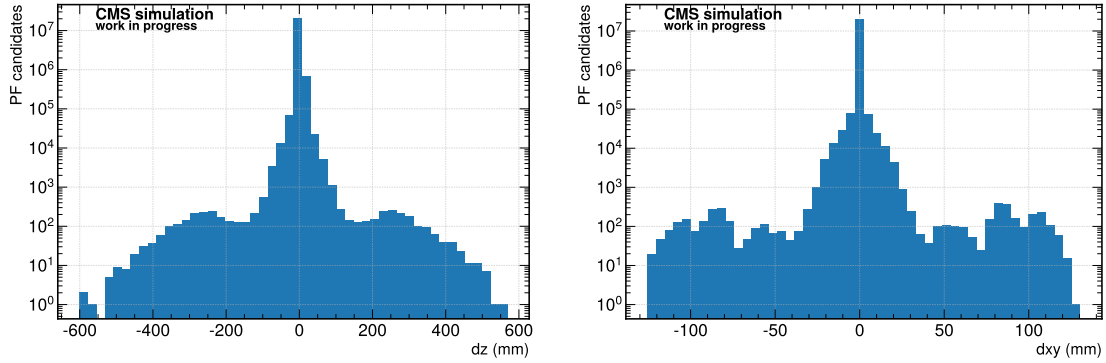




(a) Charge (left) and fromPV (right).



(b) pdgId.



(c) Impact parameter in longitudinal (left) and transverse direction (right).

Figure A.3: Further distributions of the PF candidates in the training data set for 10 000 events. About one half of the PF candidates are not charged, the other half has one positive or negative elementary charge. About one half of the PF is used in the fit of the leading vertex ( $\text{fromPV} = 3$ ) or closest to the leading vertex ( $\text{fromPV} = 2$ ). The event is dominated by hadrons ( $\text{pdgId} = 1, 2, 130, \pm 210$ ) in comparison to electrons ( $\text{pdgId} = \pm 11$ ), muons ( $\text{pdgId} = \pm 13$ ) and photons ( $\text{pdgId} = 22$ ). The impact parameter distribution shows that most PF candidates are close to the interaction point whereas a couple of hundred PF candidates are dislocated by more than 10 cm along the  $z$  axis.



## C Unscaled Resolution Plots

Figure C.1 shows the resolution plots before the application of the response scaling. As the scaling worsens the resolution of GraphMET and DeepMETResolutionTune due to their response behavior, the unscaled resolutions are even better for these two MET reconstruction methods.

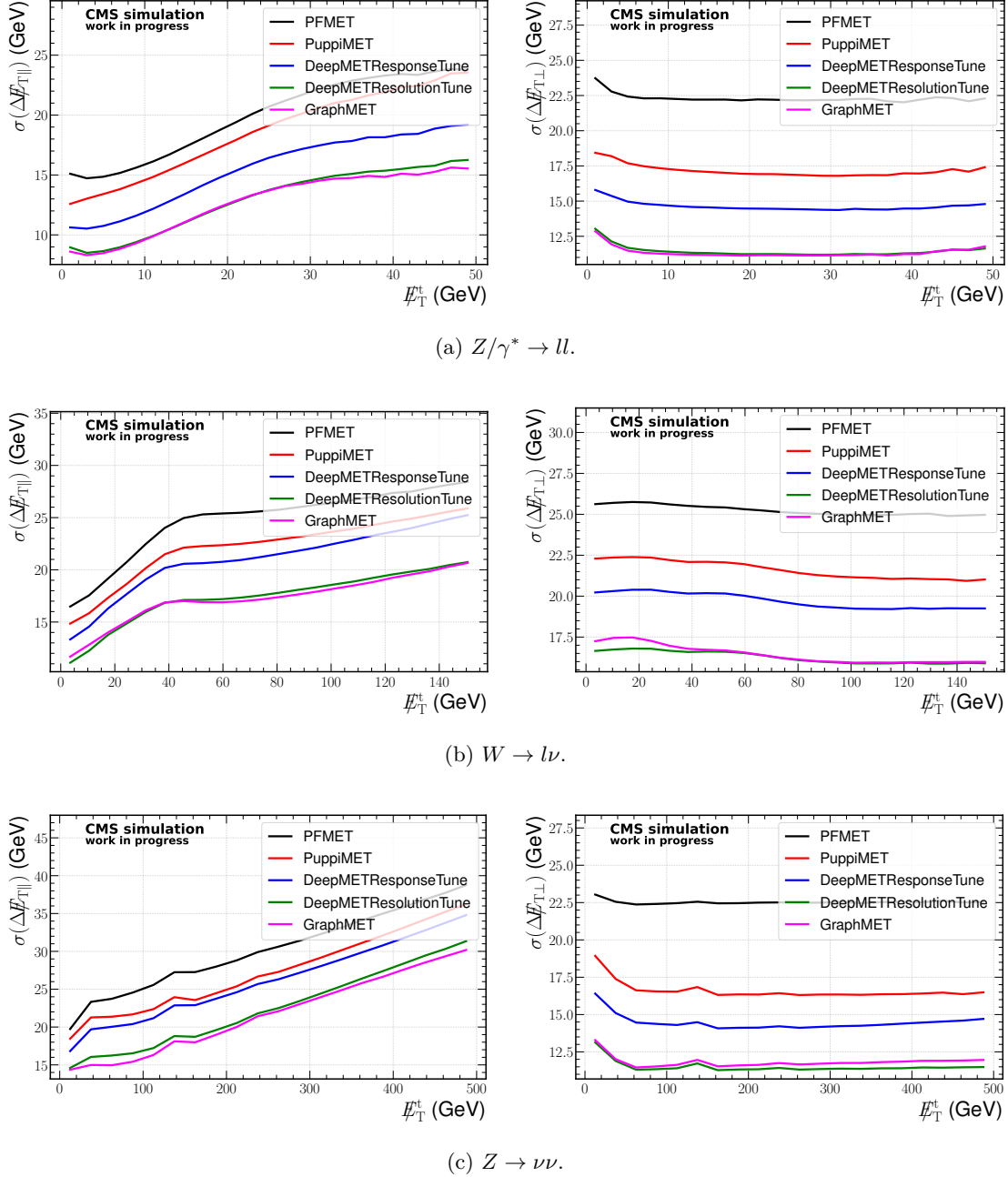
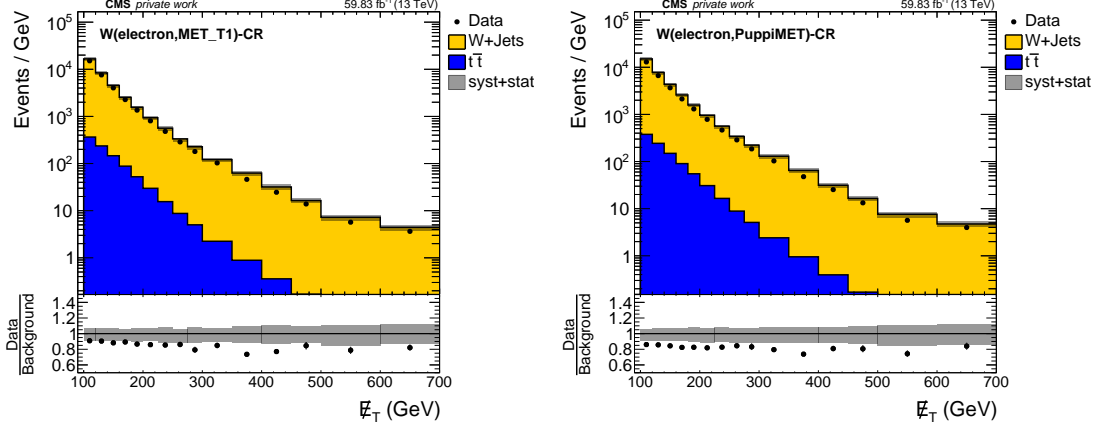


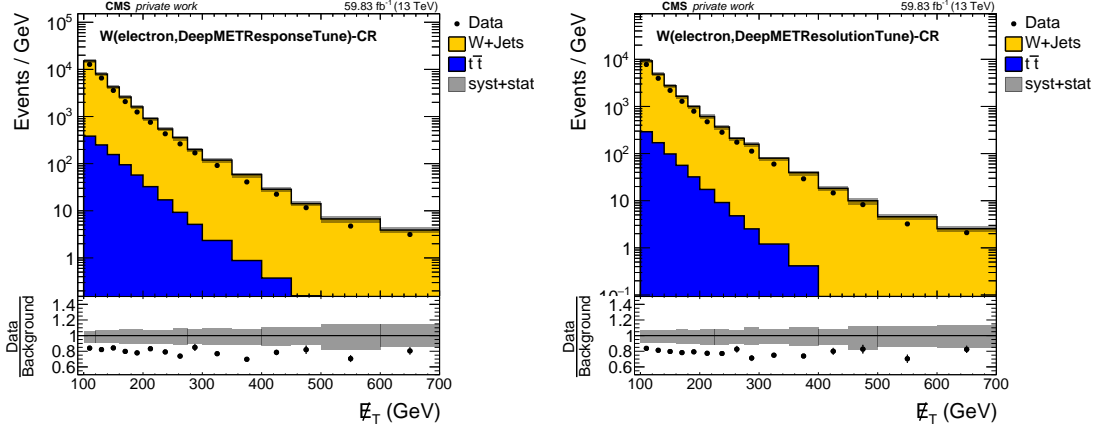
Figure C.1: Parallel (left) and perpendicular (right) resolution plots. In general, GraphMET and DeepMETResolutionTune show similar behavior with the smallest resolution values throughout all MET regions. The parallel resolution tends to increase with true MET in all samples. The perpendicular resolution is consistent in different MET regions.

## D Further Control Region and Signal Region Plots

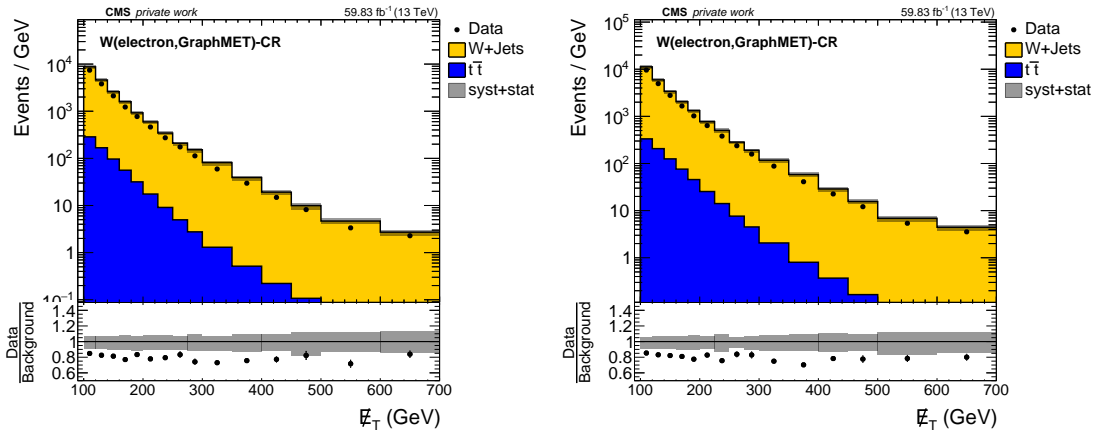
### D.1 $W$ Electron CR: MET



(a) MET-Type I (left) and PUPPI MET (right).

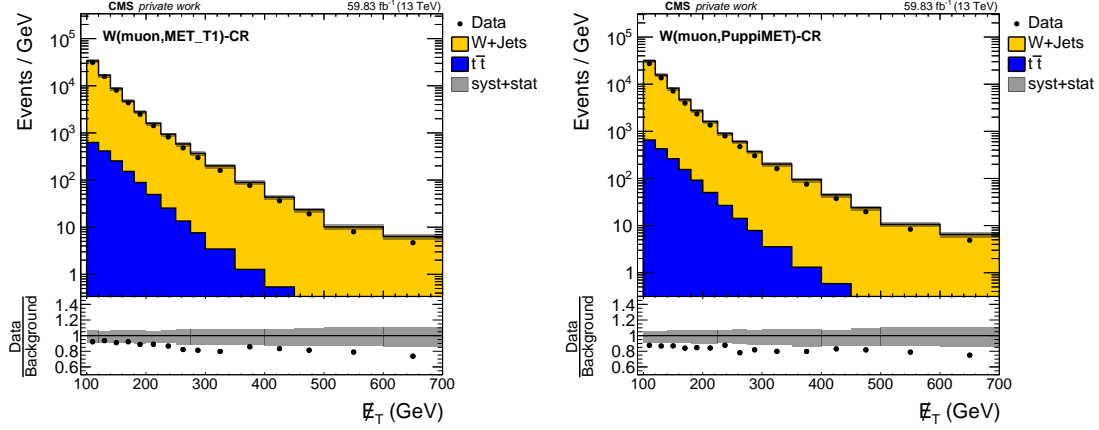


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

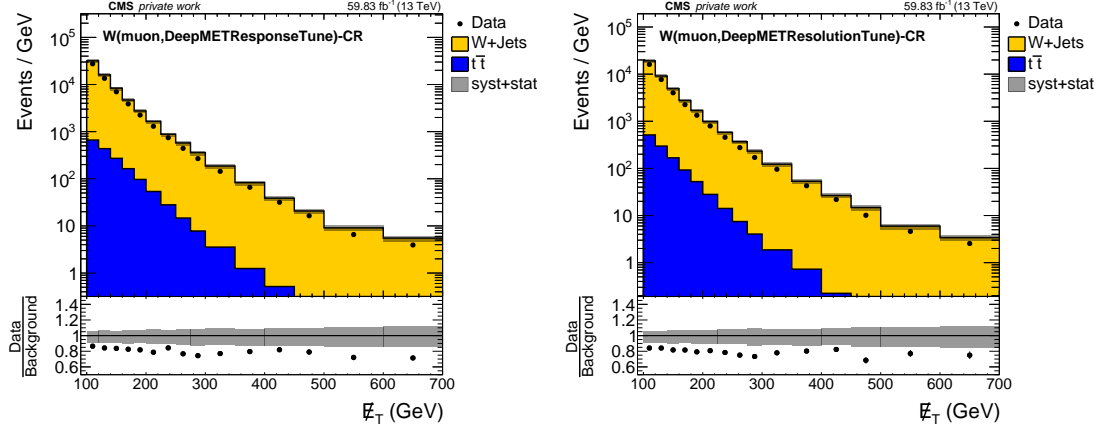


(c) GraphMET (left) and naively calibrated GraphMET (right).

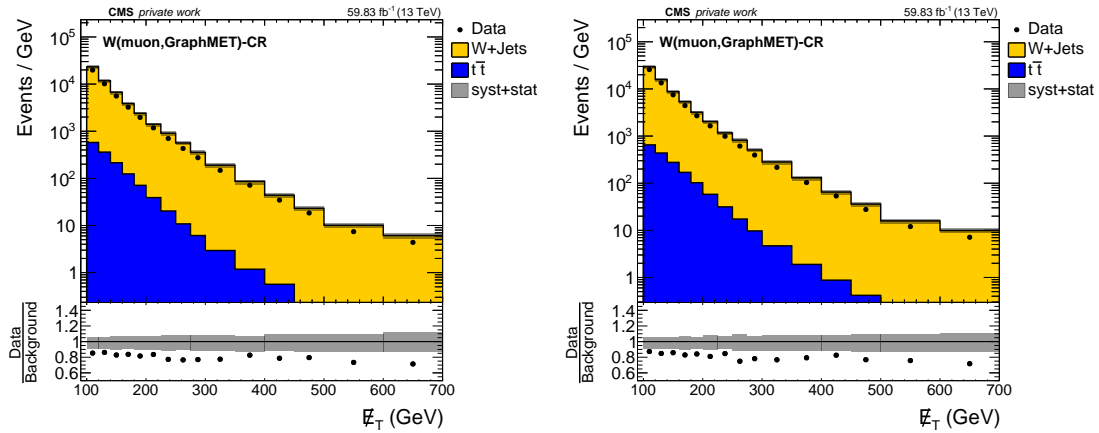
Figure D.1: MET distributions in the  $W$  CR with an electron in the final state. In general, the plots show an MC overshoot of roughly 10–20 % for all MET reconstruction methods. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods. The shapes of MET look similar when using the different reconstruction methods.

D.2  $W$  Muon CR: MET

(a) MET-Type I (left) and PUPPI MET (right).

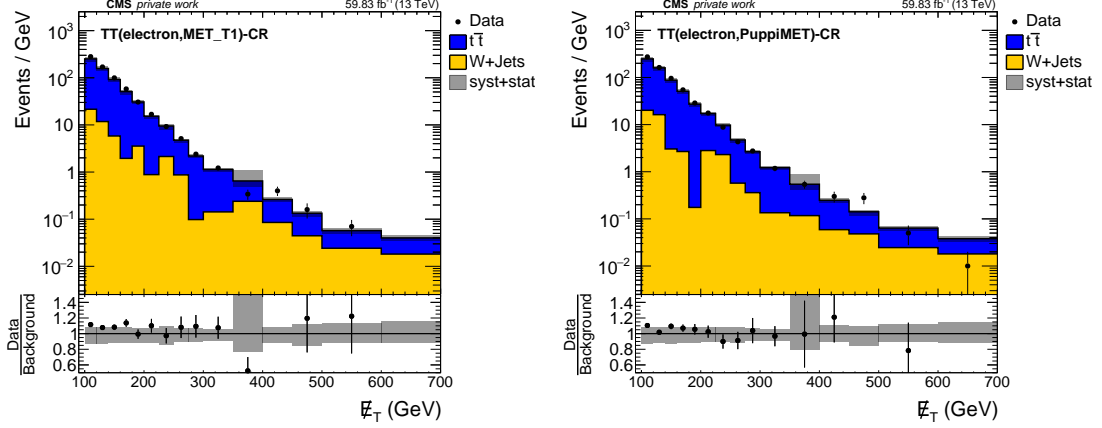


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

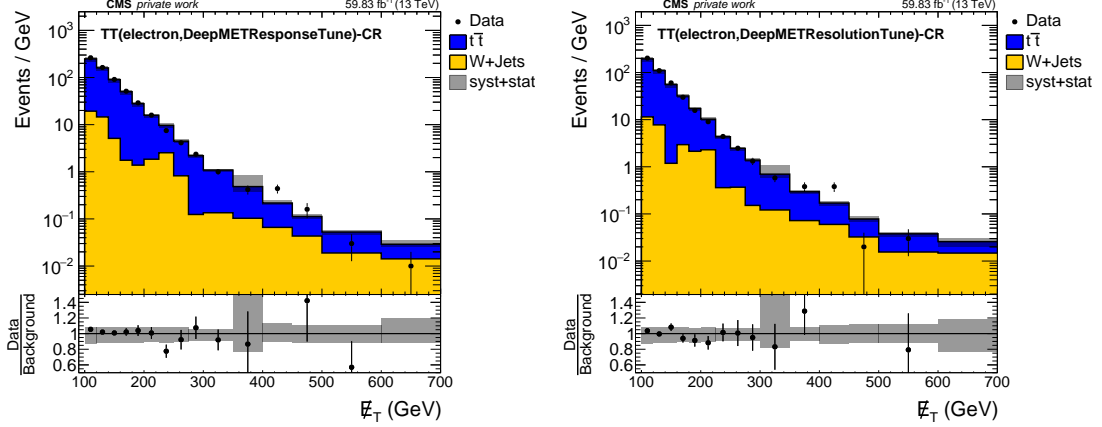


(c) GraphMET (left) and naively calibrated GraphMET (right).

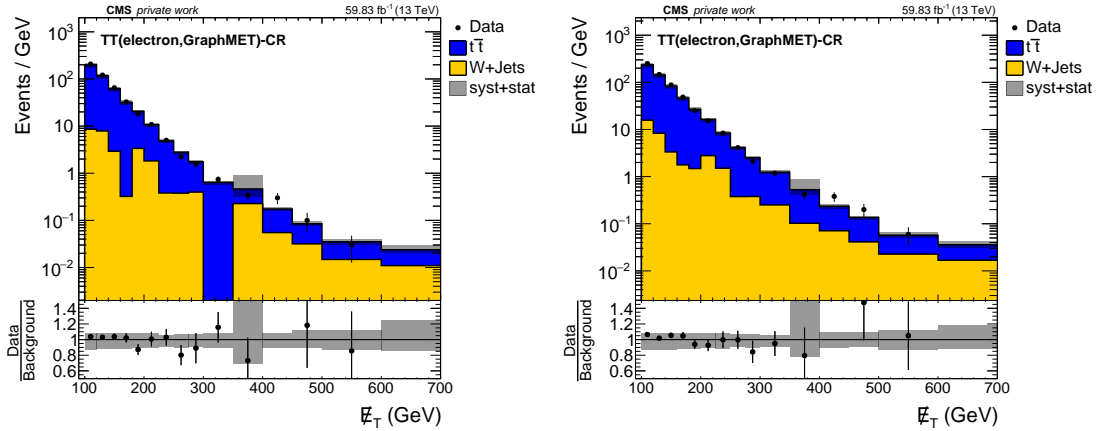
Figure D.2: MET distributions in the  $W$  CR with a muon in the final state. In general, the plots show an MC overshoot of roughly 10 – 20 % for all MET reconstruction methods. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods. The shapes of MET look similar when using the different reconstruction methods.

D.3  $t\bar{t}$  Electron CR: MET

(a) MET-Type I (left) and PUPPI MET (right).

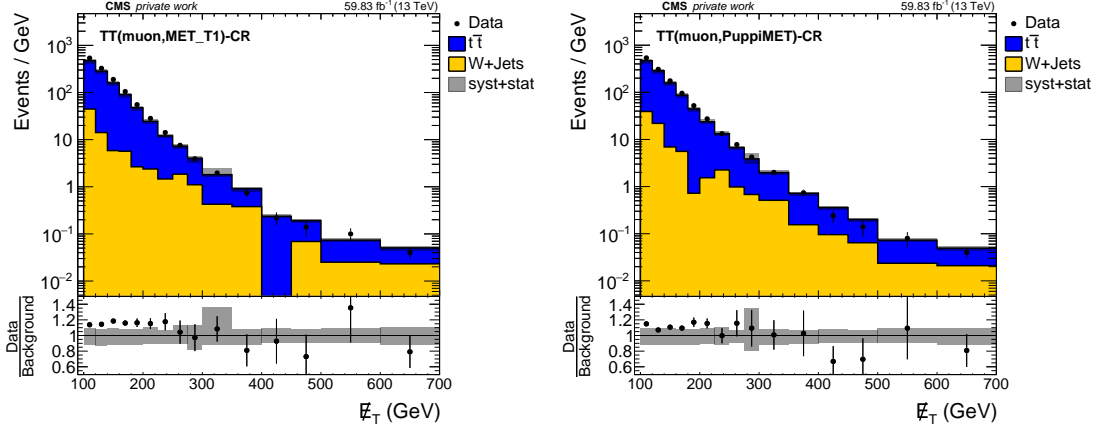


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

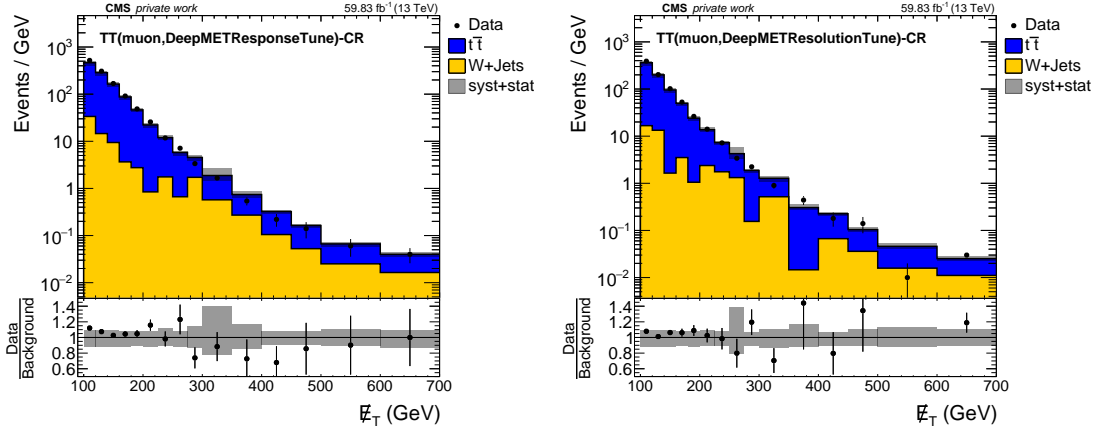


(c) GraphMET (left) and naively calibrated GraphMET (right).

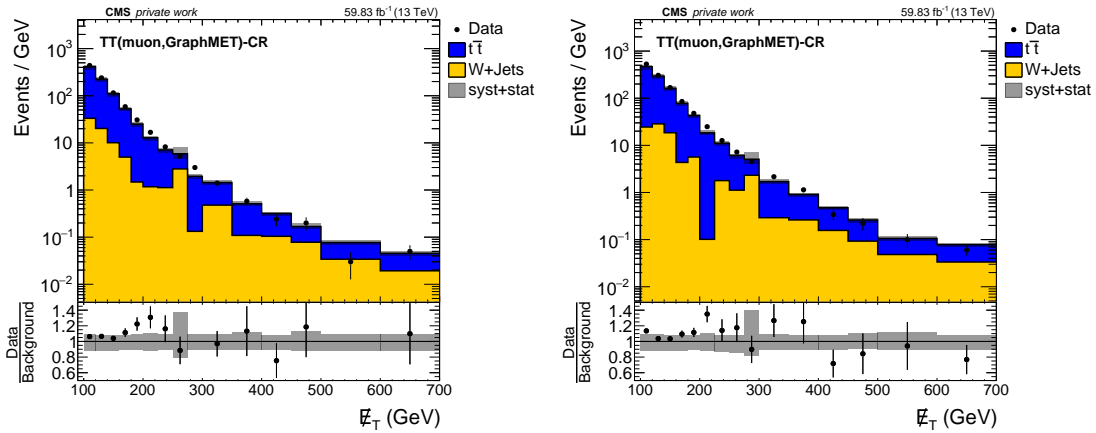
Figure D.3: MET distributions in the  $t\bar{t}$  CR with an electron in the final state. In general, the plots show a good MC description of the data points for all MET reconstruction methods. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods. The shapes of MET look similar when using the different reconstruction methods.

D.4  $t\bar{t}$  Muon CR: MET

(a) MET-Type I (left) and PUPPI MET (right).



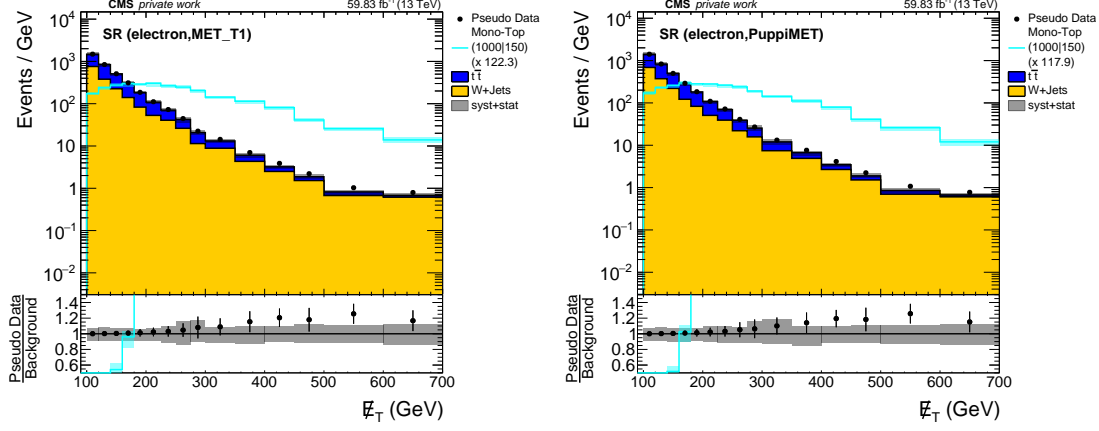
(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).



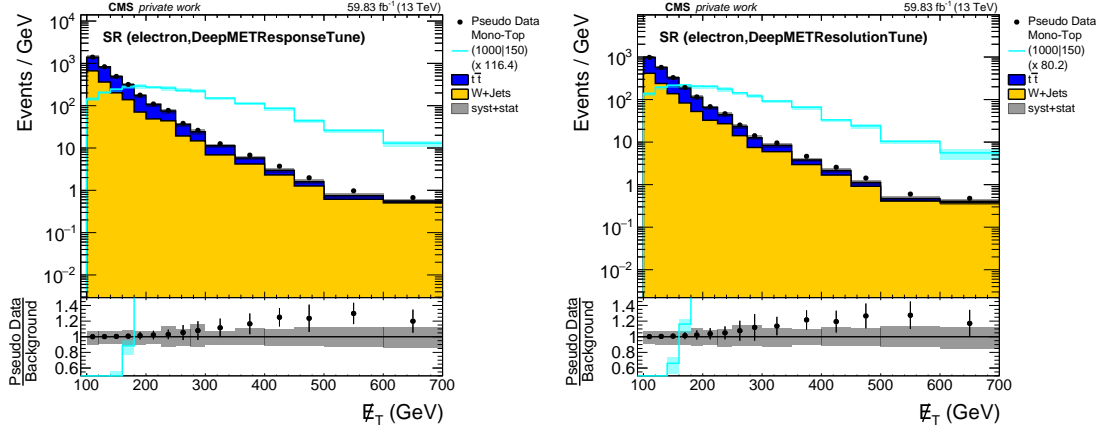
(c) GraphMET (left) and naively calibrated GraphMET (right).

Figure D.4: MET distributions in the  $t\bar{t}$  CR with a muon in the final state. Except for the first bins of MET-Type-I and PUPPI MET, the plots show good MC agreement of the data points for all MET reconstruction methods. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods. The shapes of MET look similar when using the different reconstruction methods.

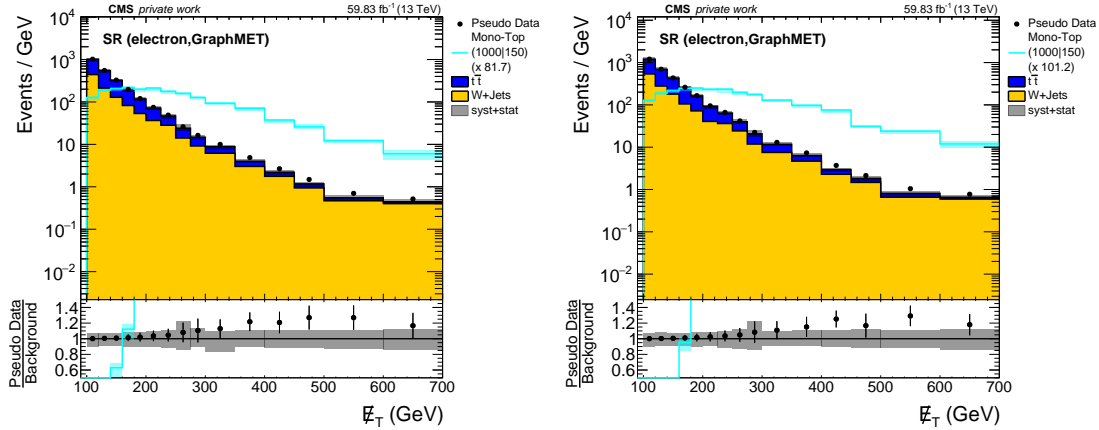
## D.5 Electron SR: MET



(a) MET-Type I (left) and PUPPI MET (right).



(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).



(c) GraphMET (left) and naively calibrated GraphMET (right).

Figure D.5: MET distributions in the SR with an electron in the final state. The signal to background ratio is best for DeepMETResolutionTune and GraphMET as is inferred from the smaller multiplication factor of the signal. The pseudodata to background ratio looks similar for all reconstruction methods.



## D.6 Muon SR: MET

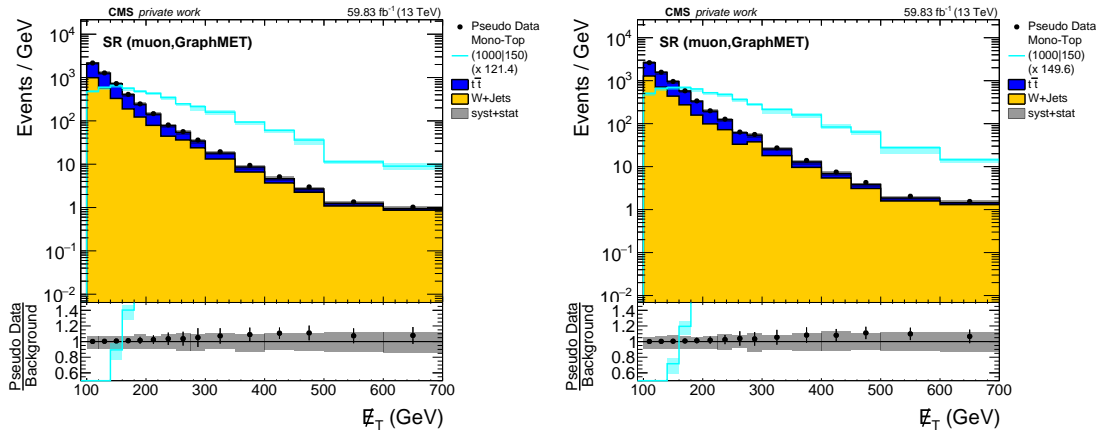
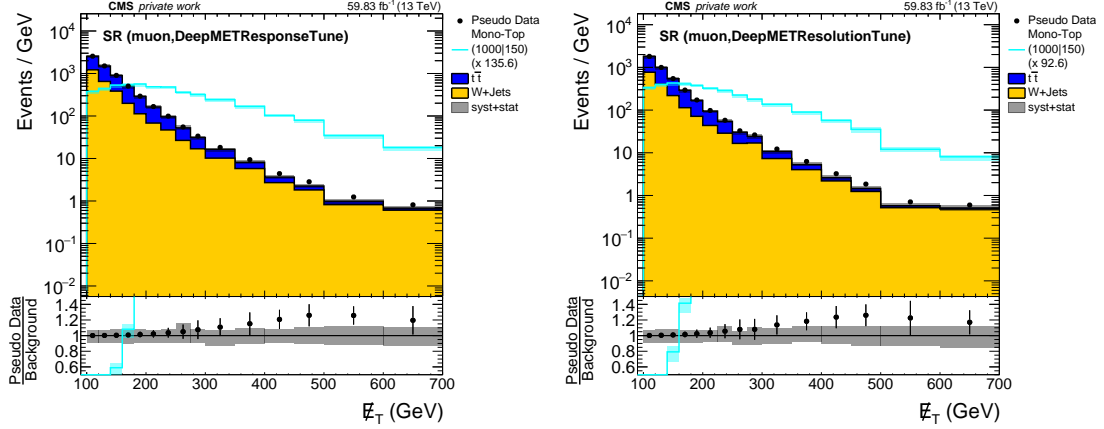
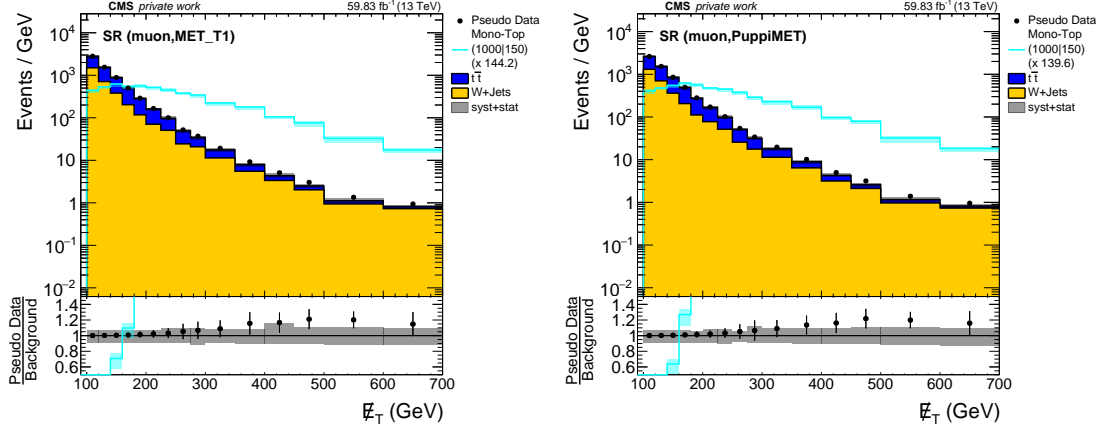
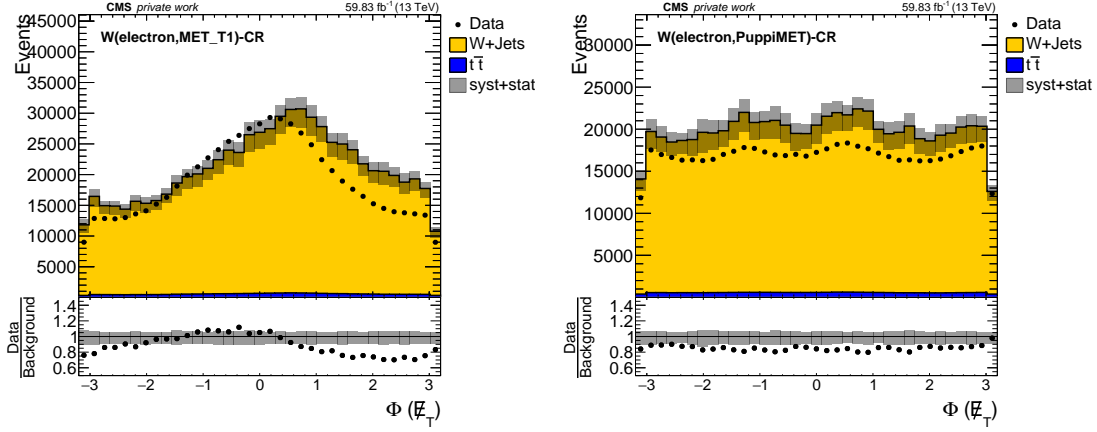
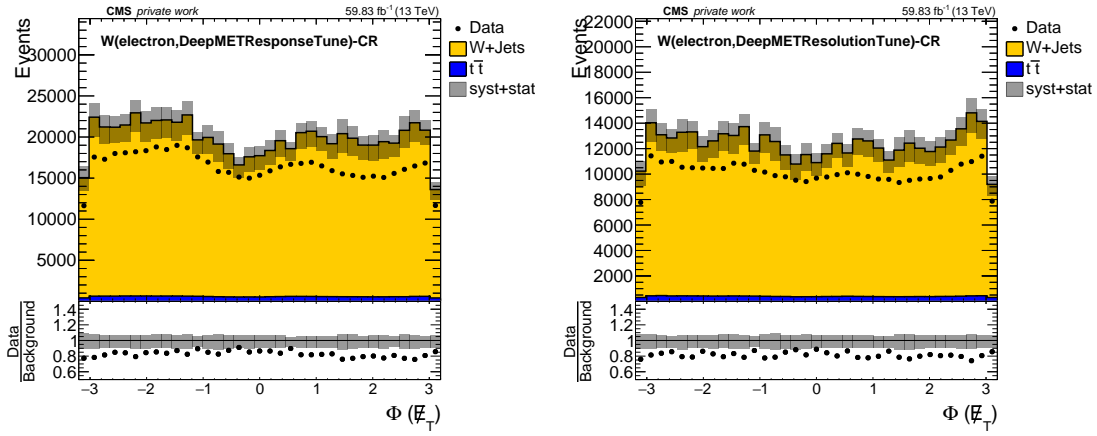


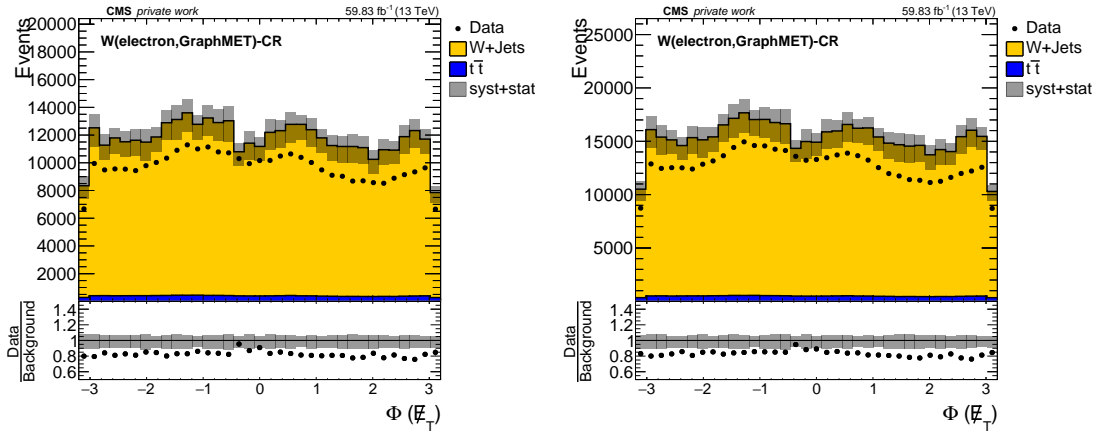
Figure D.6: MET distributions in the SR with a muon in the final state. The signal to background ratio is best for DeepMETResolutionTune as is inferred from the smaller multiplication factor of the signal. The pseudodata to background ratio looks similar for the four upper plots and considerably worse for the versions of GraphMET.

D.7  $W$  Electron CR:  $\phi(\vec{\cancel{E}}_{\text{T}})$ 

(a) MET-Type I (left) and PUPPI MET (right).

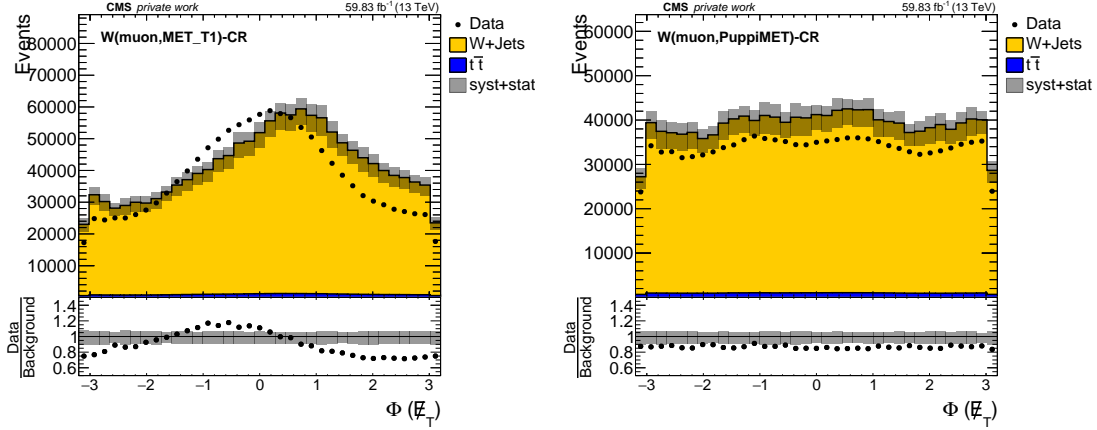


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

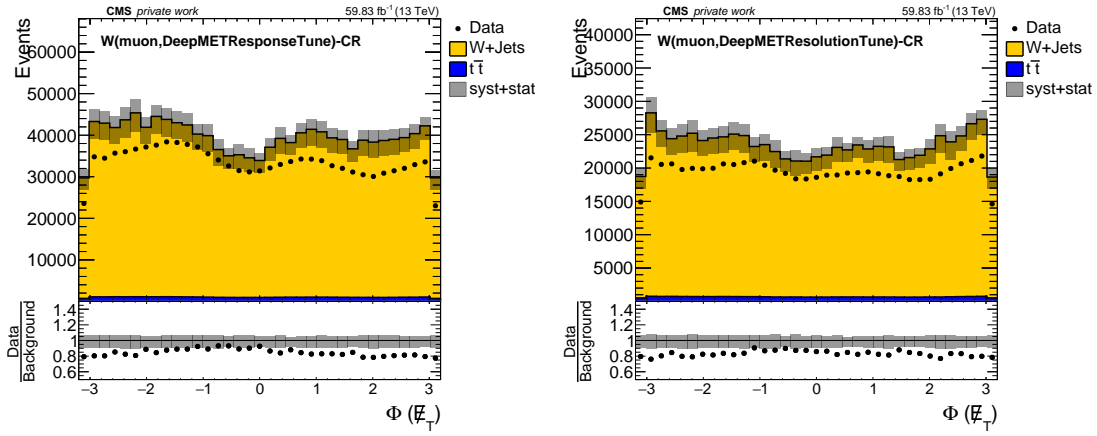


(c) GraphMET (left) and naively calibrated GraphMET (right).

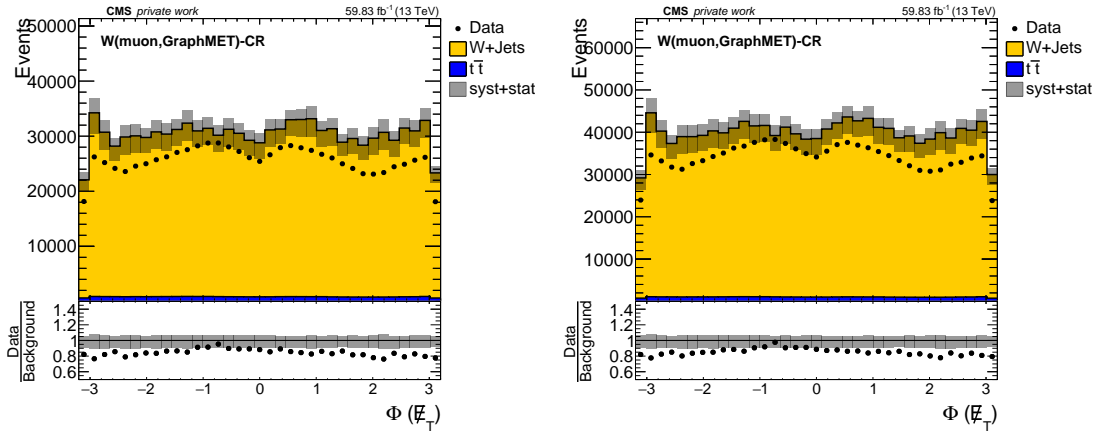
Figure D.7:  $\phi(\vec{\cancel{E}}_{\text{T}})$  distributions in the  $W$  CR with an electron in the final state. Except for MET-Type-I, which shows a modulation in the data over background ratio, the plots only show an MC offset of about 10 – 20%. Also except for MET-Type-I, the shapes look similar with some modulation in data which is described by MC except for said offset. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods.

D.8  $W$  Muon CR:  $\phi(\vec{\cancel{E}}_T)$ 

(a) MET-Type I (left) and PUPPI MET (right).



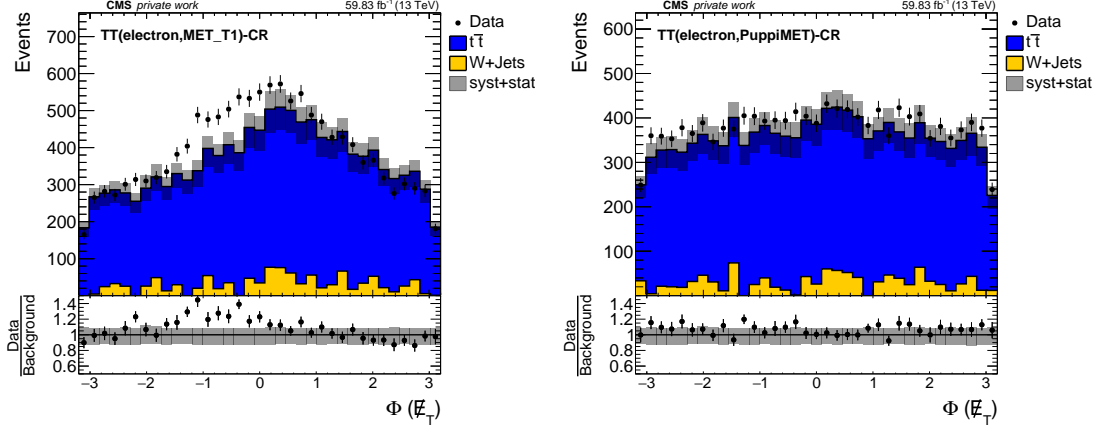
(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).



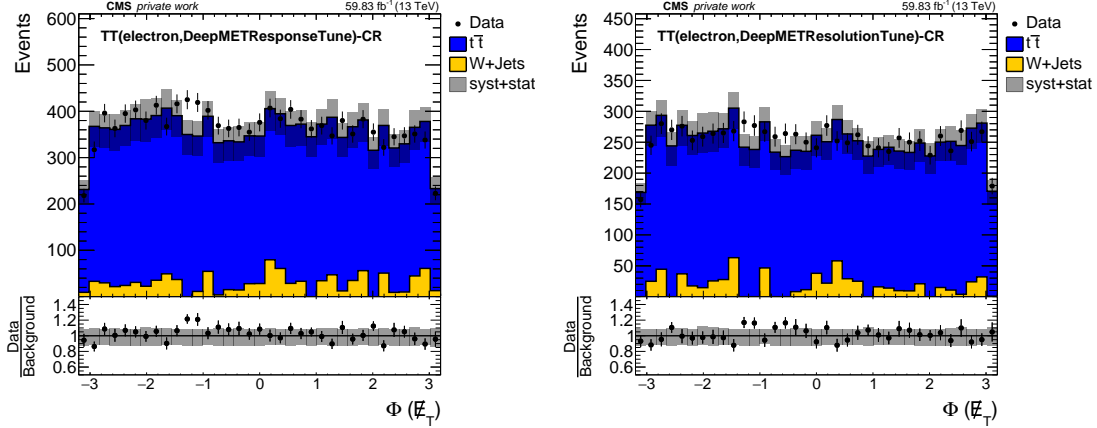
(c) GraphMET (left) and naively calibrated GraphMET (right).

Figure D.8:  $\phi(\vec{\cancel{E}}_T)$  distributions in the  $W$  CR with a muon in the final state. Except for MET-Type-I, which shows a modulation in the data over background ratio, the plots only show an MC offset of about 10 – 20%. Again, except for MET-Type-I, the shapes look similar with some modulation in data which is described by MC except for said offset. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods.

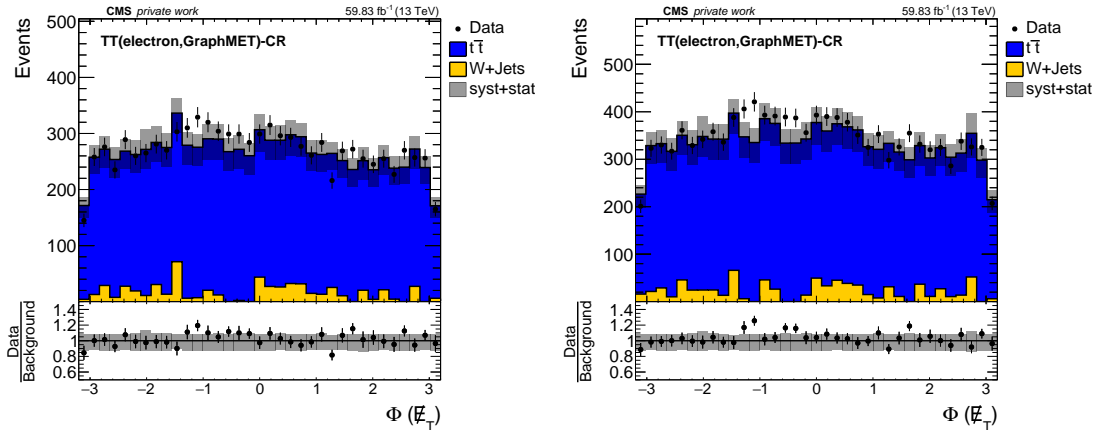
### D.9 $t\bar{t}$ Electron CR: $\phi(\vec{\cancel{E}}_T)$



(a) MET-Type I (left) and PUPPI MET (right).

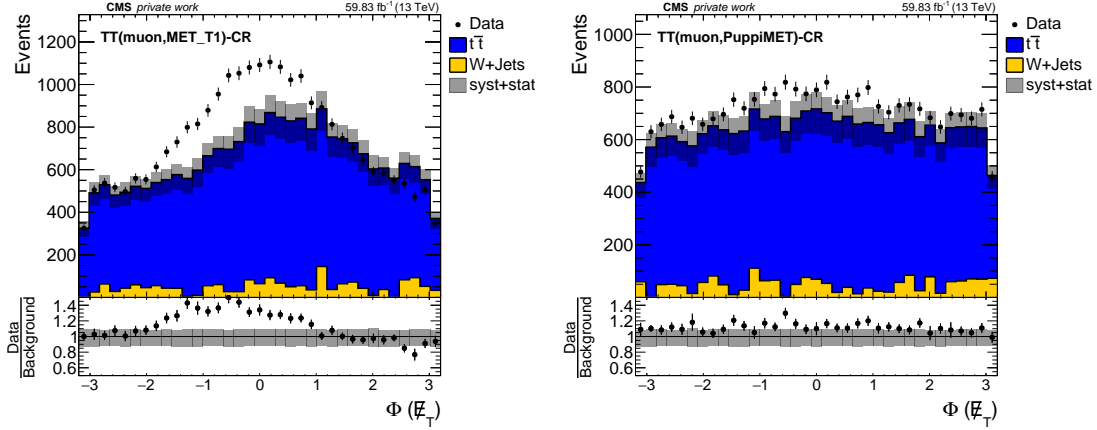


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

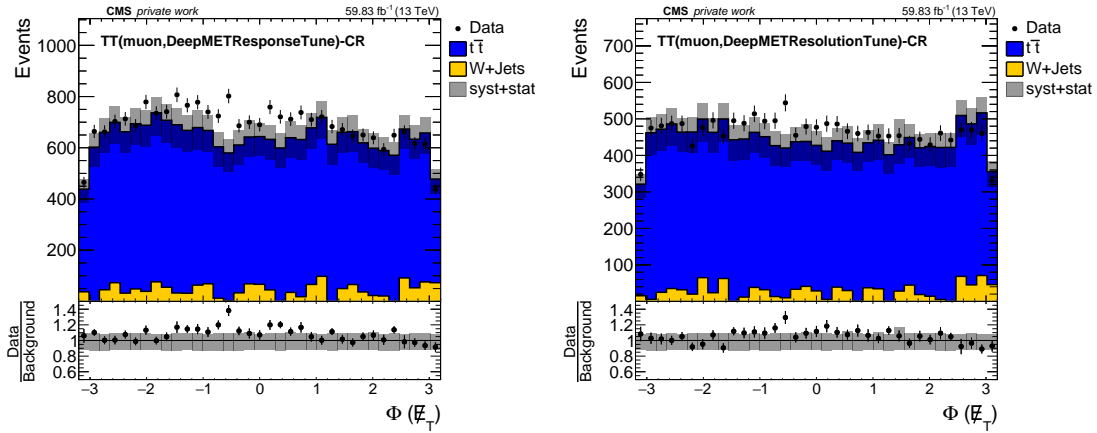


(c) GraphMET (left) and naively calibrated GraphMET (right).

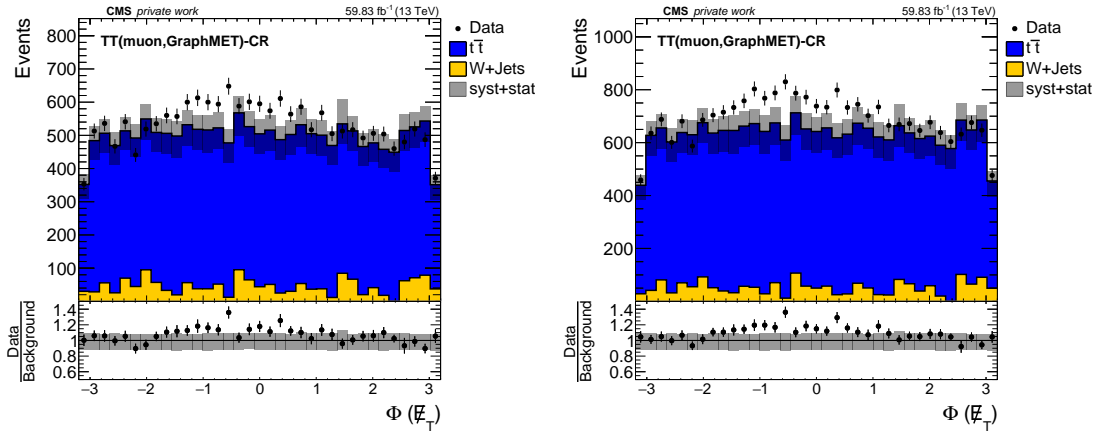
Figure D.9:  $\phi(\vec{\cancel{E}}_T)$  distributions in the  $t\bar{t}$  CR with an electron in the final state. Except for MET-Type-I, which shows a modulation in the data over background ratio, the plots show a good data-MC agreement. The event yields of the uncalibrated GraphMET and DeepMETResolutionTune are smaller than with the other MET reconstruction methods.

D.10  $t\bar{t}$  Muon CR:  $\phi(\vec{\cancel{E}}_T)$ 

(a) MET-Type I (left) and PUPPI MET (right).

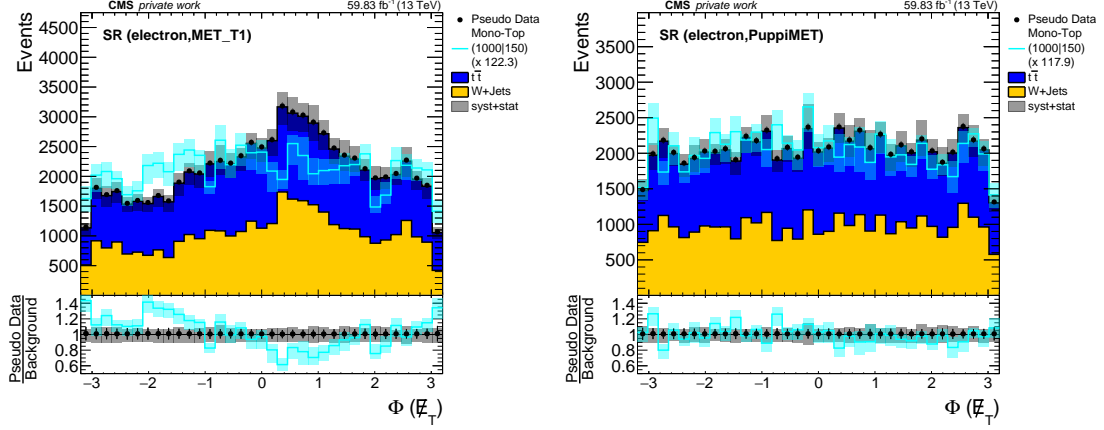


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

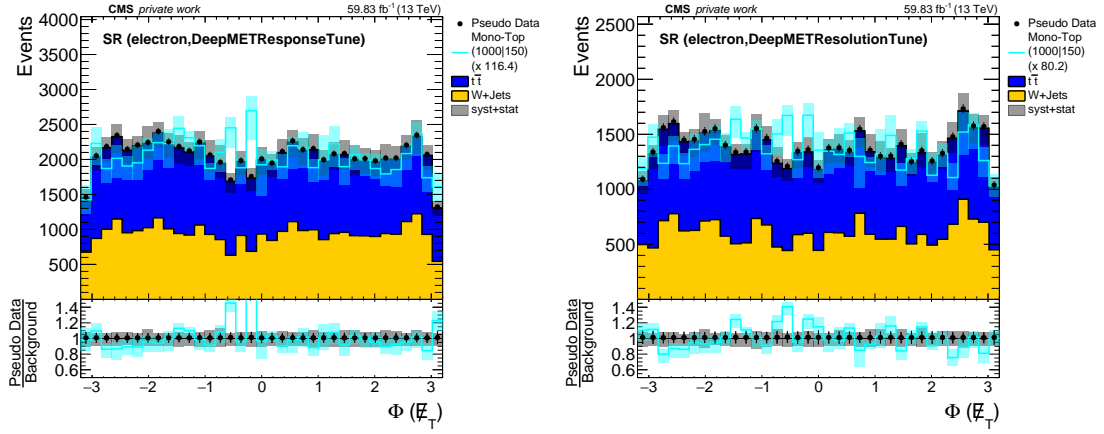


(c) GraphMET (left) and naively calibrated GraphMET (right).

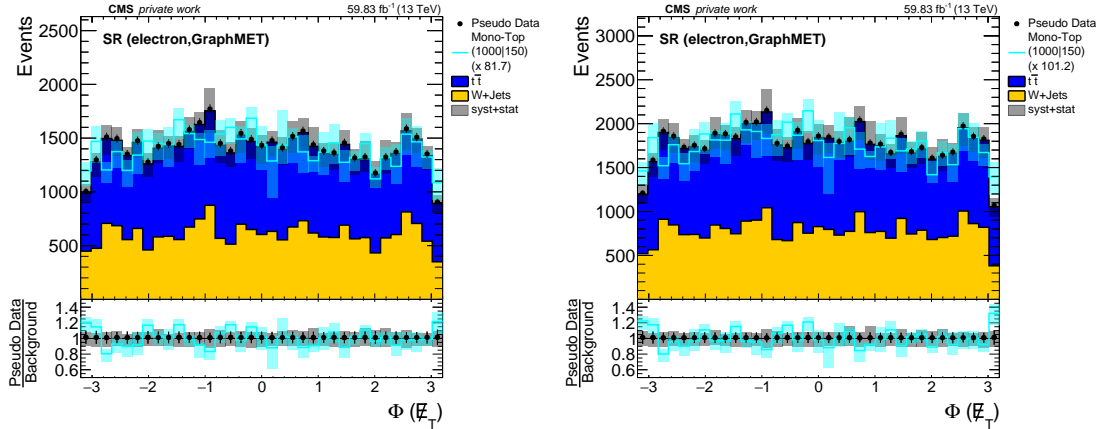
Figure D.10:  $\phi(\vec{\cancel{E}}_T)$  distributions in the  $t\bar{t}$  CR with a muon in the final state. Except for MET-Type-I, which shows a modulation in the data over background ratio, the plots show a good data-MC agreement in most bins. However, the MC undershoot in the interval  $[-2, 1]$  data is not reproduced by MC by all reconstruction methods. The event yields of the uncalibrated GraphMET and DeepMETResolutionTune are smaller than with the other MET reconstruction methods.

D.11 Electron SR:  $\phi(\vec{\cancel{E}}_T)$ 

(a) MET-Type I (left) and PUPPI MET (right).

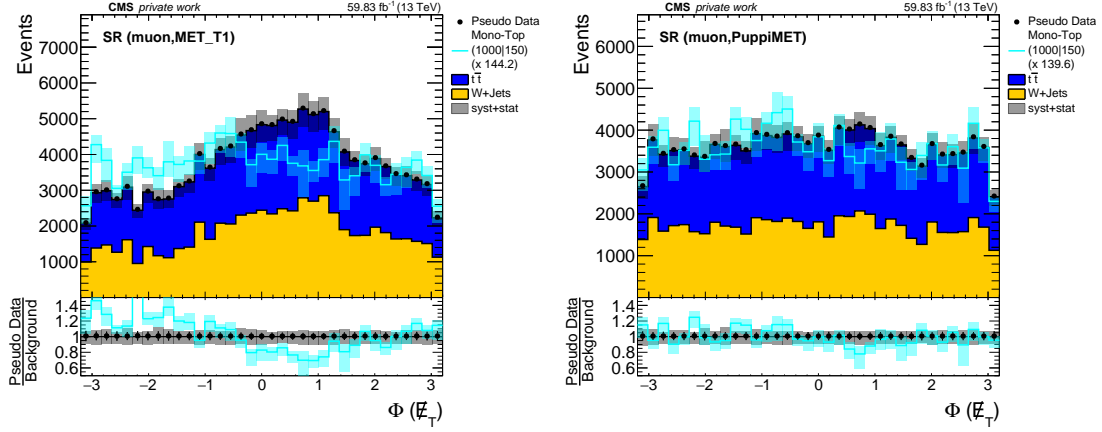


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

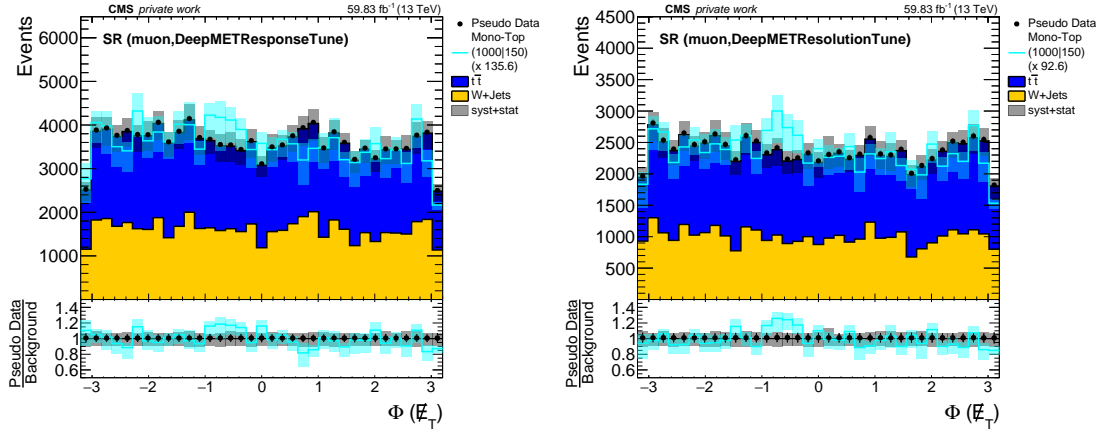


(c) GraphMET (left) and naively calibrated GraphMET (right).

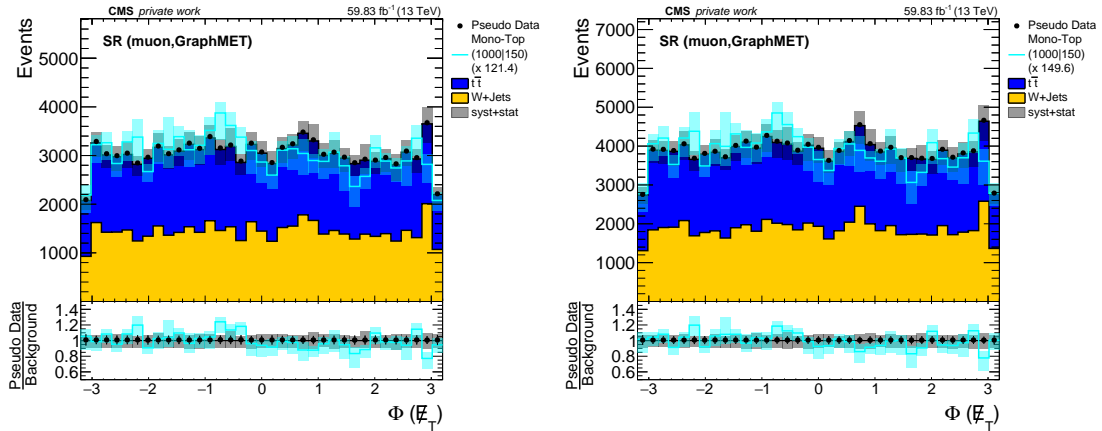
Figure D.11:  $\phi(\vec{\cancel{E}}_T)$  distributions in the SR with an electron in the final state. The shapes look similar except for MET-Type-I which shows a strong modulation in the background but not in signal. The event yields of the uncalibrated GraphMET and DeepMETResolutionTune are smaller than with the other MET reconstruction methods. The signal multiplication factor indicates the best signal-to-background ratio for the uncalibrated GraphMET and DeepMETResolutionTune.

D.12 Muon SR:  $\phi(\vec{\cancel{E}}_{\text{T}})$ 

(a) MET-Type I (left) and PUPPI MET (right).

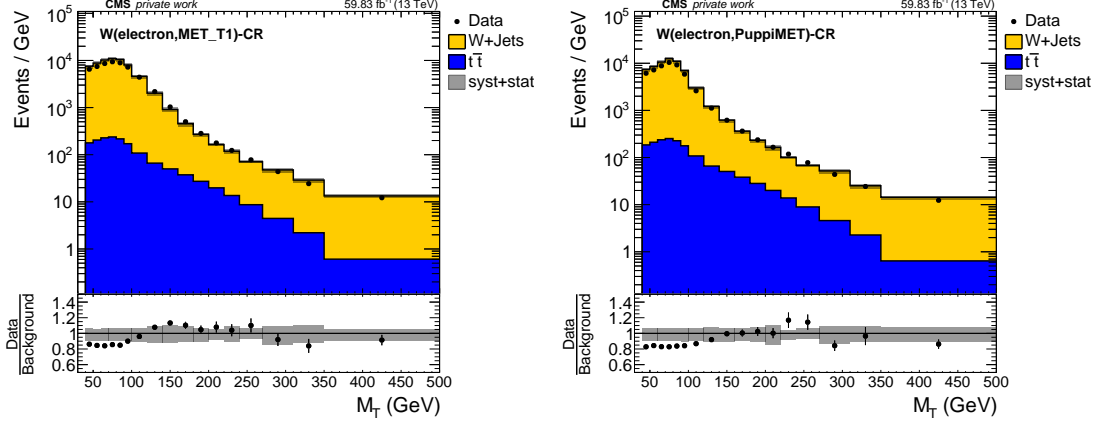


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

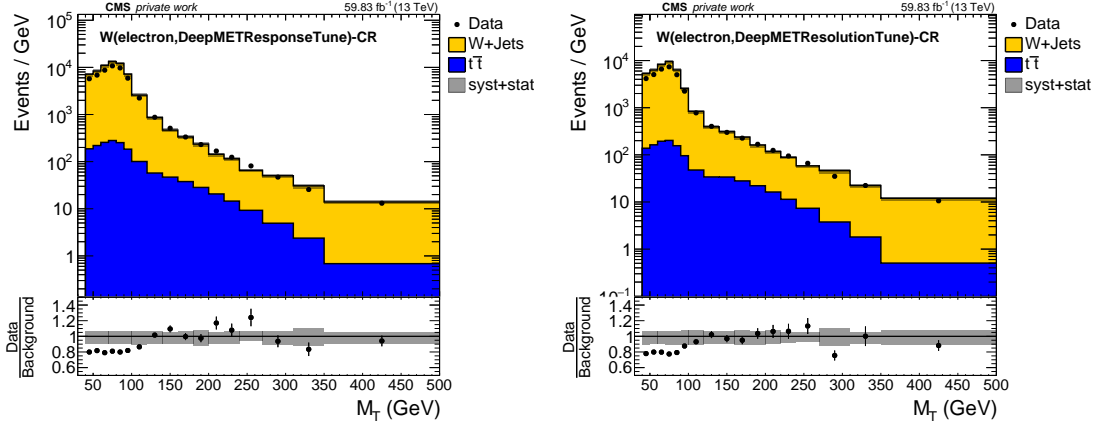


(c) GraphMET (left) and naively calibrated GraphMET (right).

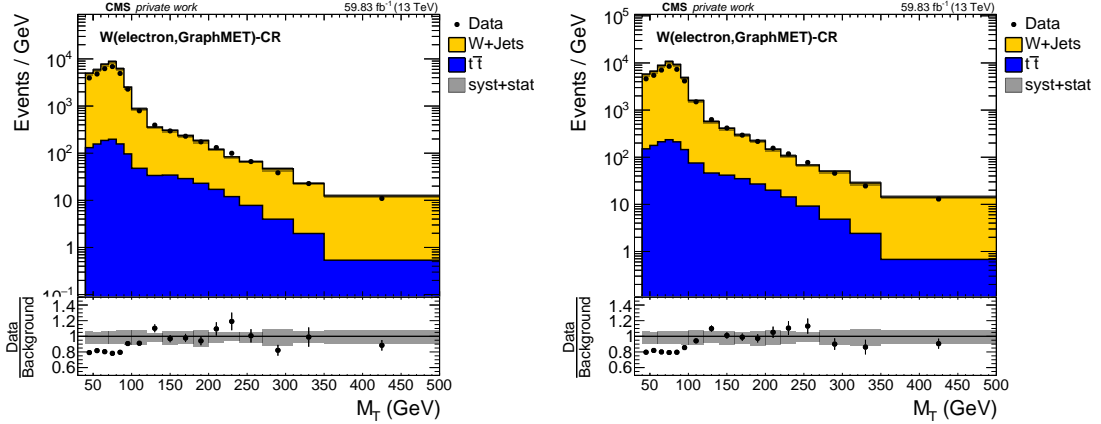
Figure D.12:  $\phi(\vec{\cancel{E}}_{\text{T}})$  distributions in the SR with a muon in the final state. The shapes look similar except for MET-Type-I which shows a strong modulation in the background but not in signal. The event yields of the uncalibrated GraphMET and DeepMETResolutionTune are smaller than with the other MET reconstruction methods. The signal multiplication factor indicates the best signal-to-background ratio for DeepMETResolutionTune.

D.13  $W$  Electron CR:  $M_T$ 

(a) MET-Type I (left) and PUPPI MET (right).



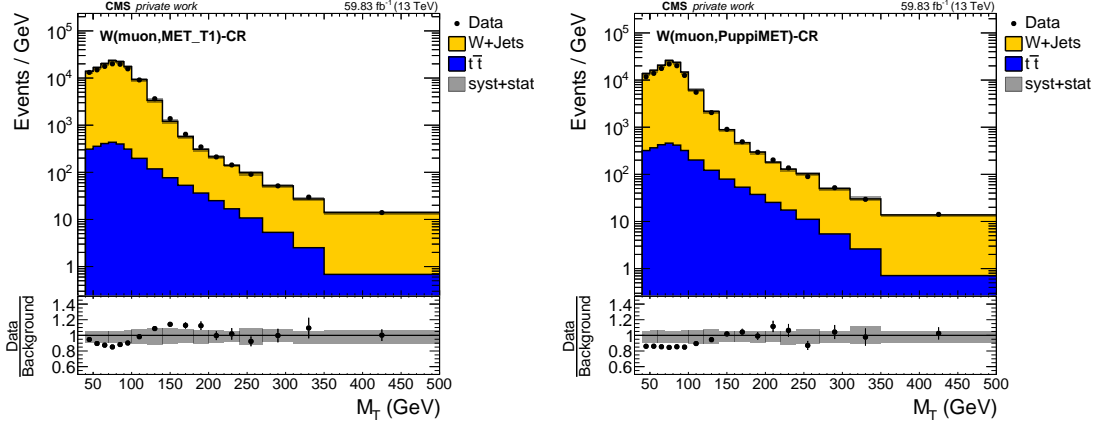
(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).



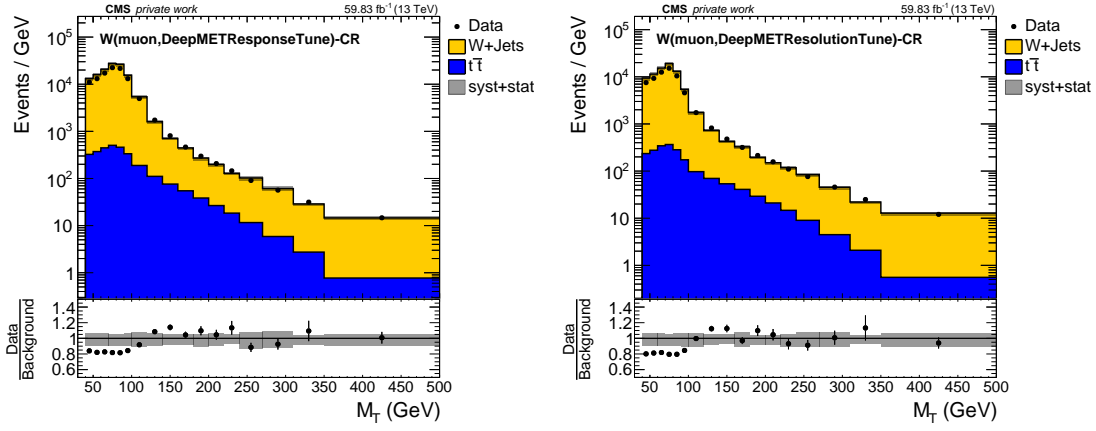
(c) GraphMET (left) and naively calibrated GraphMET (right).

Figure D.13:  $M_T$  distributions in the  $W$  CR with an electron in the final state. A modulation in data over background can be seen using any of the reconstruction methods. It is least dominant for PUPPI MET and most dominant for both DeepMET versions and both GraphMET versions. The Jacobian edge is least visible for MET-Type-I and most visible for GraphMET and DeepMETResolutionTune. The event yields are again smaller for GraphMET and DeepMETResolutionTune.

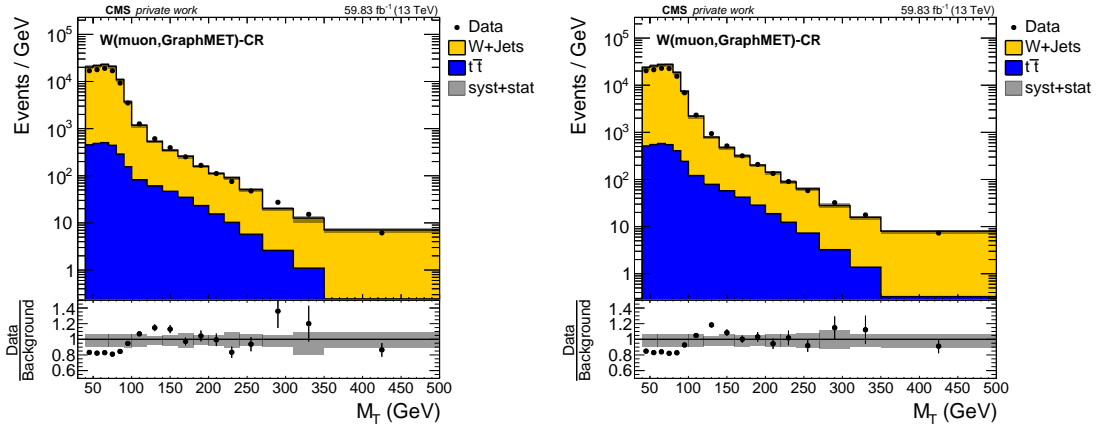


D.14  $W$  Muon CR:  $M_T$ 

(a) MET-Type I (left) and PUPPI MET (right).

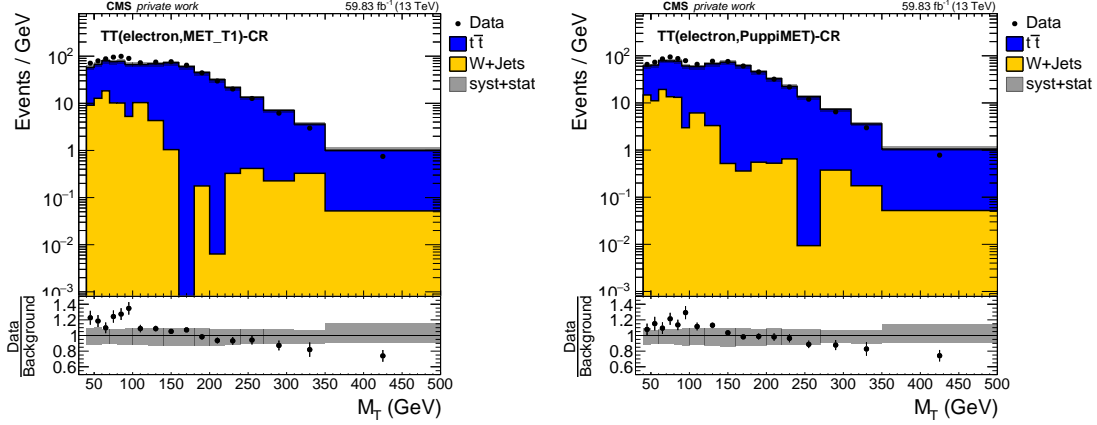


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

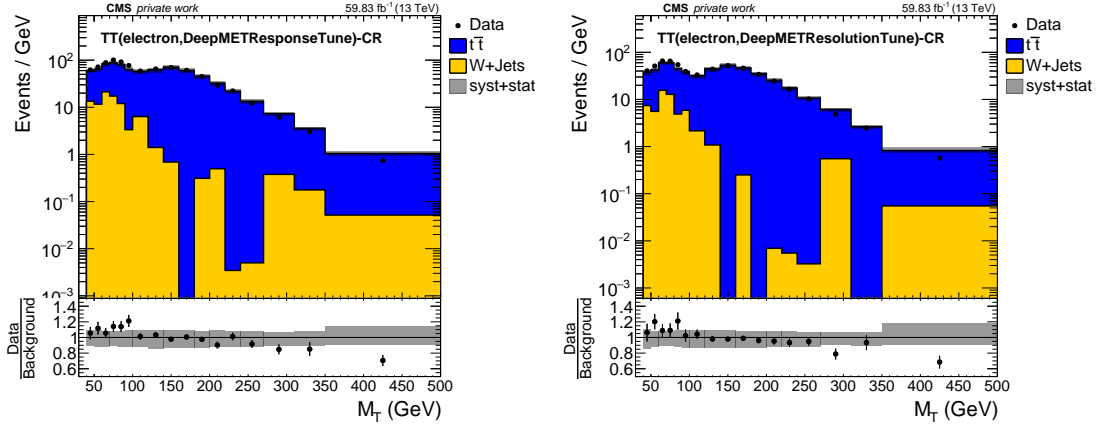


(c) GraphMET (left) and naively calibrated GraphMET (right).

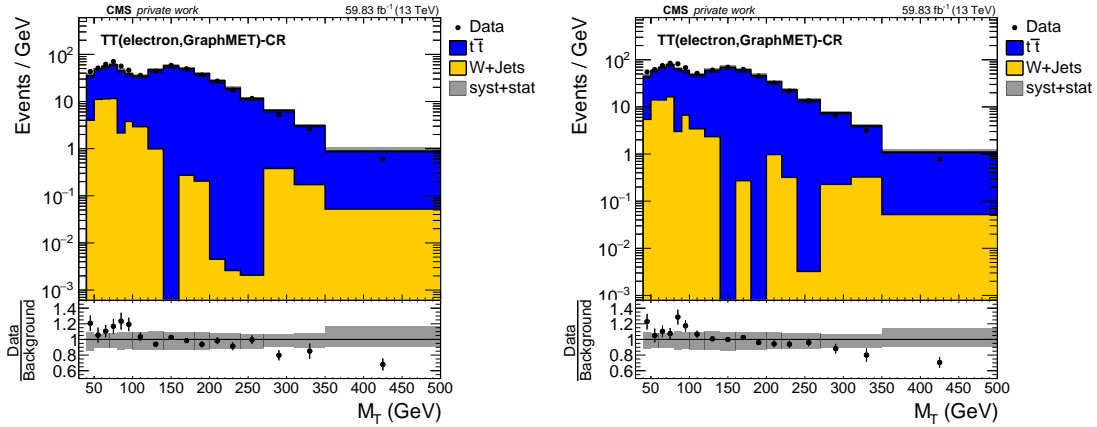
Figure D.14:  $M_T$  distributions in the  $W$  CR with a muon in the final state. A modulation in data over background can be seen using any of the reconstruction methods. It is least dominant for PUPPI MET and most dominant for both DeepMET versions and both GraphMET versions. The Jacobian edge is least visible for MET-Type-I and most visible for GraphMET and DeepMETResolutionTune. The event yields are again smaller for GraphMET and DeepMETResolutionTune.

D.15  $t\bar{t}$  Electron CR:  $M_T$ 

(a) MET-Type I (left) and PUPPI MET (right).

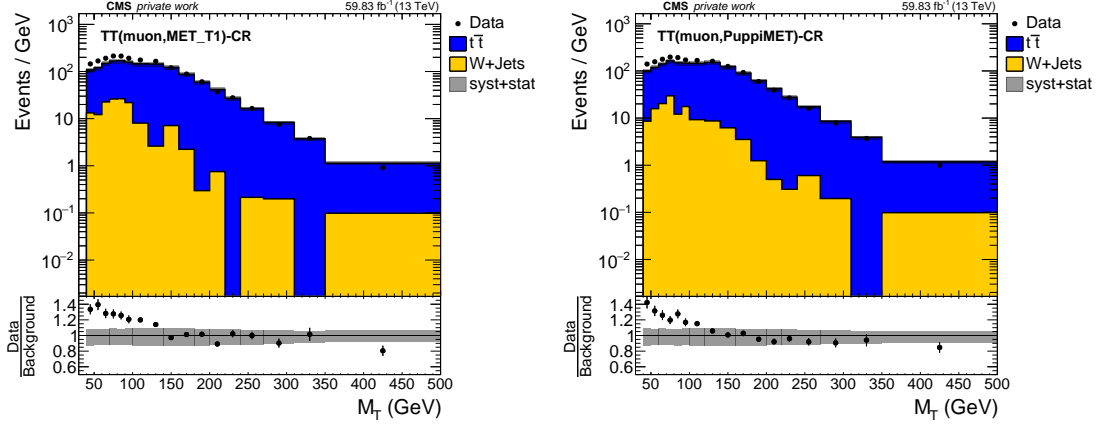


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

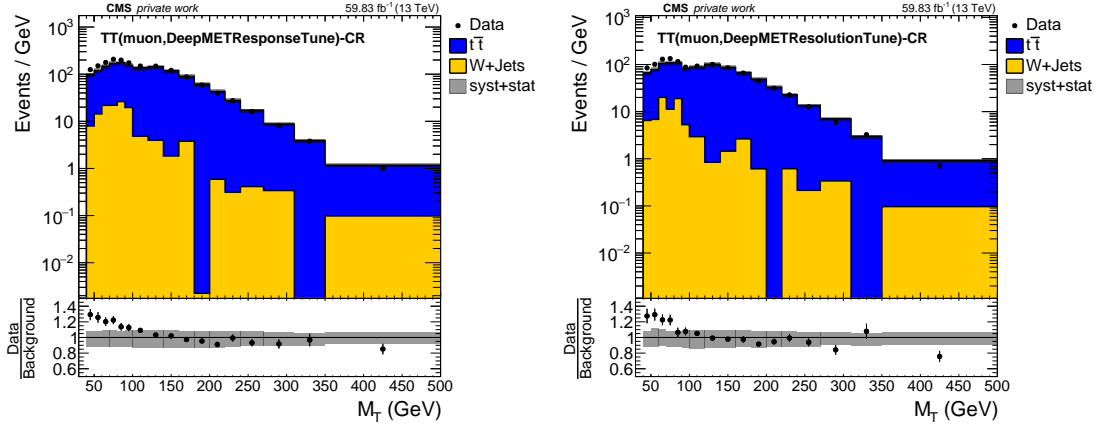


(c) GraphMET (left) and naively calibrated GraphMET (right).

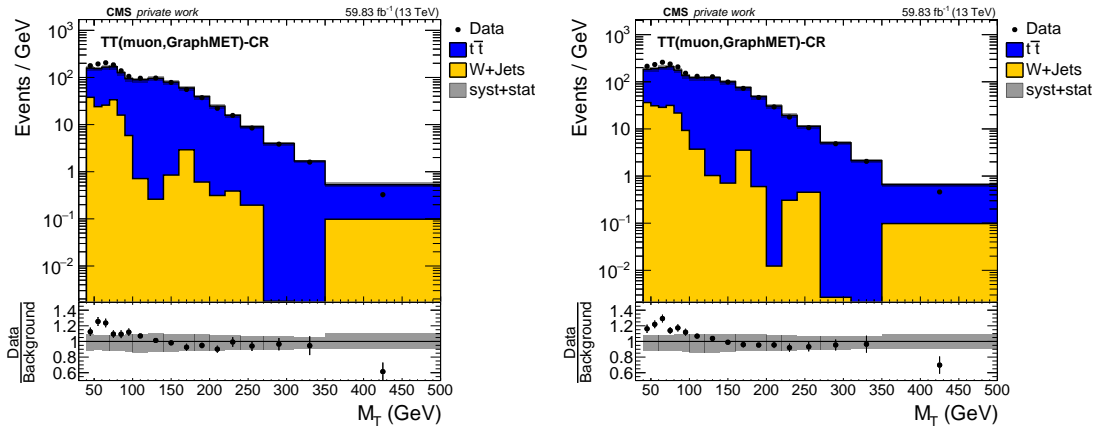
Figure D.15:  $M_T$  distributions in the  $t\bar{t}$  CR with an electron in the final state. The data-MC agreement is good for larger  $M_T$ . For smaller  $M_T$  values, there is a tendency for an MC undershoot. The Jacobian edge is still visible but much less dominant than in the  $W$  CRs. It is the least visible for MET-Type-I and PUPPI MET. The event yields are again smaller for GraphMET and DeepMETResolutionTune.

D.16  $t\bar{t}$  Muon CR:  $M_T$ 

(a) MET-Type I (left) and PUPPI MET (right).

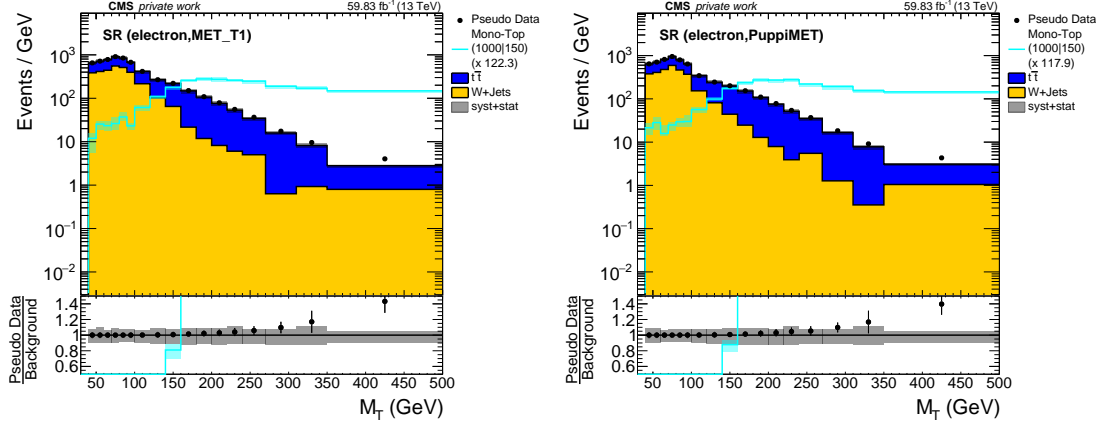


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

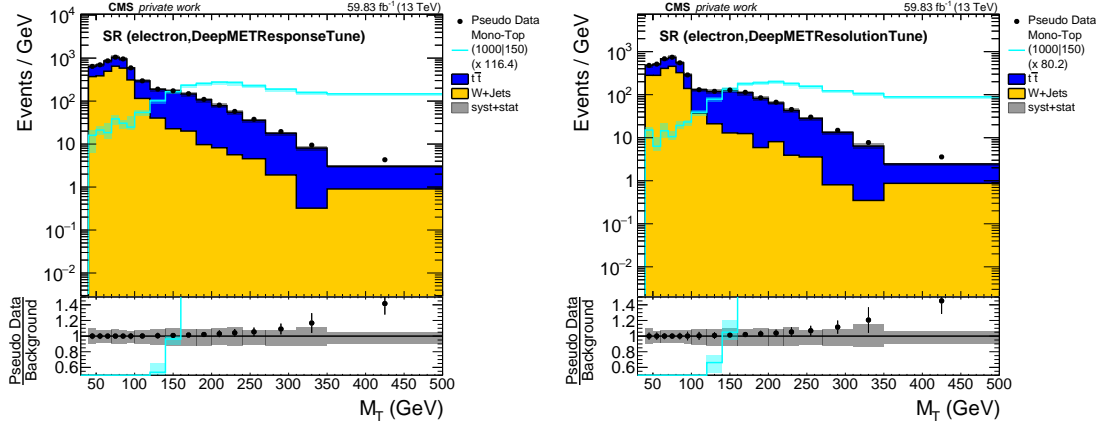


(c) GraphMET (left) and naively calibrated GraphMET (right).

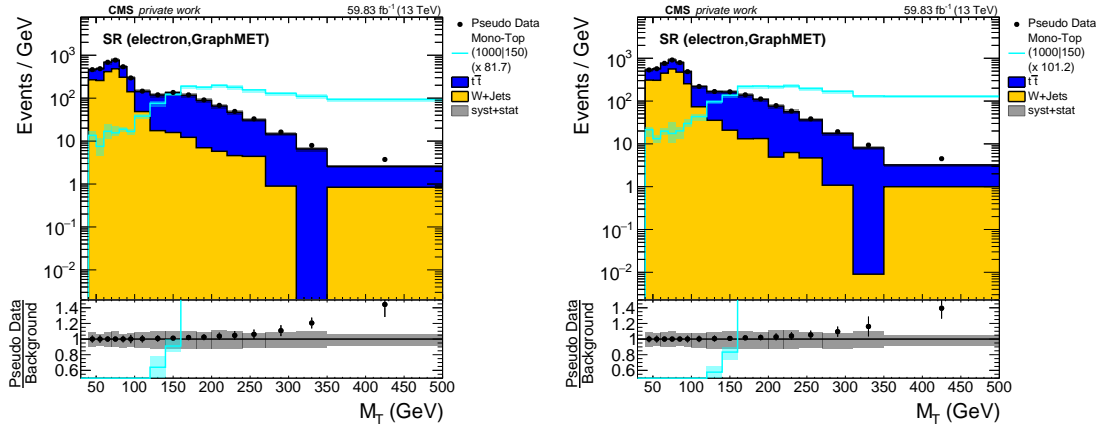
Figure D.16:  $M_T$  distributions in the  $t\bar{t}$  CR with a muon in the final state. The data-MC agreement is good for larger  $M_T$ . For smaller  $M_T$  values, there is a tendency for an MC undershoot. The Jacobian edge is still visible but much less dominant than in the  $W$  CRs. It is the least visible for MET-Type-I and PUPPI MET. The event yields are again smaller for GraphMET and DeepMETResolutionTune.

D.17 Electron SR:  $M_T$ 

(a) MET-Type I (left) and PUPPI MET (right).

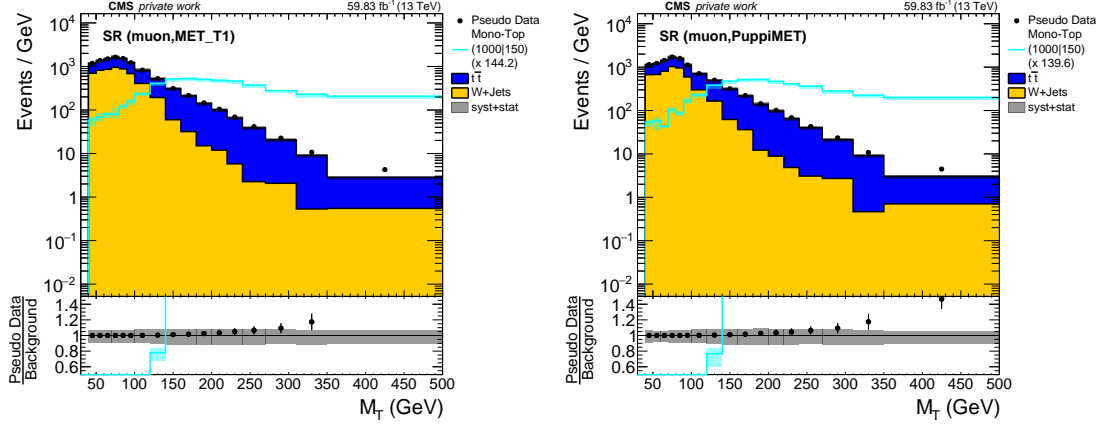


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

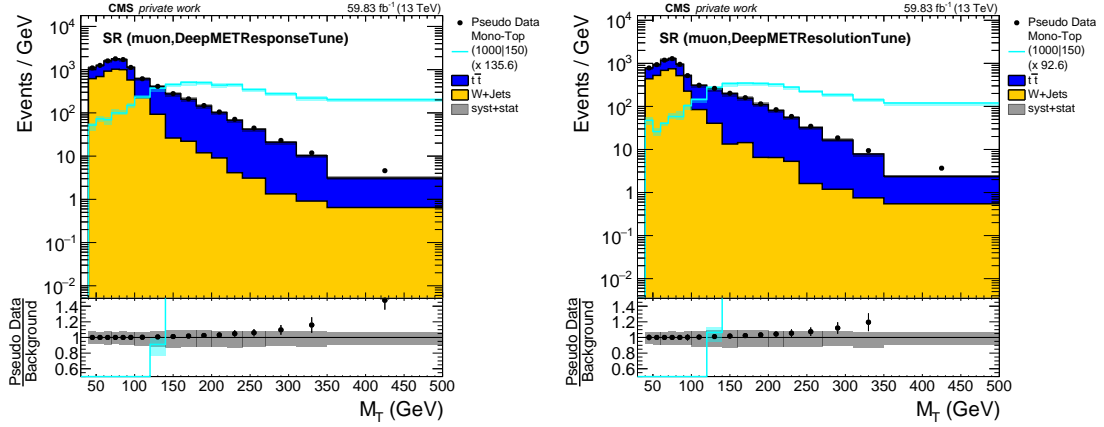


(c) GraphMET (left) and naively calibrated GraphMET (right).

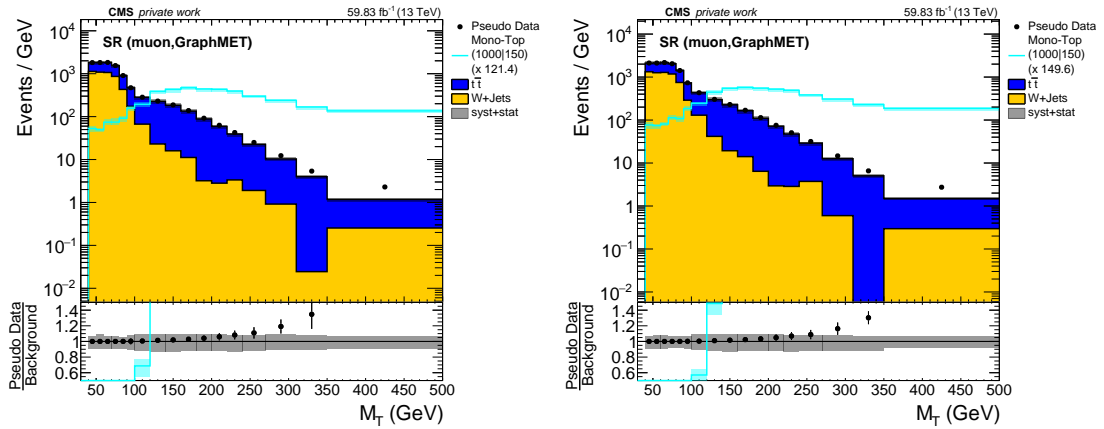
Figure D.17:  $M_T$  distributions in the SR with an electron in the final state. The Jacobian edge is again less visible for MET-Type-I and PUPPI MET, the signal-to-background ratio is best for DeepMETResolutionTune and GraphMET, judging by the signal multiplication factor. The pseudo-data over background ratio is best for GraphMET as can be seen from the second to last bin showing a significant excess already.

D.18 Muon SR:  $M_T$ 

(a) MET-Type I (left) and PUPPI MET (right).

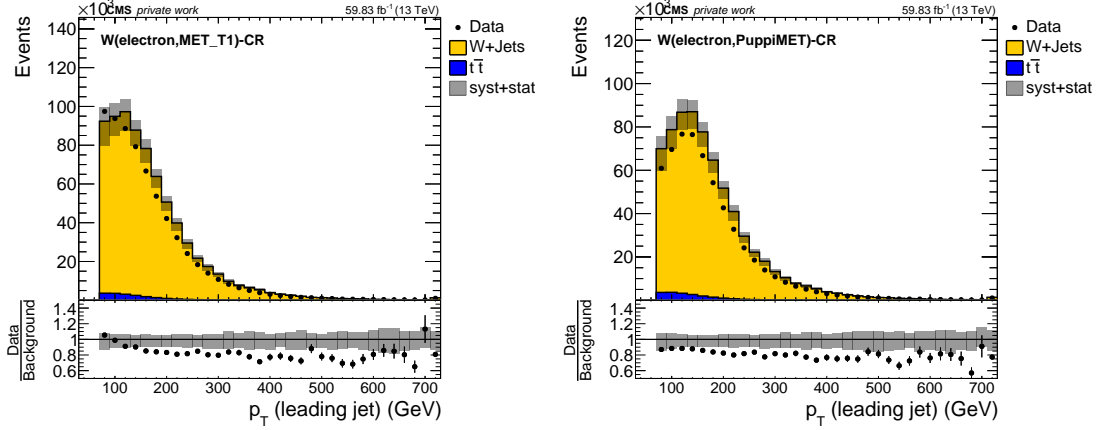


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

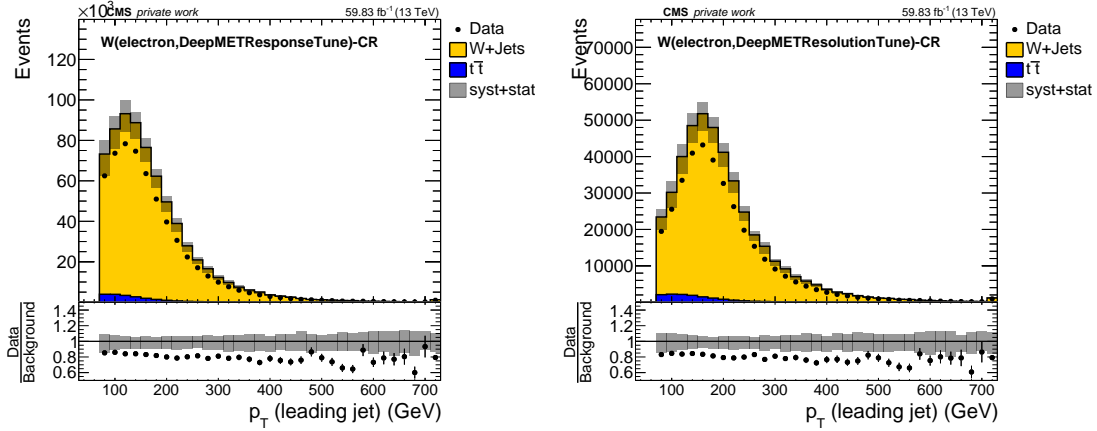


(c) GraphMET (left) and naively calibrated GraphMET (right).

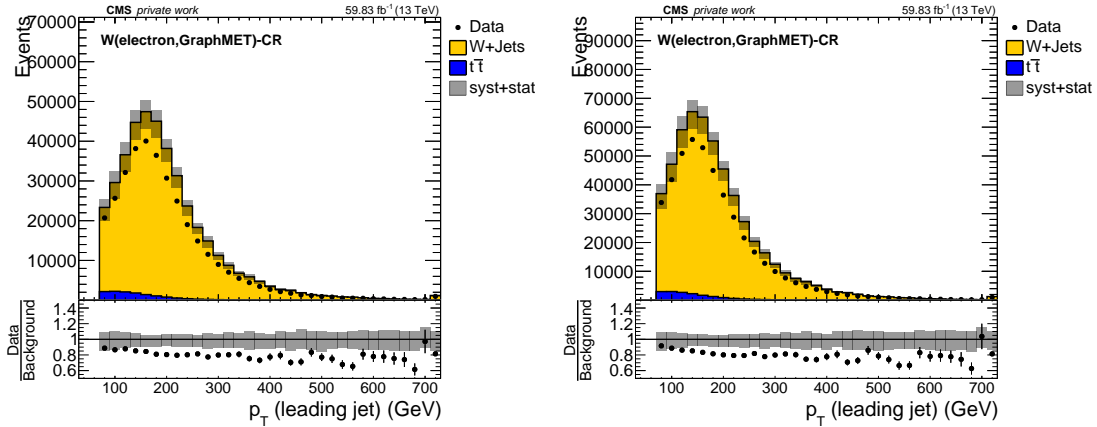
Figure D.18:  $M_T$  distributions in the SR with a muon in the final state. The Jacobian edge is again less visible for MET-Type-I and PUPPI MET. It is most visible for both GraphMET versions while the signal to background ratio is best for DeepMETResolutionTune, judging by the signal multiplication factor. The pseudodata over background ratio is best for GraphMET as can be seen from the third to last bin showing a significant excess already.

D.19  $W$  Electron CR:  $p_T$ (leading jet)

(a) MET-Type I (left) and PUPPI MET (right).

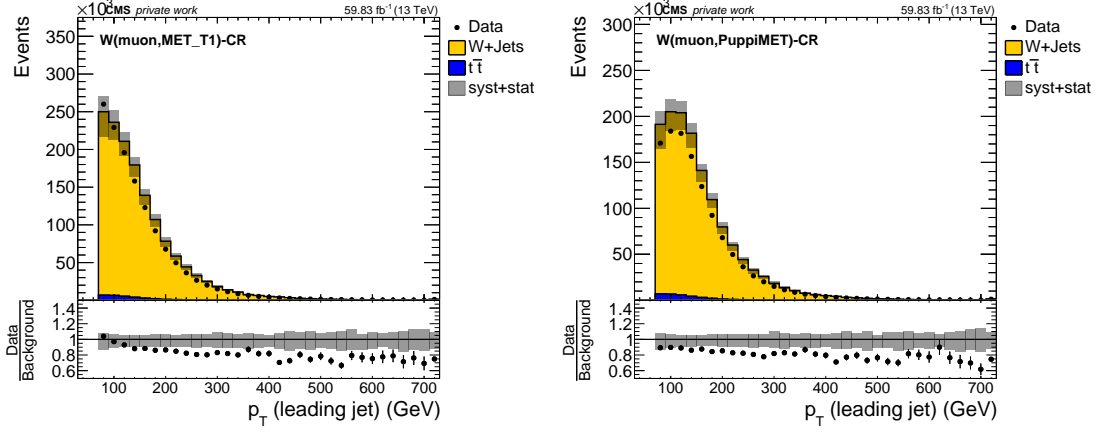


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

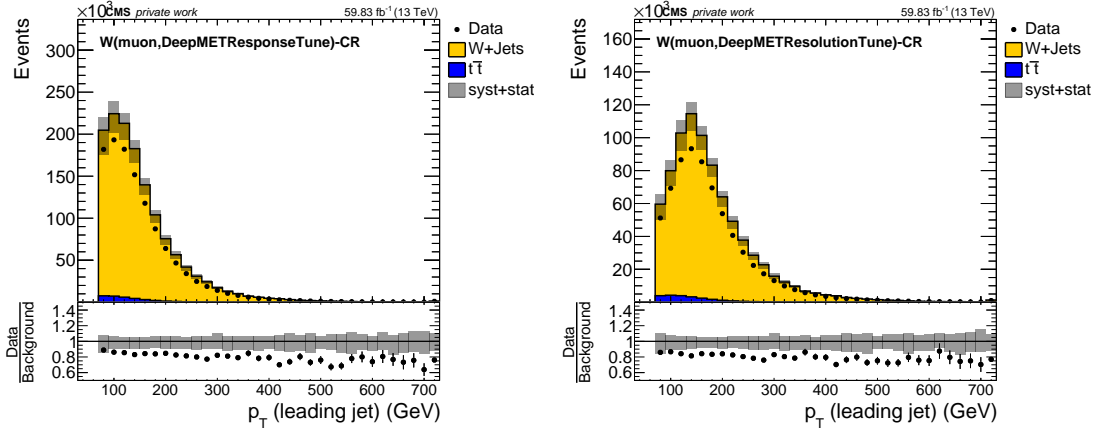


(c) GraphMET (left) and naively calibrated GraphMET (right).

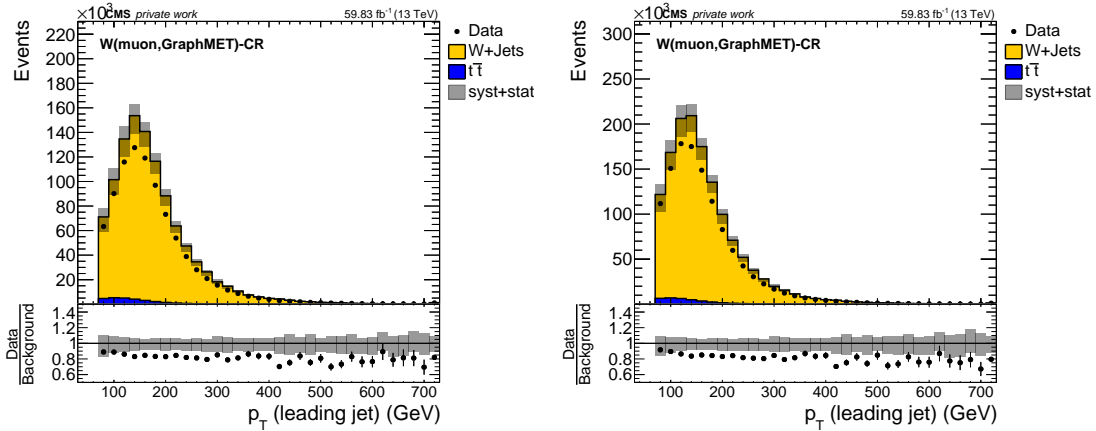
Figure D.19:  $p_T$ (leading jet) distributions in the  $W$  CR with an electron in the final state. The plots show an MC overshoot of roughly 10 – 20 % for all MET reconstruction methods. The event yields of GraphMET and DeepMETResolutionTune are notably smaller than those of the other methods. The distribution of MET-Type-I has more events at small  $p_T$ . This might be due to the worse MET resolution letting more events with small MET pass the criteria and due to momentum conservation also events with small jet momenta.

D.20  $W$  Muon CR:  $p_T(\text{leading jet})$ 

(a) MET-Type I (left) and PUPPI MET (right).



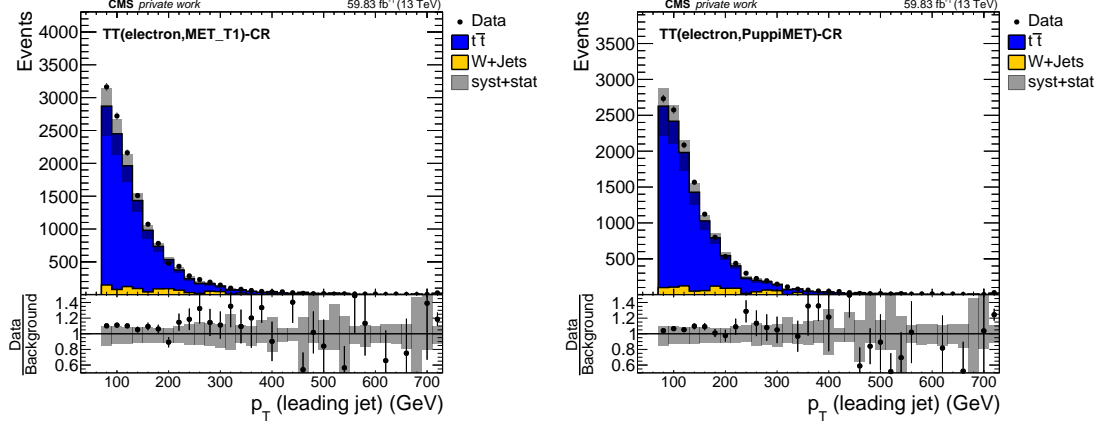
(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).



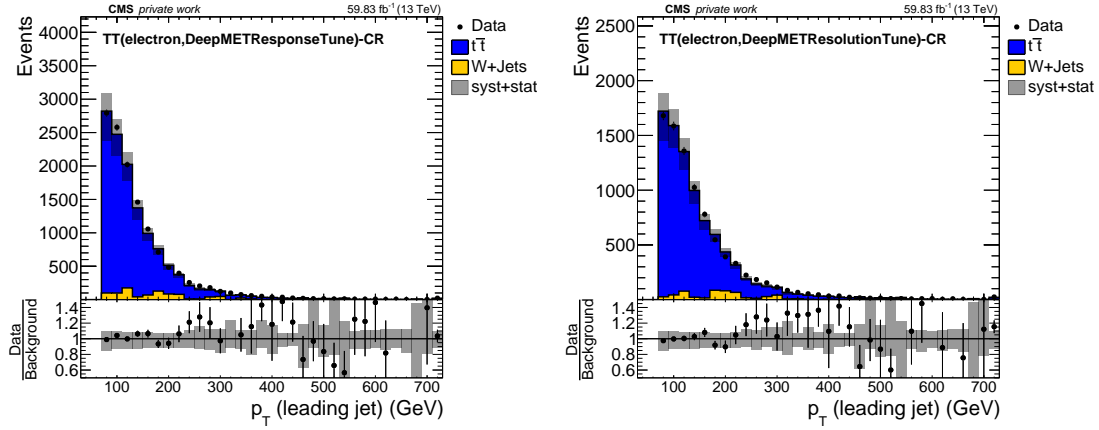
(c) GraphMET (left) and naively calibrated GraphMET (right).

Figure D.20:  $p_T(\text{leading jet})$  distributions in the  $W$  CR with a muon in the final state. The plots show an MC overshoot of roughly 10 – 20 % for all MET reconstruction methods. The event yields of GraphMET and DeepMETResolutionTune are notably smaller than those of the other methods. The distribution of MET-Type-I has more events at small  $p_T$ . This might be due to the worse MET resolution letting more events with small MET pass the criteria and due to momentum conservation also events with small jet momenta.

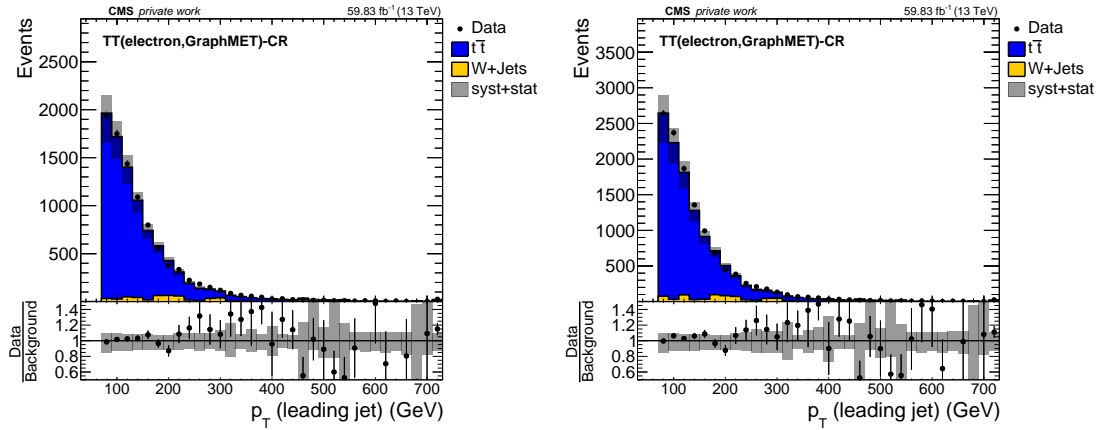
## D.21 $t\bar{t}$ Electron CR: $p_T(\text{leading jet})$



(a) MET-Type I (left) and PUPPI MET (right).



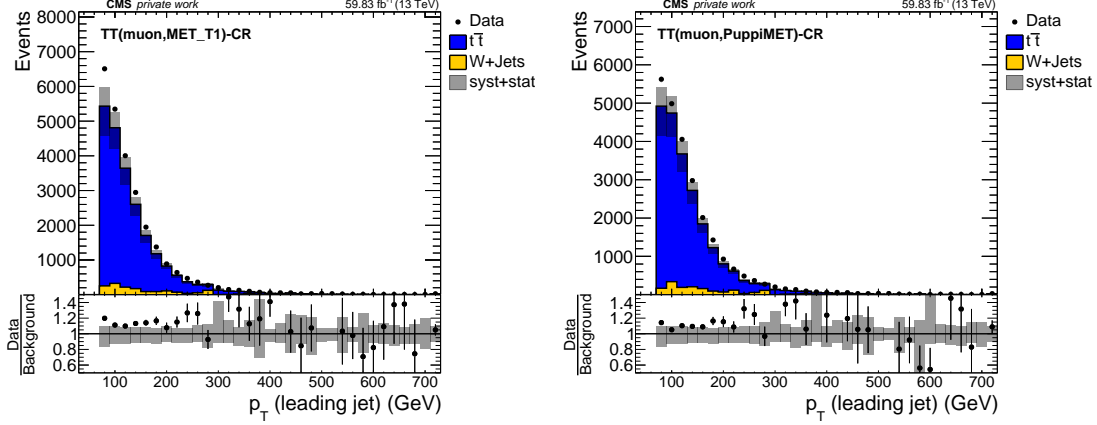
(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).



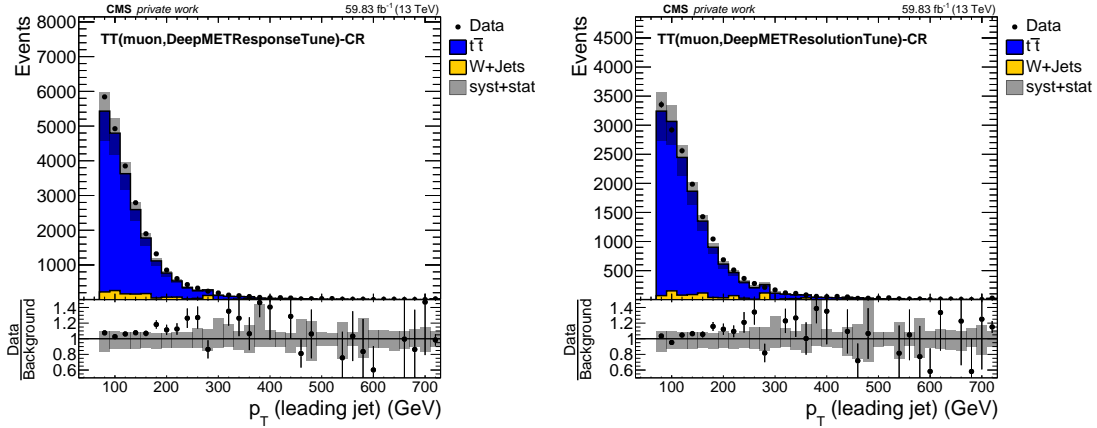
(c) GraphMET (left) and naively calibrated GraphMET (right).

Figure D.21:  $p_T(\text{leading jet})$  distributions in the  $t\bar{t}$  CR with an electron in the final state. The data-MC agreement is reasonable considering the large uncertainties. Except for the event yields being smaller for GraphMET and DeepMETResolutionTune, there is no big difference.

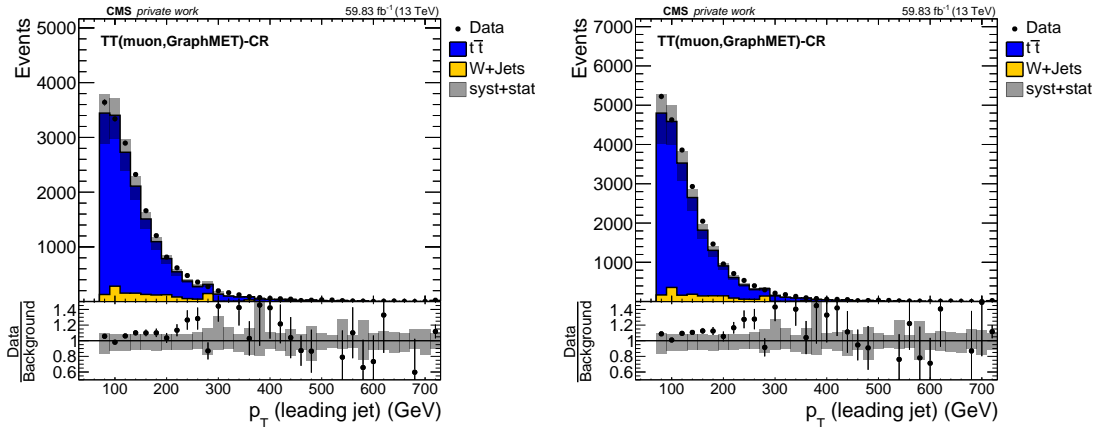


D.22  $t\bar{t}$  Muon CR:  $p_T(\text{leading jet})$ 

(a) MET-Type I (left) and PUPPI MET (right).

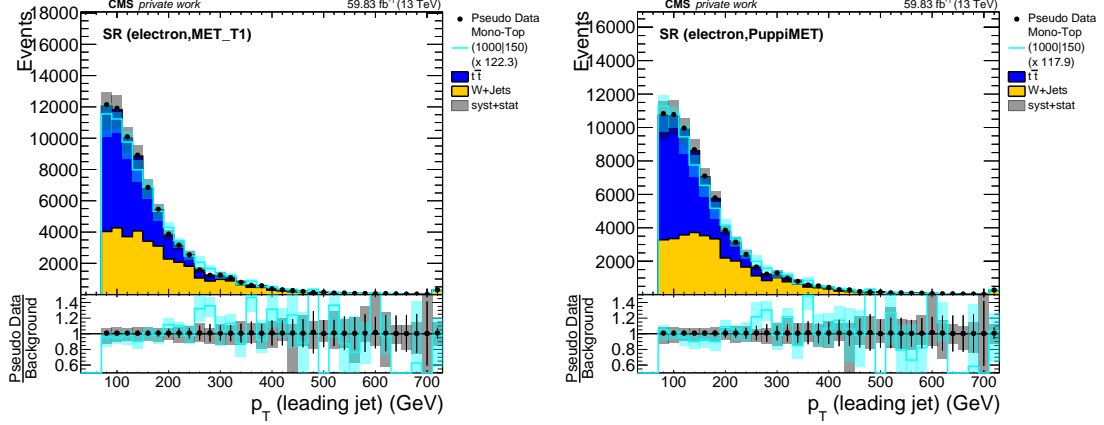


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

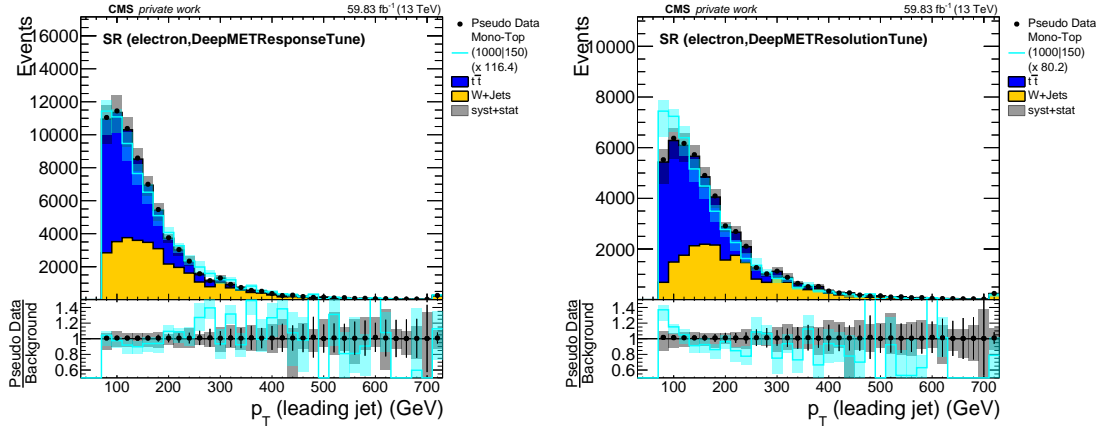


(c) GraphMET (left) and naively calibrated GraphMET (right).

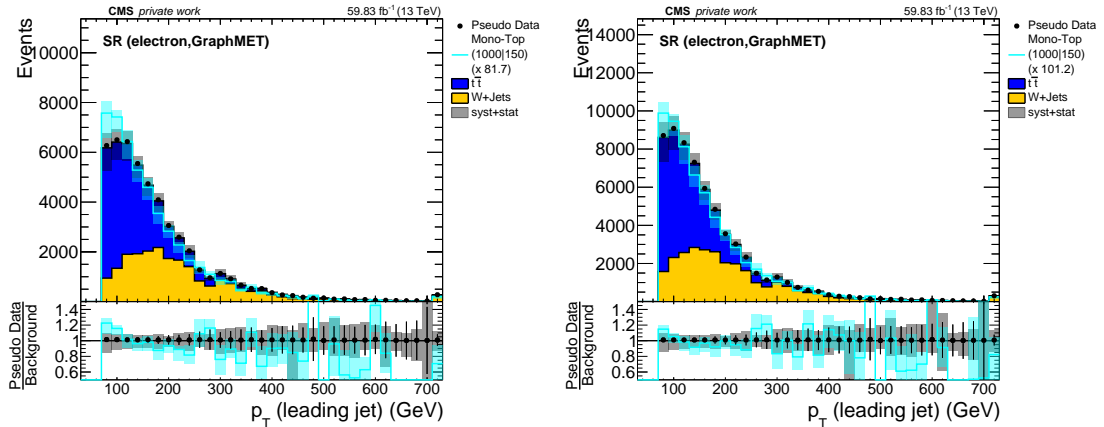
Figure D.22:  $p_T(\text{leading jet})$  distributions in the  $t\bar{t}$  CR with a muon in the final state. The simulation tends to undershoot compared to data. Again, except for the event yields being smaller for GraphMET and DeepMETResolutionTune, there is no big difference between the distributions.

D.23 Electron SR:  $p_T$ (leading jet)

(a) MET-Type I (left) and PUPPI MET (right).

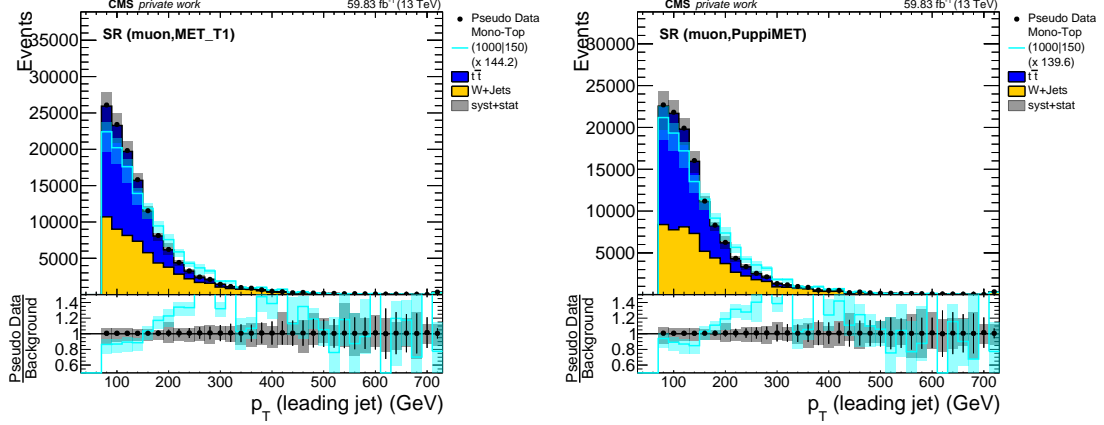


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

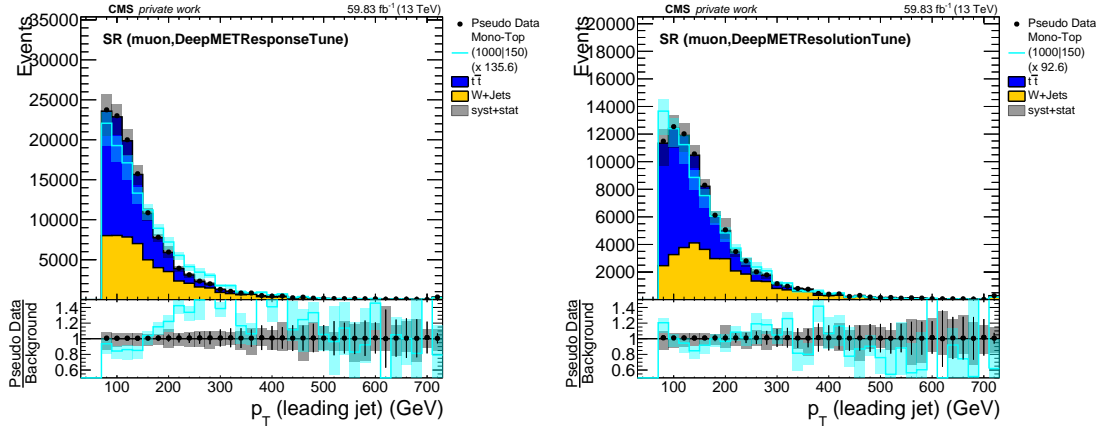


(c) GraphMET (left) and naively calibrated GraphMET (right).

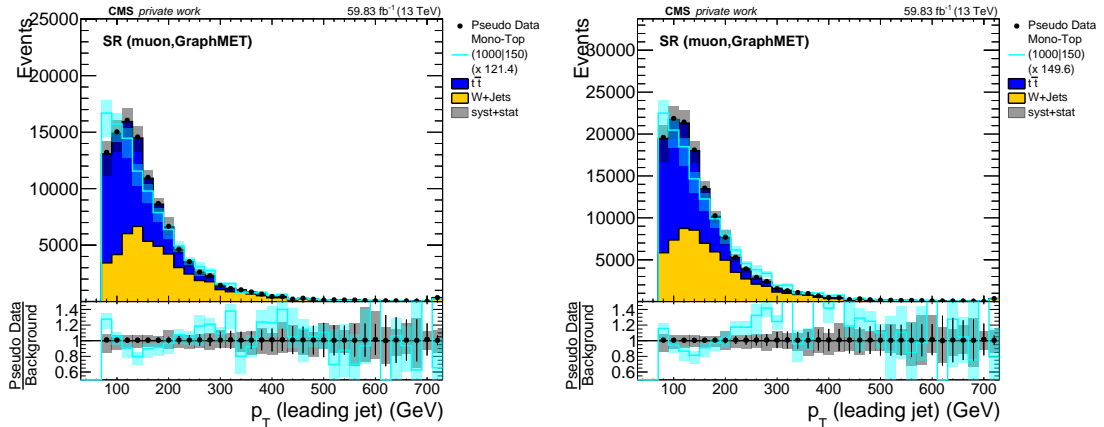
Figure D.23:  $p_T$ (leading jet) distributions in the SR with an electron in the final state. Except for the suppression of the  $W$ +Jets background for the MET reconstruction methods with good resolution, and the smaller event yields for GraphMET and DeepMETResolutionTune, the shapes look similar.

D.24 Muon SR:  $p_T(\text{leading jet})$ 

(a) MET-Type I (left) and PUPPI MET (right).

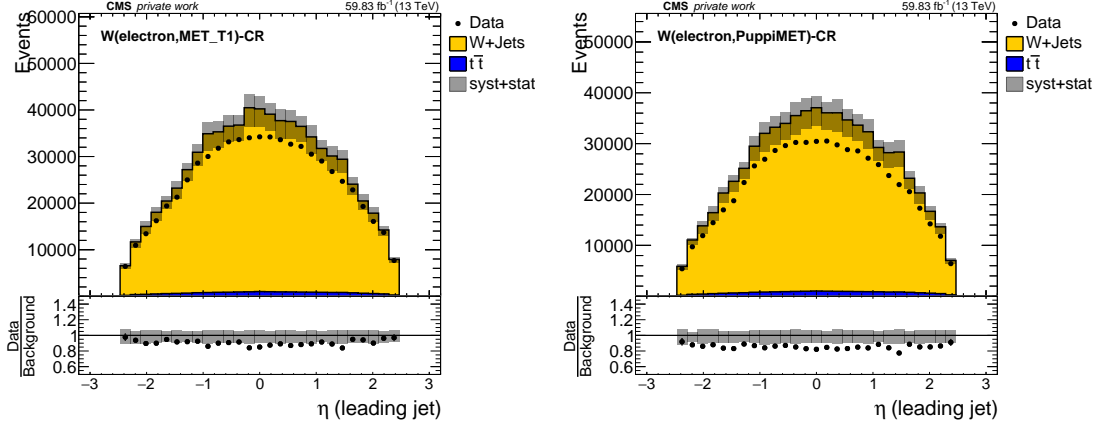


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

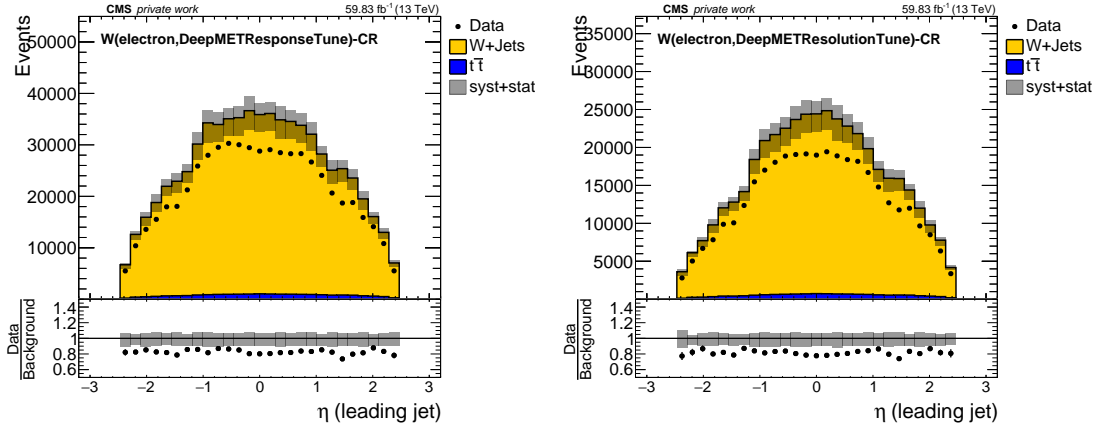


(c) GraphMET (left) and naively calibrated GraphMET (right).

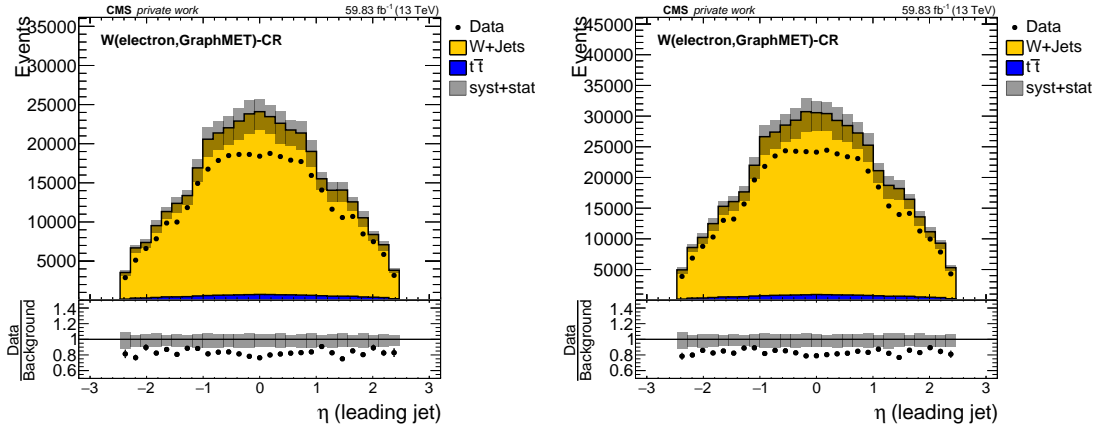
Figure D.24:  $p_T(\text{leading jet})$  distributions in the SR with a muon in the final state. Except for the suppression of the  $W+Jets$  background for the MET reconstruction methods with good resolution, and the smaller event yields for GraphMET and DeepMETResolutionTune, the shapes look similar.

D.25  $W$  Electron CR:  $\eta$ (leading jet)

(a) MET-Type I (left) and PUPPI MET (right).

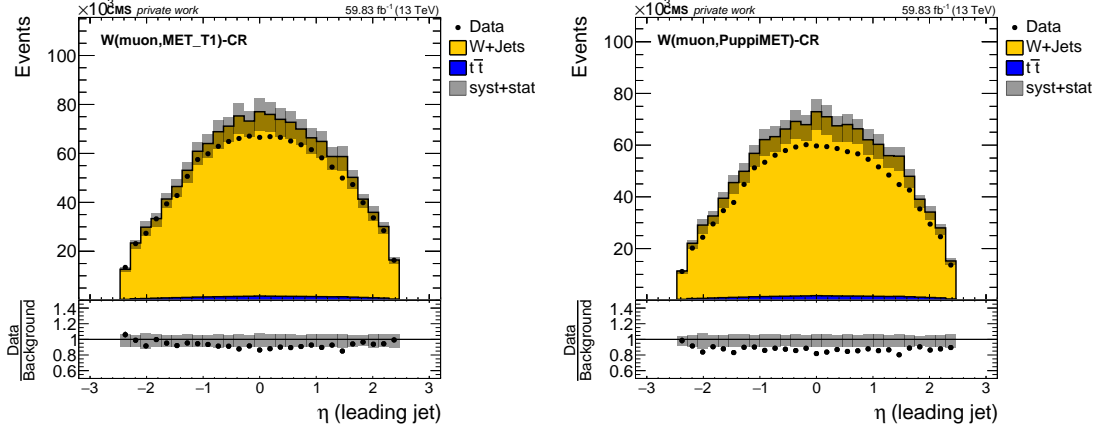


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

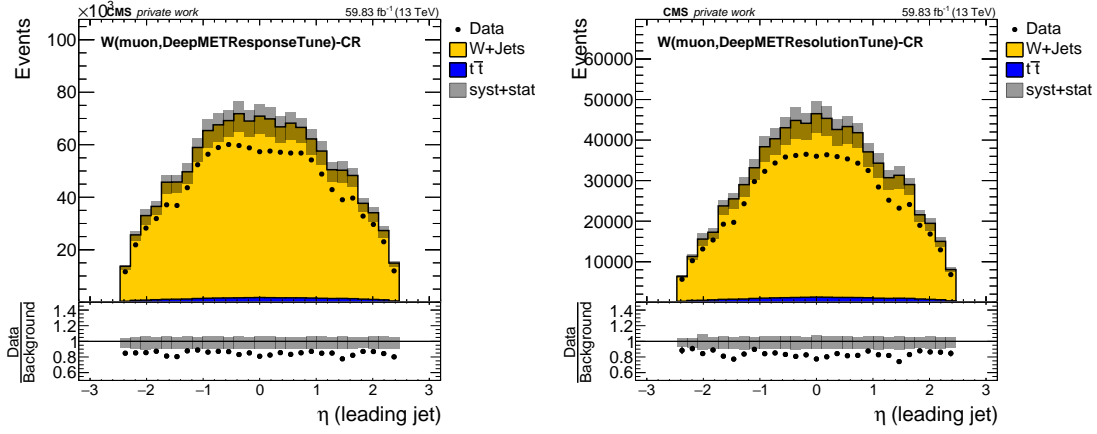


(c) GraphMET (left) and naively calibrated GraphMET (right).

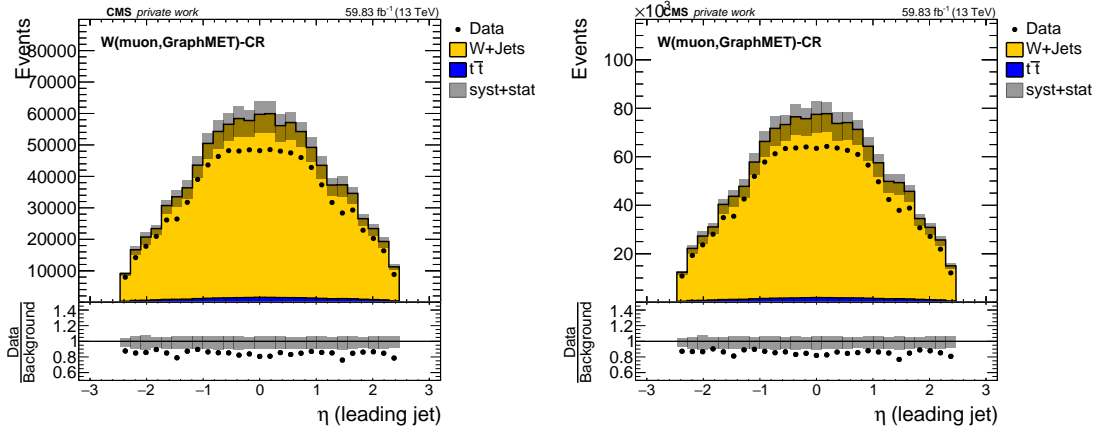
Figure D.25:  $\eta$ (leading jet) distributions in the  $W$  CR with an electron in the final state. In general, the plots show an MC overshoot of roughly 10 – 20 % for all MET reconstruction methods. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other methods. The shapes of DeepMETResolutionTune and GraphMET have sharper edges than the other shapes. This might be due to the better resolution and worse response, letting fewer events with low true MET pass the criterion.

D.26  $W$  Muon CR:  $\eta(\text{leading jet})$ 

(a) MET-Type I (left) and PUPPI MET (right).

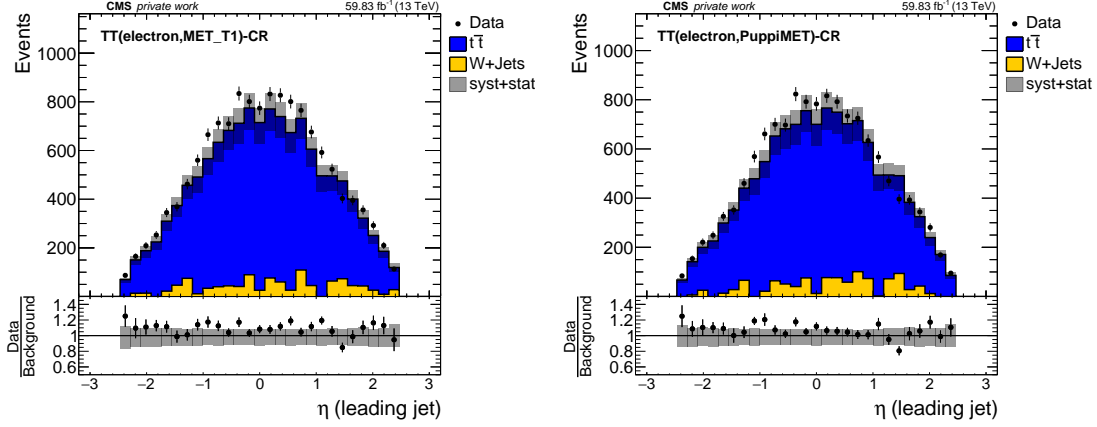


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

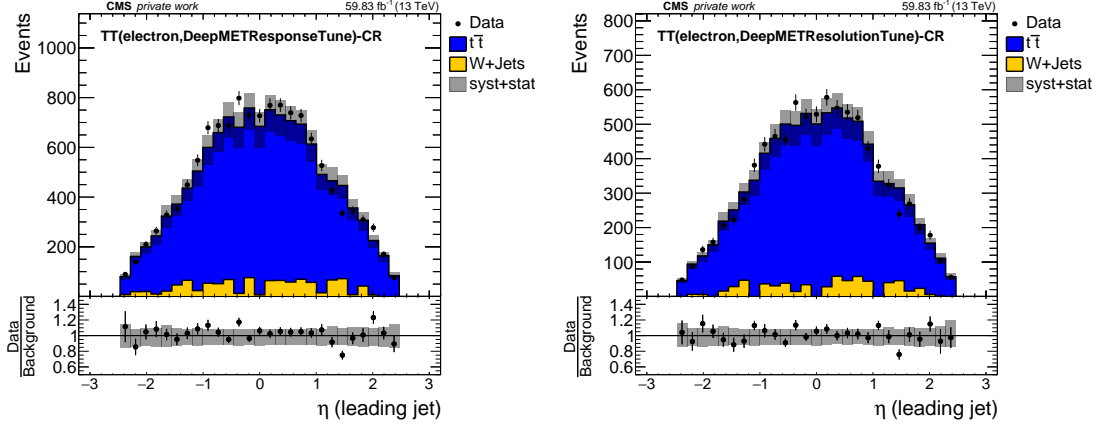


(c) GraphMET (left) and naively calibrated GraphMET (right).

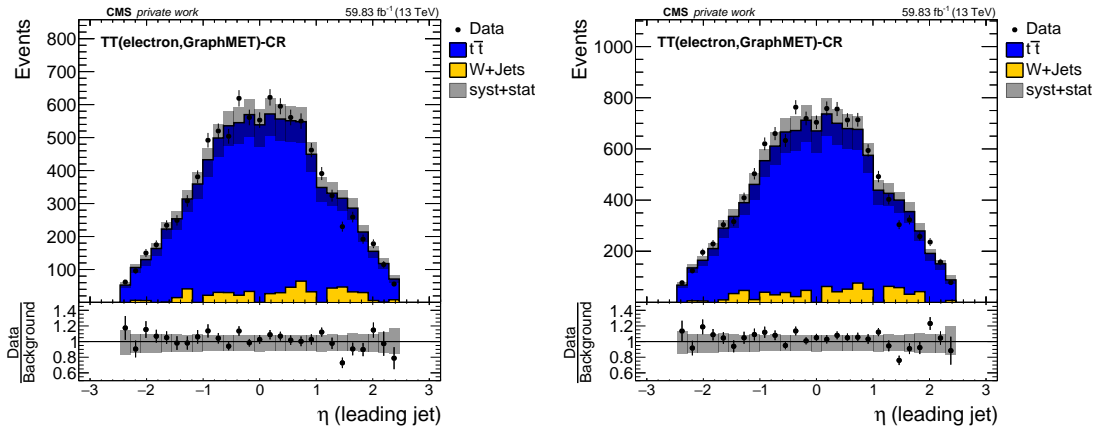
Figure D.26:  $\eta(\text{leading jet})$  distributions in the  $W$  CR with a muon in the final state. In general, the plots show an MC overshoot of roughly 10 – 20 % for all MET reconstruction methods. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other methods. The shapes of DeepMETResolutionTune and GraphMET have sharper edges than the other shapes. This might be due to the better resolution and worse response, letting fewer events with low true MET pass the criterion.

D.27  $t\bar{t}$  Electron CR:  $\eta(\text{leading jet})$ 

(a) MET-Type I (left) and PUPPI MET (right).

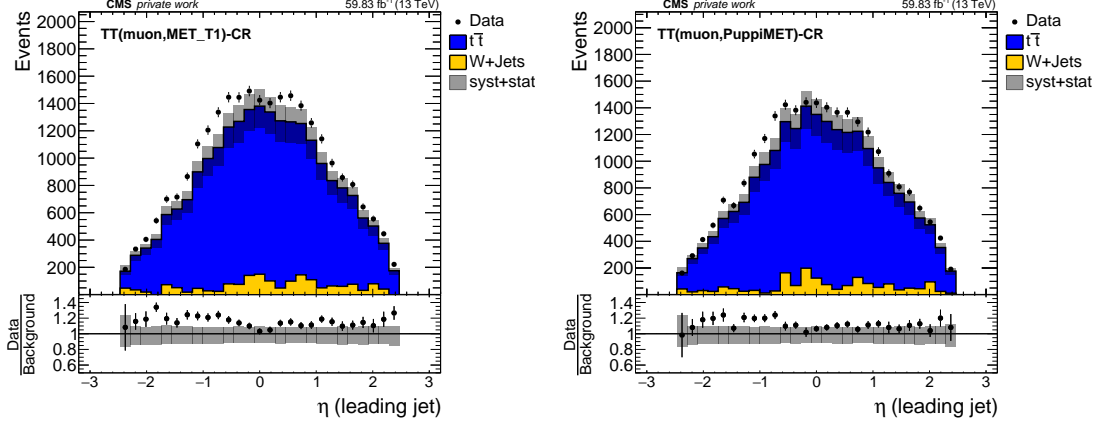


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

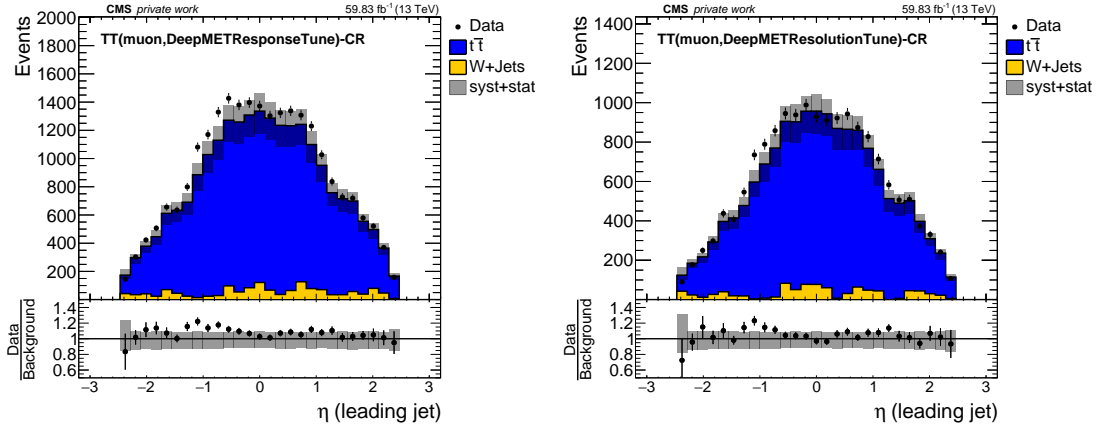


(c) GraphMET (left) and naively calibrated GraphMET (right).

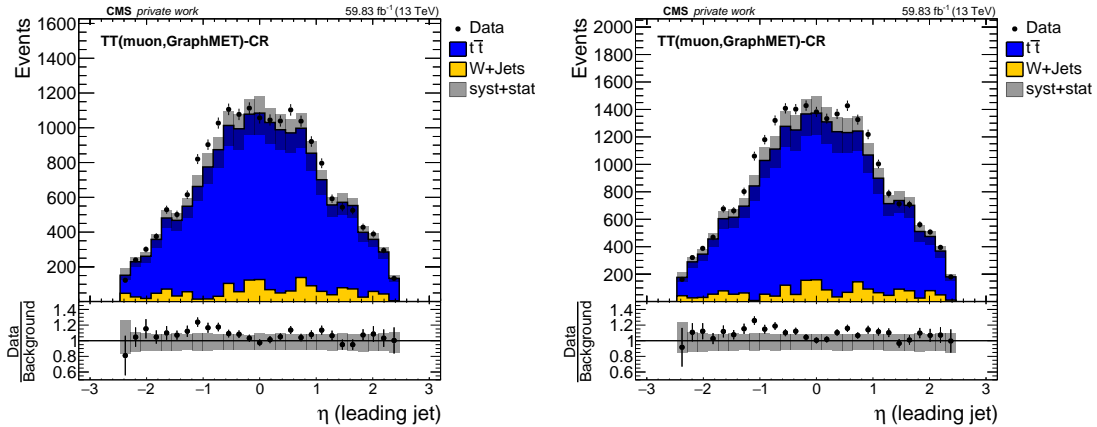
Figure D.27:  $\eta(\text{leading jet})$  distributions in the  $t\bar{t}$  CR with an electron in the final state. In general, the plots show good data-MC agreement and a trend to have sharper edges for GraphMET and DeepMETResolutionTune due to their good resolution and tendency to underestimate MET. Again, the event yields are smaller for these two MET reconstruction methods.

D.28  $t\bar{t}$  Muon CR:  $\eta(\text{leading jet})$ 

(a) MET-Type I (left) and PUPPI MET (right).

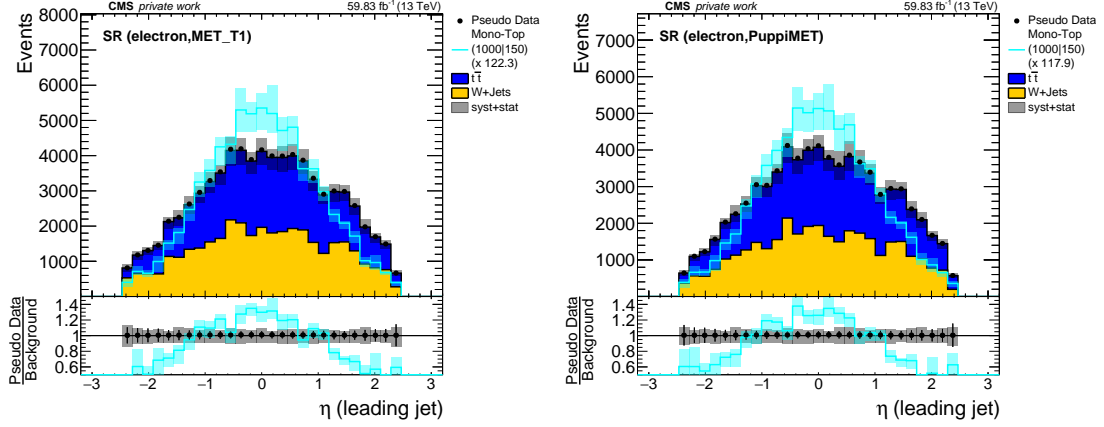


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

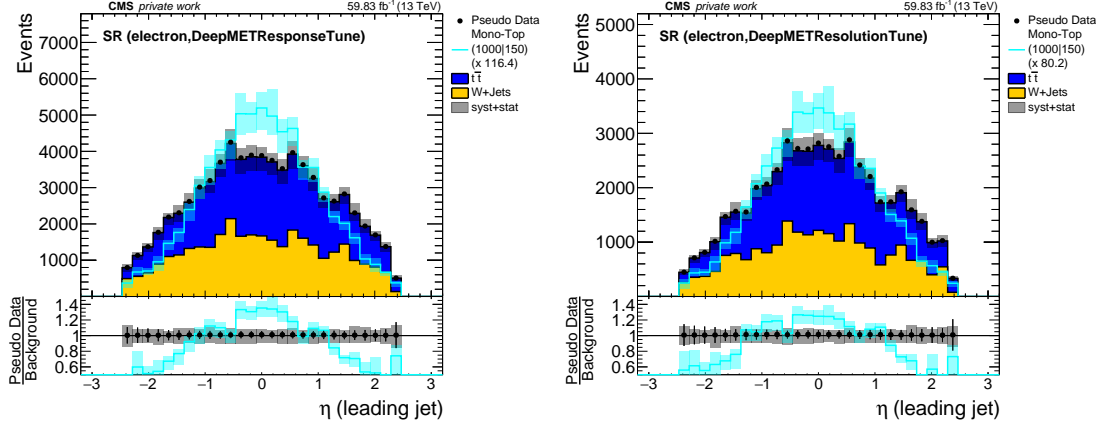


(c) GraphMET (left) and naively calibrated GraphMET (right).

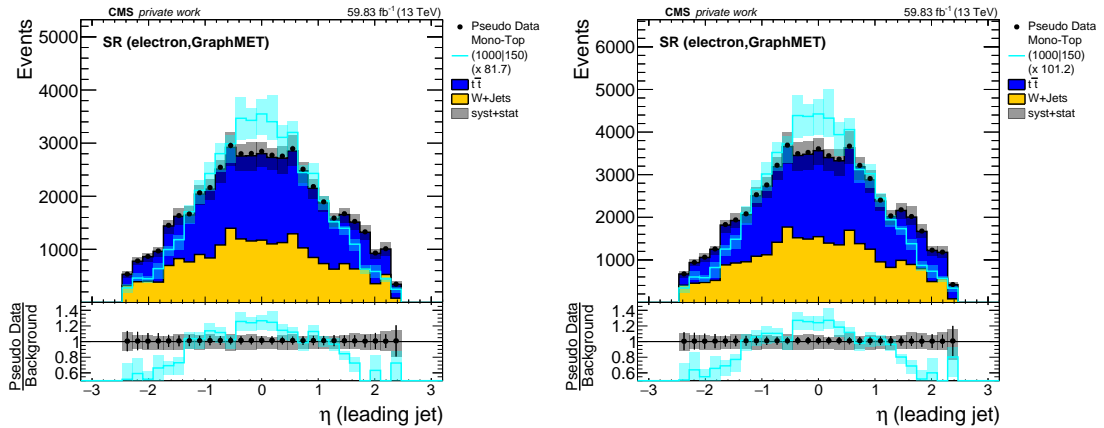
Figure D.28:  $\eta(\text{leading jet})$  distributions in the  $t\bar{t}$  CR with a muon in the final state. Apart from a slight MC undershoot, the event yields and the sharper distribution for GraphMET and DeepMET, the shapes look similar. Again, the sharper edges might be explained by the better resolution and response below one.

D.29 Electron SR:  $\eta$ (leading jet)

(a) MET-Type I (left) and PUPPI MET (right).



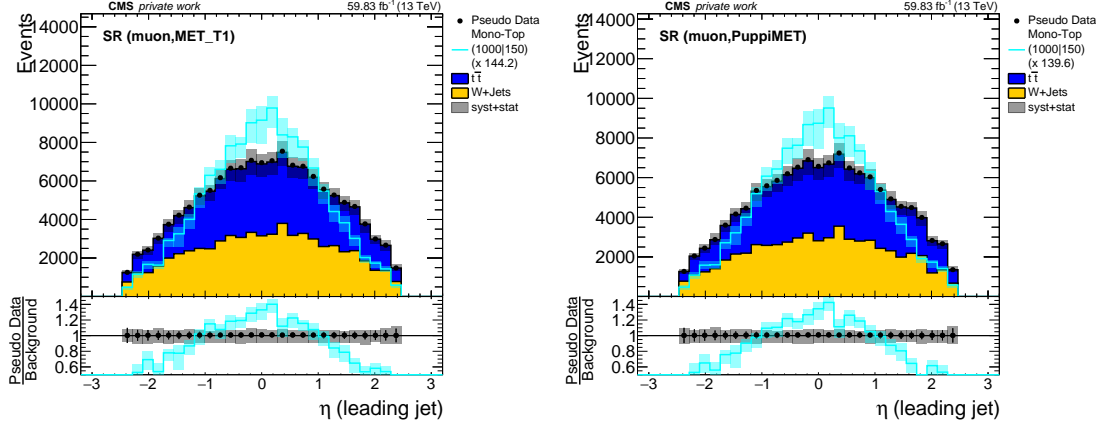
(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).



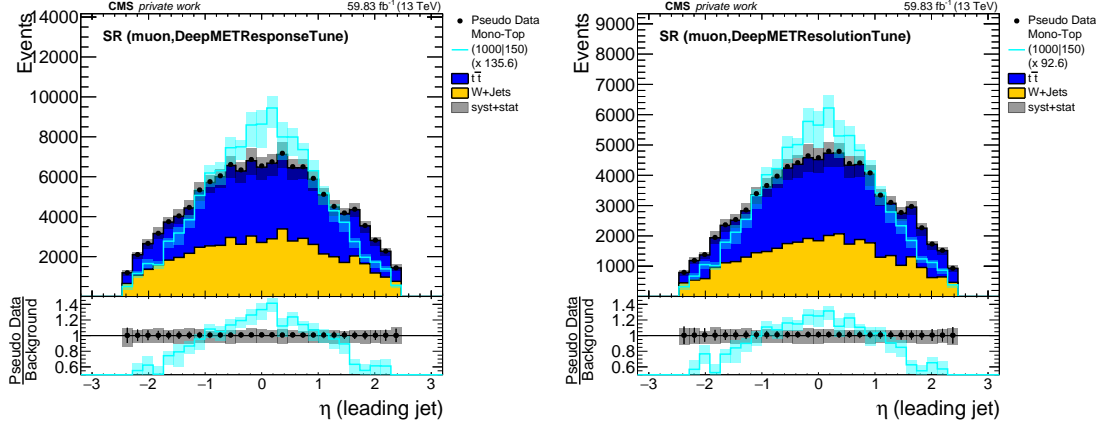
(c) GraphMET (left) and naively calibrated GraphMET (right).

Figure D.29:  $\eta$ (leading jet) distributions in the SR with an electron in the final state. Due to the signal events containing larger amounts of MET, their distributions have sharper edges than the background distributions. However, this effect is far from propagating significantly to pseudodata over background. Again, there are smaller event yields for GraphMET and DeepMETResolutionTune.

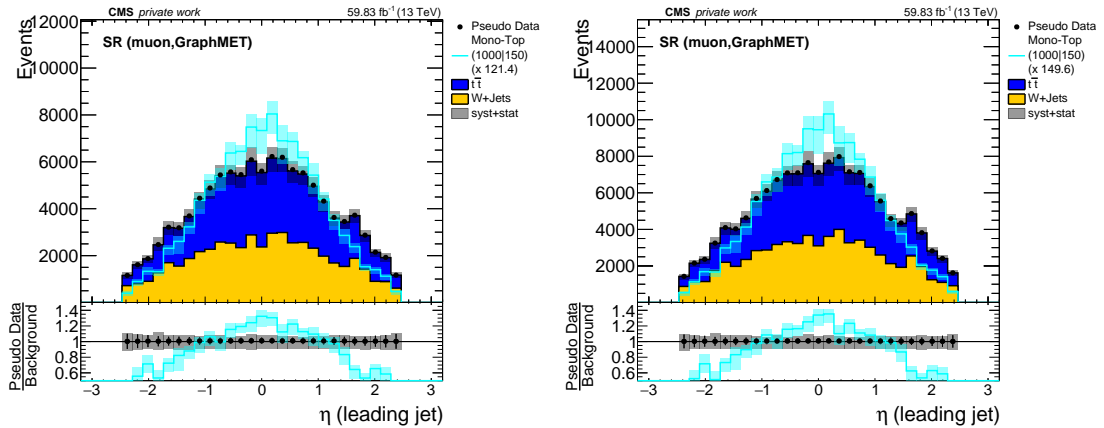


D.30 Muon SR:  $\eta(\text{leading jet})$ 

(a) MET-Type I (left) and PUPPI MET (right).

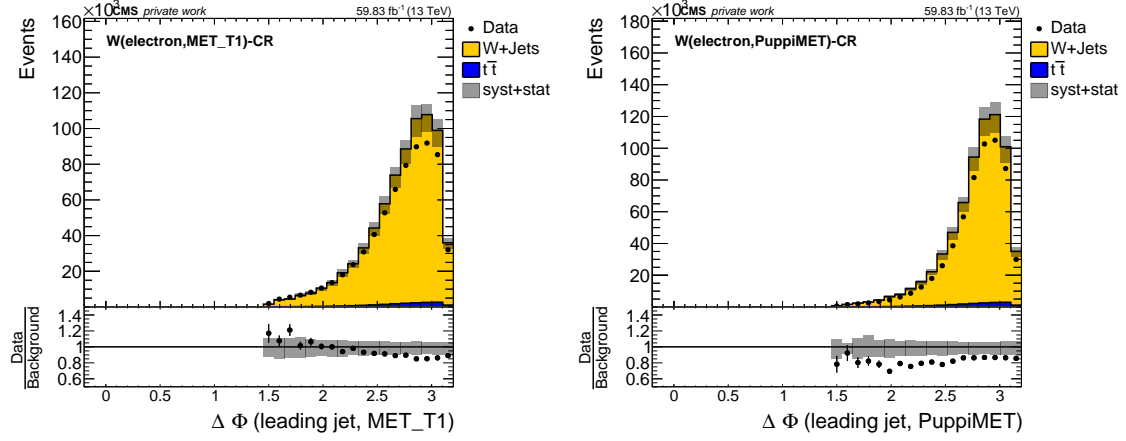


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

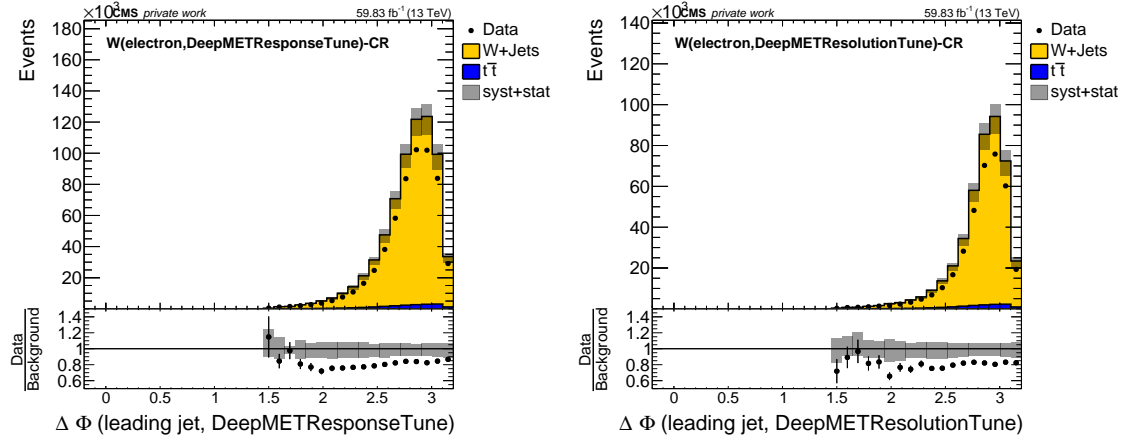


(c) GraphMET (left) and naively calibrated GraphMET (right).

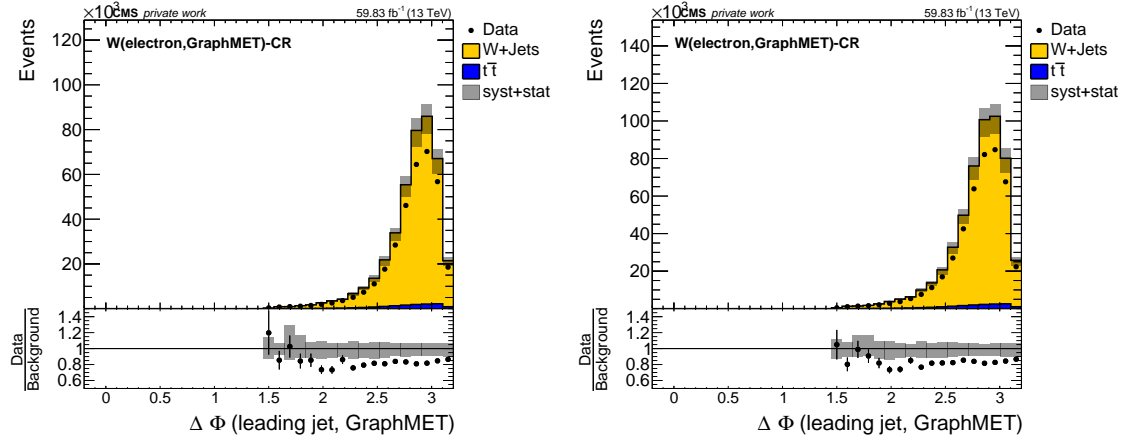
Figure D.30:  $\eta(\text{leading jet})$  distributions in the SR with a muon in the final state. Due to the signal events containing larger amounts of MET, their distributions are sharper than the background. However, this effect is far from propagating significantly to pseudodata over background. Again, there are smaller event yields for GraphMET and DeepMETResolutionTune.

D.31  $W$  Electron CR:  $\Delta\phi(\text{leading jet, MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).

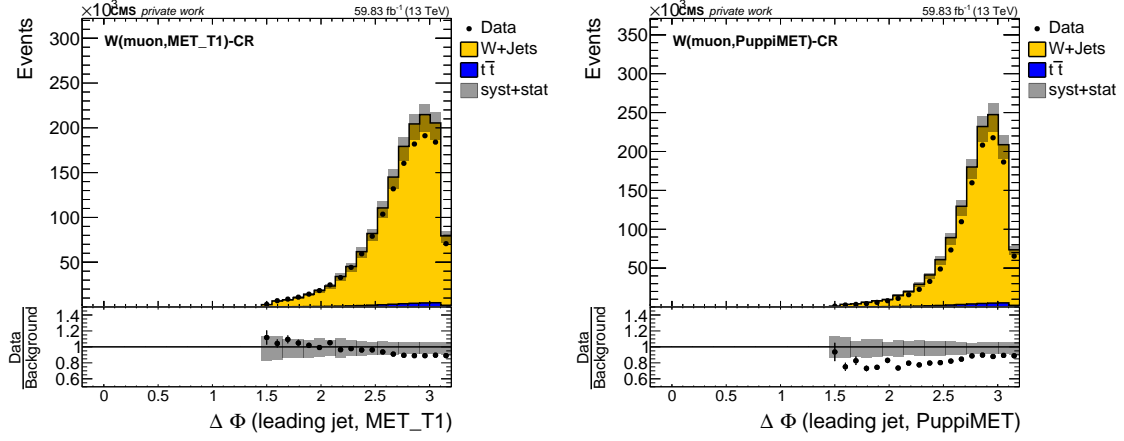


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

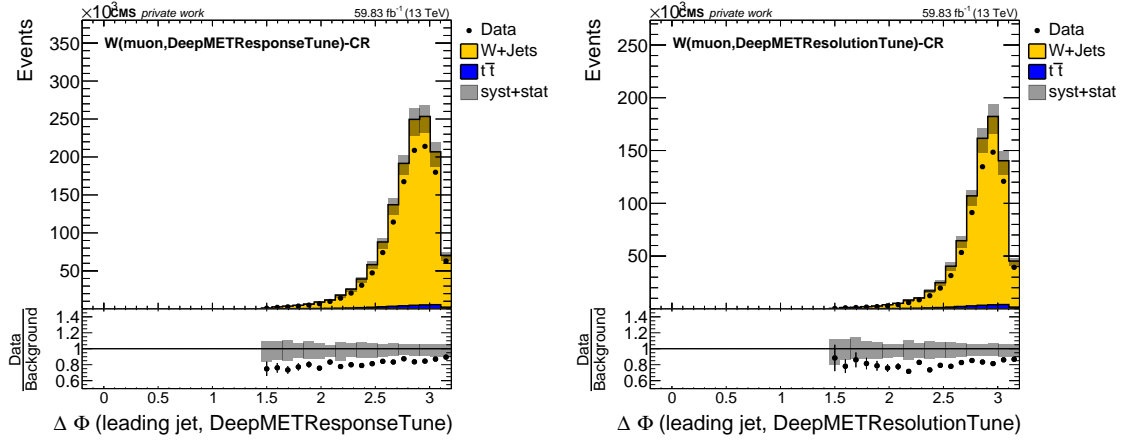


(c) GraphMET (left) and naively calibrated GraphMET (right).

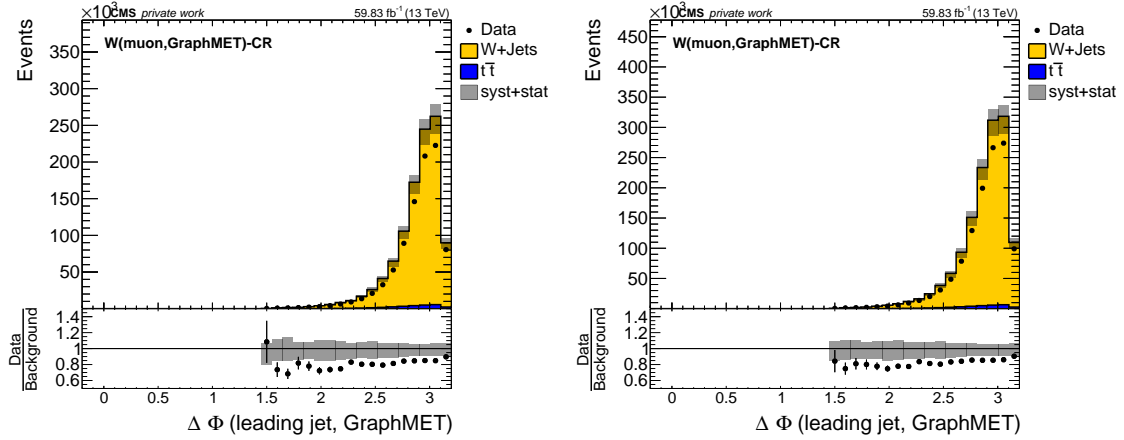
Figure D.31:  $\Delta\phi(\text{leading jet, MET})$  distributions in the  $W$  CR with an electron in the final state. The plots show an MC overshoot of roughly 10–20 % in most bins. The distributions of DeepMET and GraphMET have more entries at larger  $\Delta\phi$  than the other methods as a result of the good resolution reducing smearing effects. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods.

D.32  $W$  Muon CR:  $\Delta\phi(\text{leading jet, MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).



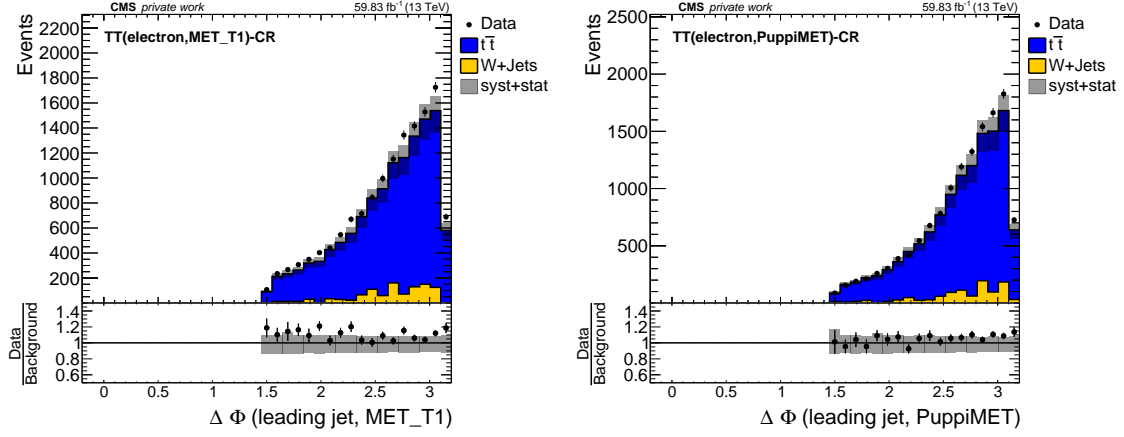
(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).



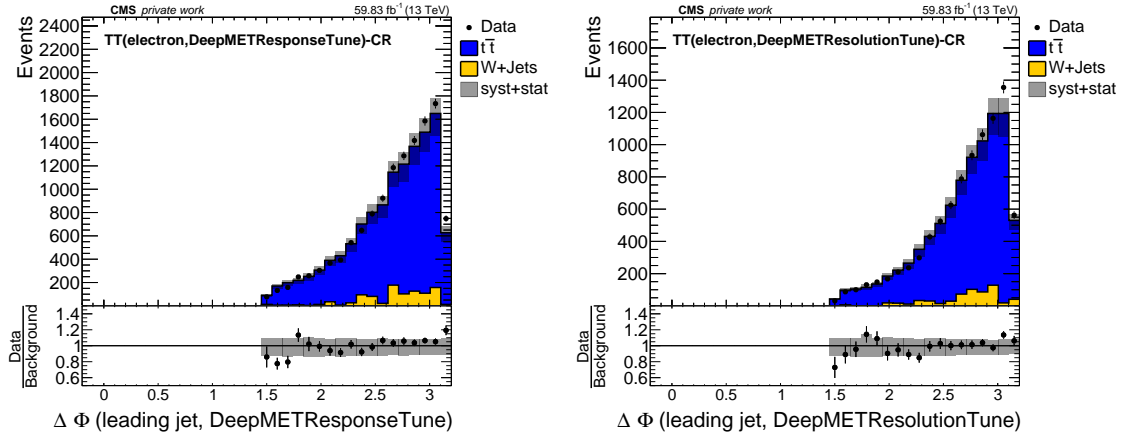
(c) GraphMET (left) and naively calibrated GraphMET (right).

Figure D.32:  $\Delta\phi(\text{leading jet, MET})$  distributions in the  $W$  CR with a muon in the final state. The plots show an MC overshoot of roughly 10–20 % in most bins. The distributions of DeepMET and GraphMET have more entries at larger  $\Delta\phi$  than the other methods as a result of the good resolution reducing smearing effects. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods.

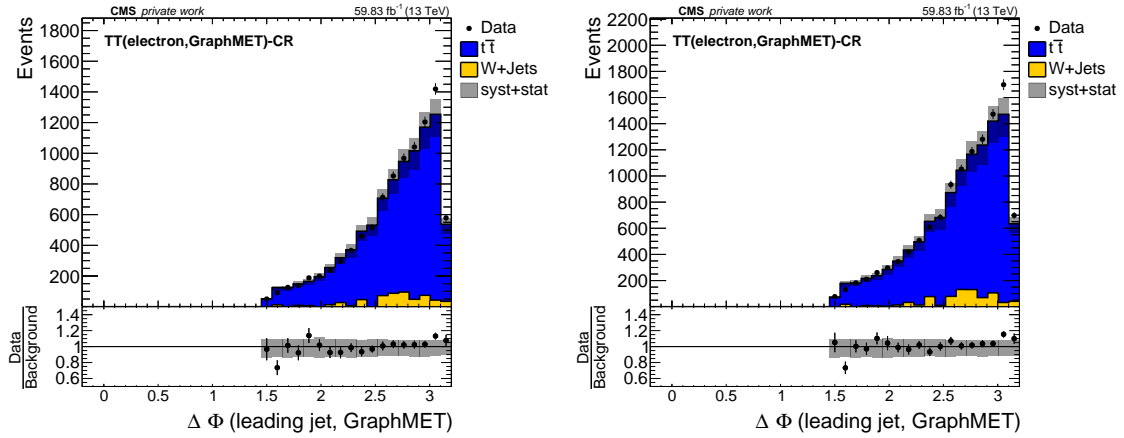
### D.33 $t\bar{t}$ Electron CR: $\Delta\phi(\text{leading jet, MET})$



(a) MET-Type I (left) and PUPPI MET (right).

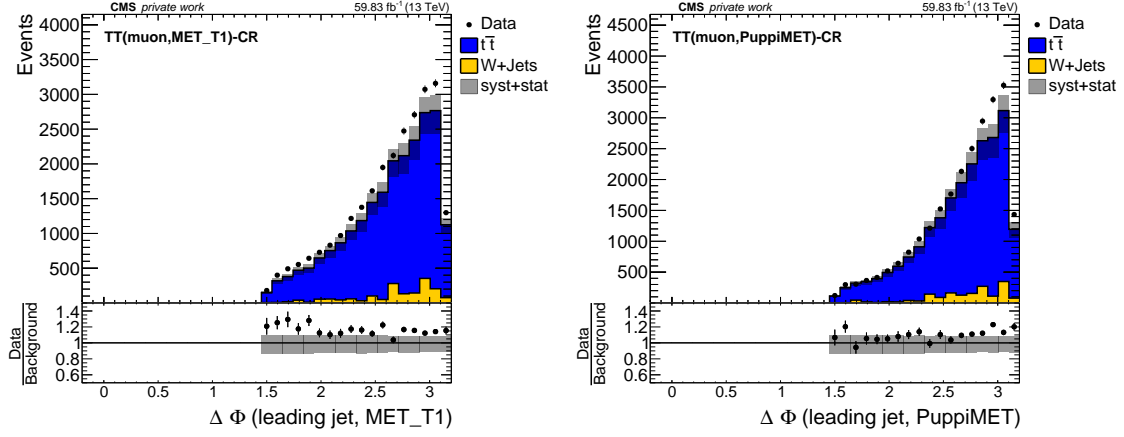


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

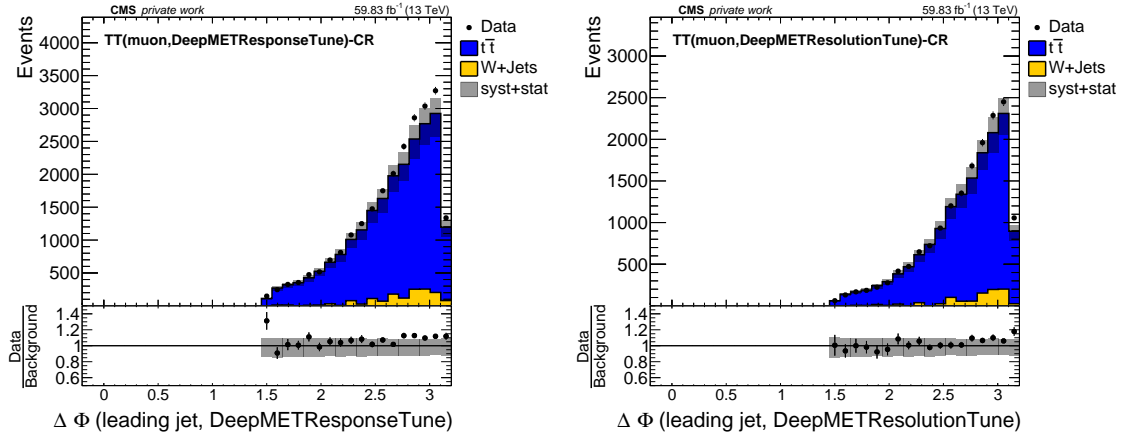


(c) GraphMET (left) and naively calibrated GraphMET (right).

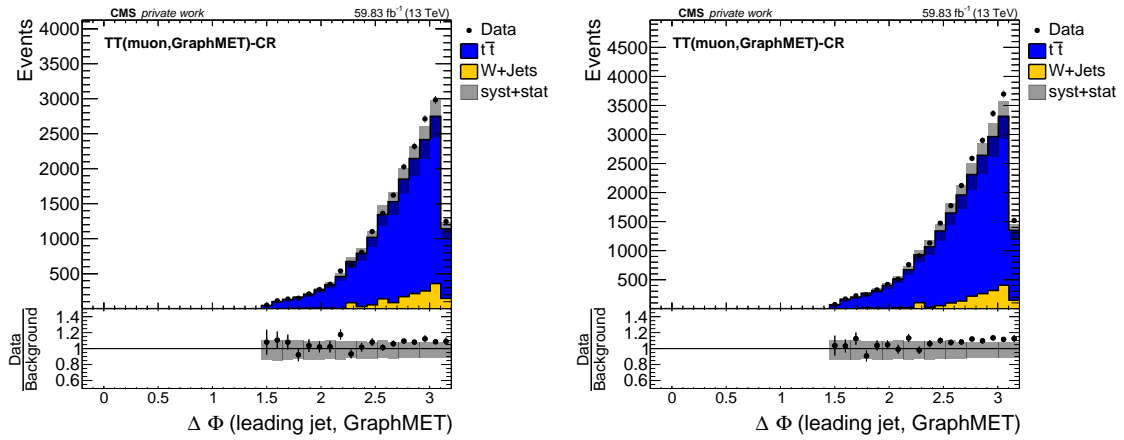
Figure D.33:  $\Delta\phi(\text{leading jet, MET})$  distributions in the  $t\bar{t}$  CR with an electron in the final state. The plots show a reasonable data-MC agreement. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods. The distributions of DeepMET and GraphMET have more entries at larger  $\Delta\phi$  than the other methods, again being a result of the good resolution reducing smearing effects.

D.34  $t\bar{t}$  Muon CR:  $\Delta\phi(\text{leading jet, MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).

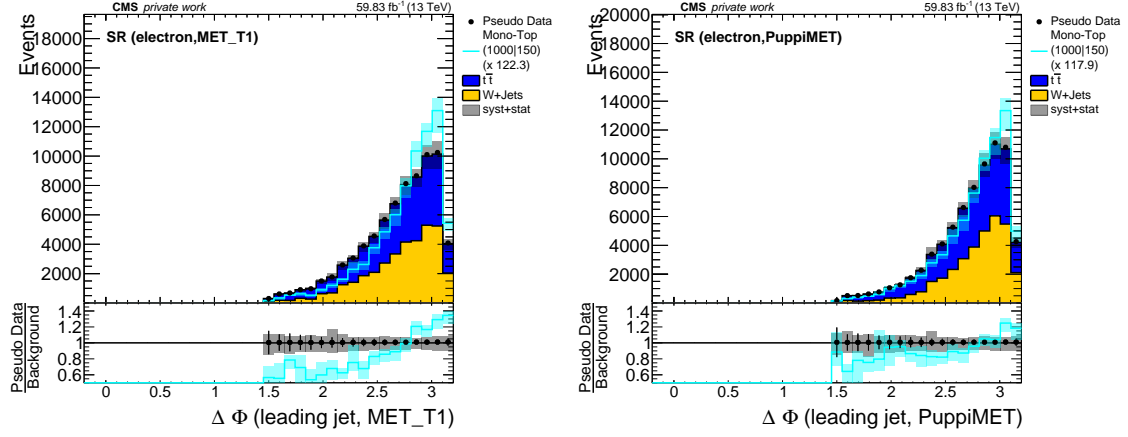


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

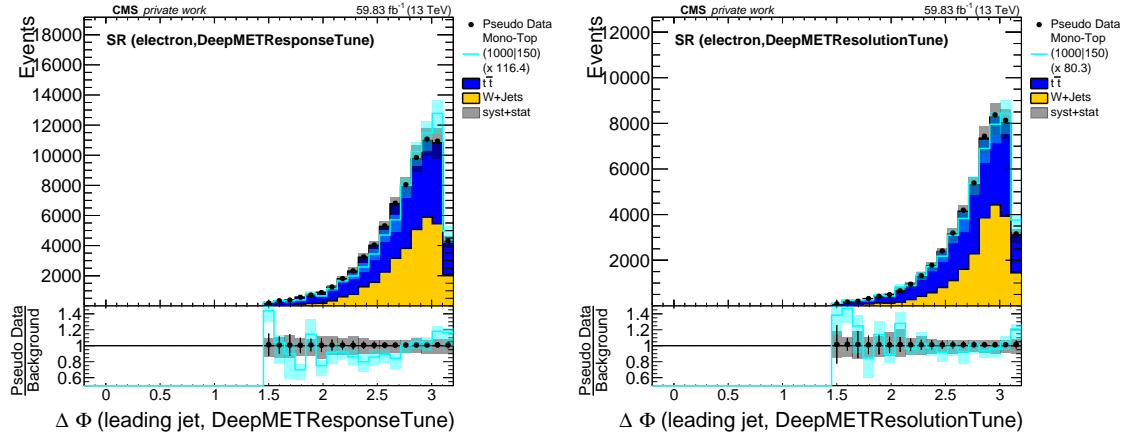


(c) GraphMET (left) and naively calibrated GraphMET (right).

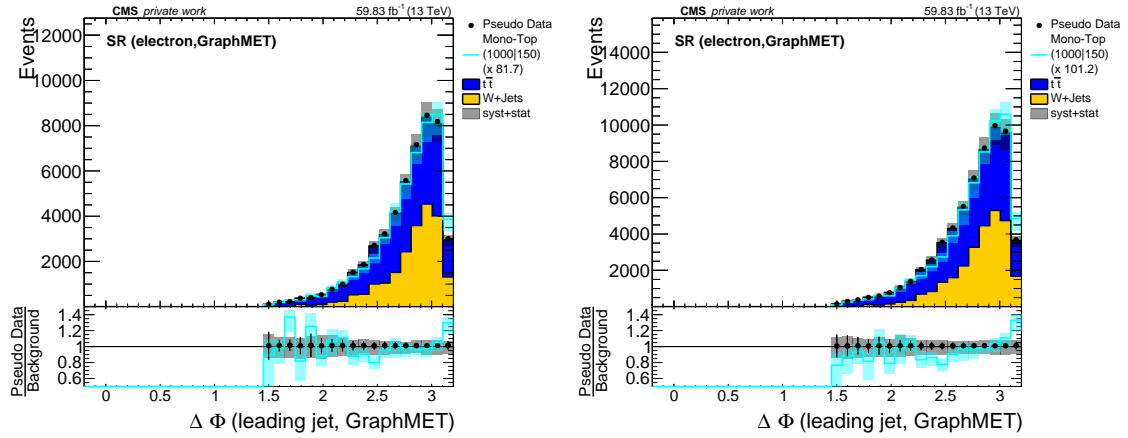
Figure D.34:  $\Delta\phi(\text{leading jet, MET})$  distributions in the  $t\bar{t}$  CR with a muon in the final state. The plots show a good data-MC agreement, except for MET-Type-I. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods. The distributions of DeepMET and GraphMET have more entries at larger  $\Delta\phi$  than the other methods, again being a result of the good resolution reducing smearing effects.

D.35 Electron SR:  $\Delta\phi(\text{leading jet, MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).

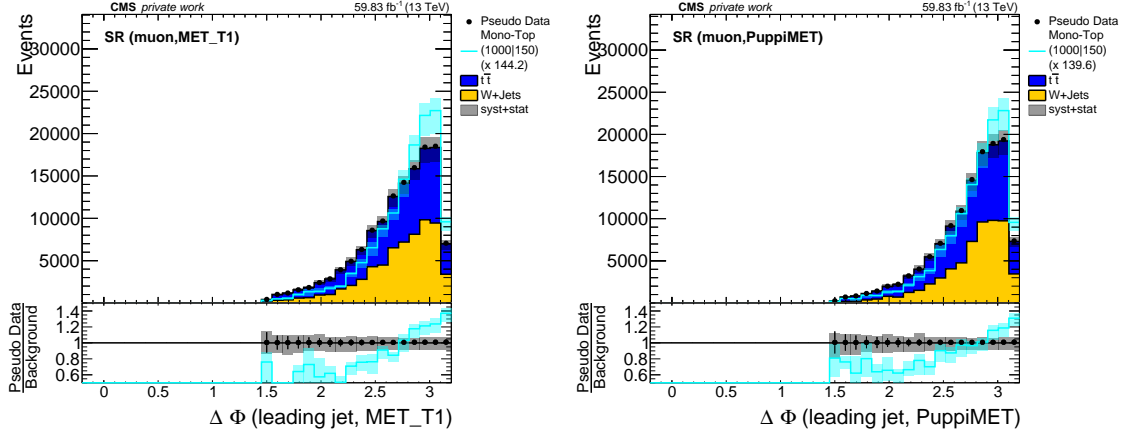


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

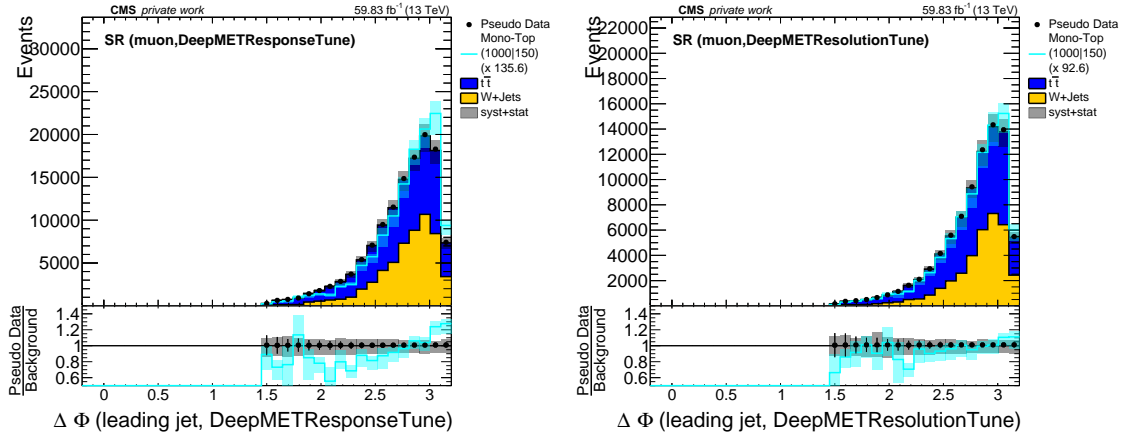


(c) GraphMET (left) and naively calibrated GraphMET (right).

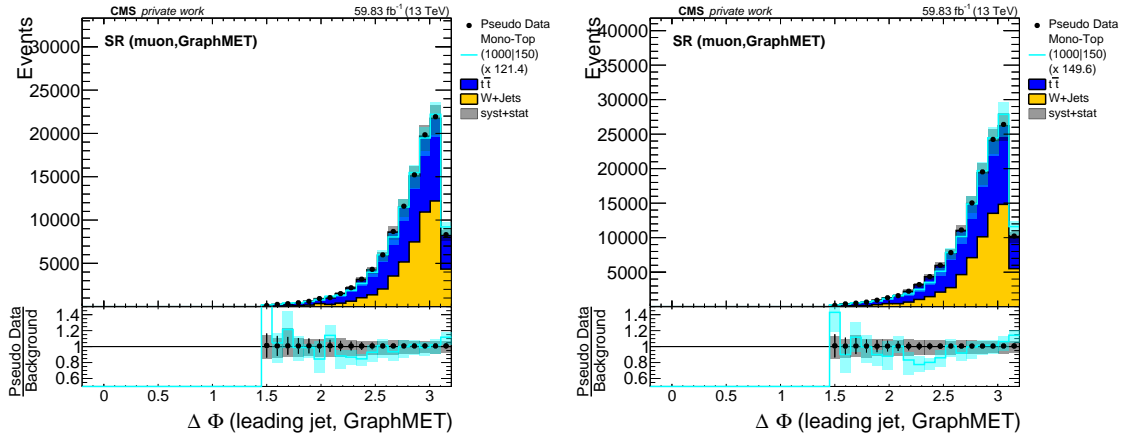
Figure D.35:  $\Delta\phi(\text{leading jet, MET})$  distributions in the SR with an electron in the final state. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods. There is close to no discriminatory power of this variable but the distributions of GraphMET and DeepMET have more entries at larger  $\Delta\phi$  than the other methods, likely as a result of the improved resolution.

D.36 Muon SR:  $\Delta\phi(\text{leading jet, MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).

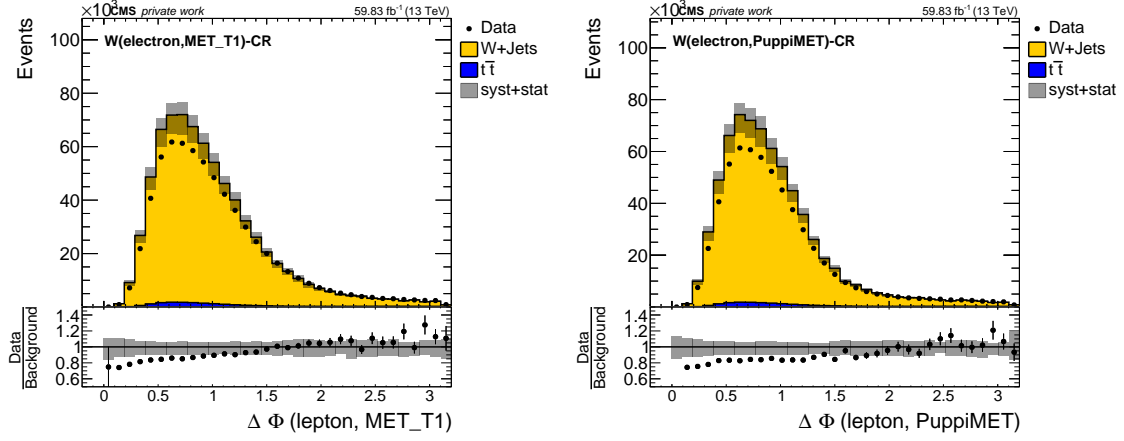


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

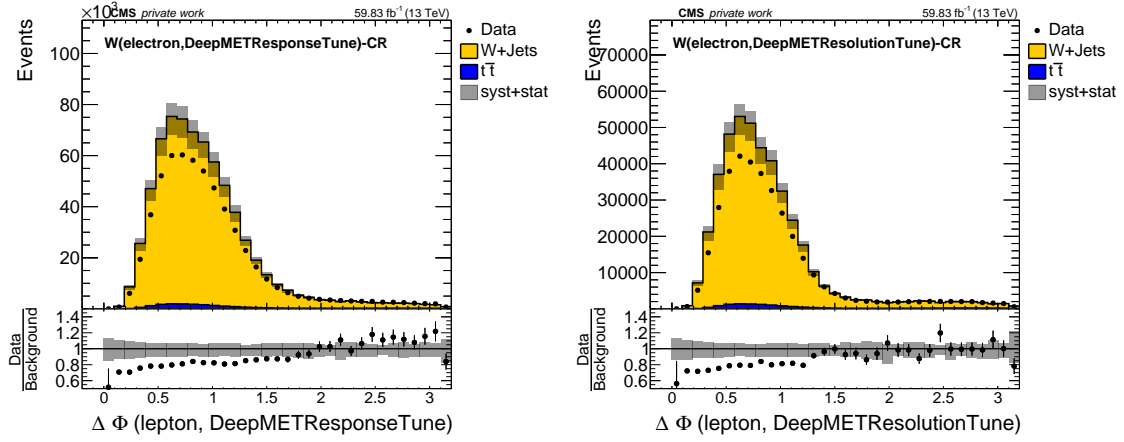


(c) GraphMET (left) and naively calibrated GraphMET (right).

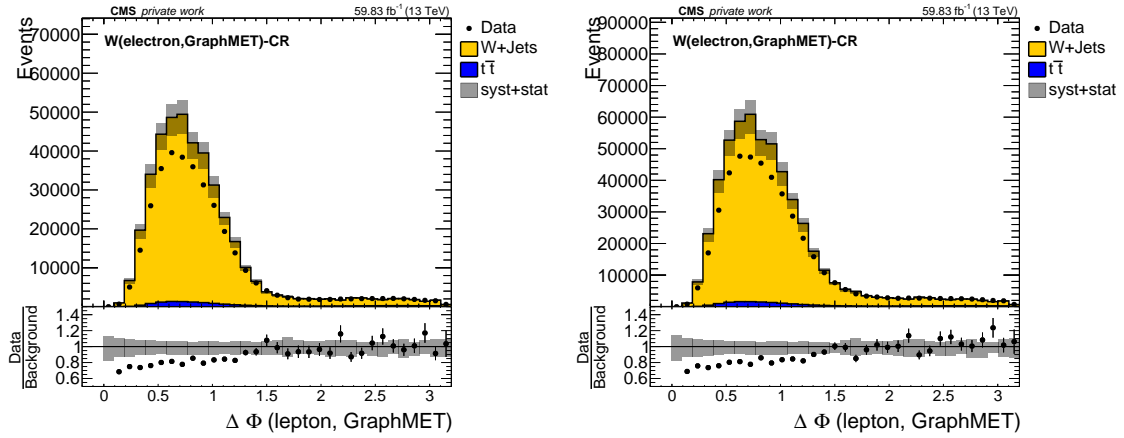
Figure D.36:  $\Delta\phi(\text{leading jet, MET})$  distributions in the SR with a muon in the final state. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods. There is close to no discriminatory power of this variable but the distributions of GraphMET and DeepMET have more entries at larger  $\Delta\phi$  than the other methods, likely as a result of the improved resolution.

D.37  $W$  Electron CR:  $\Delta\phi(\text{lepton}, \text{MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).



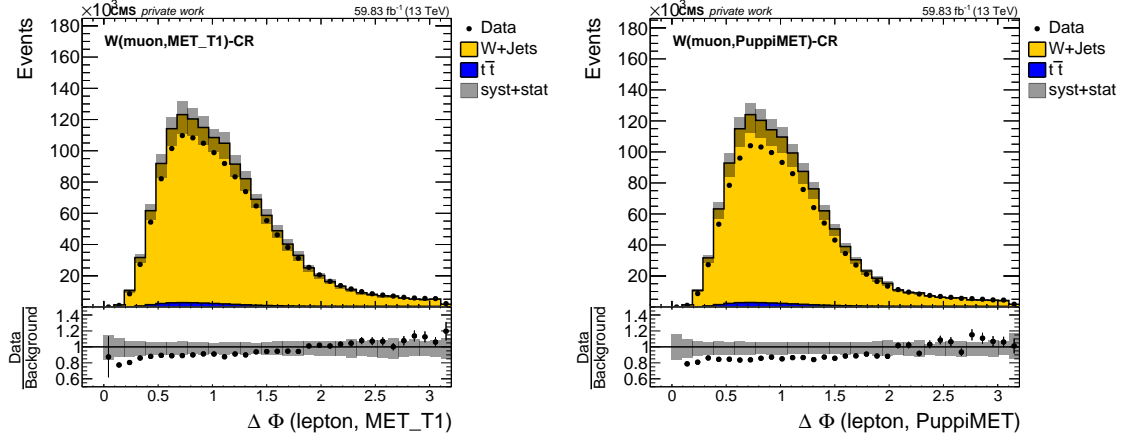
(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).



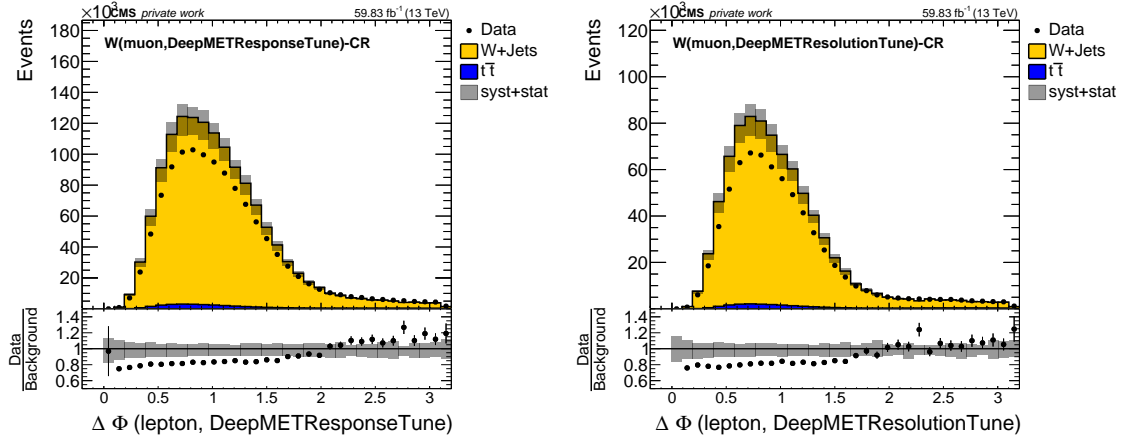
(c) GraphMET (left) and naively calibrated GraphMET (right).

Figure D.37:  $\Delta\phi(\text{lepton}, \text{MET})$  distributions in the  $W$  CR with an electron in the final state. With respect to all events, the plots show an MC overshoot of roughly 10 – 20 % for all methods as the larger bins contain only few events. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other reconstruction methods. The distribution of MET-Type-I and PUPPI MET seem to be more smeared as a result of the worse resolution.

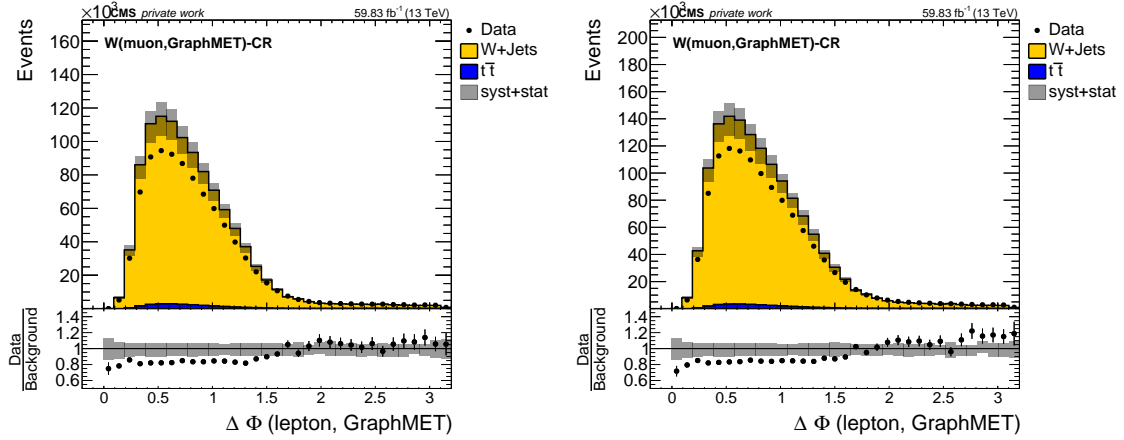


D.38  $W$  Muon CR:  $\Delta\phi(\text{lepton}, \text{MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).

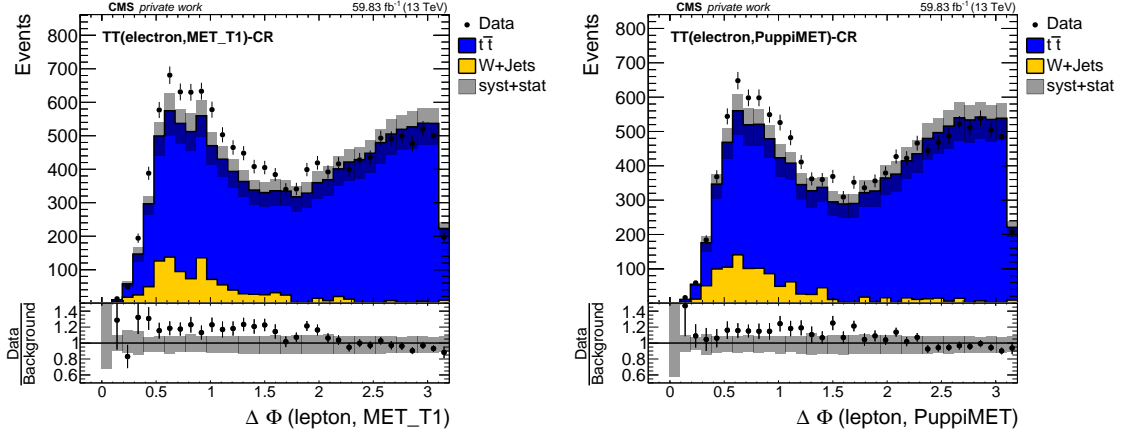


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

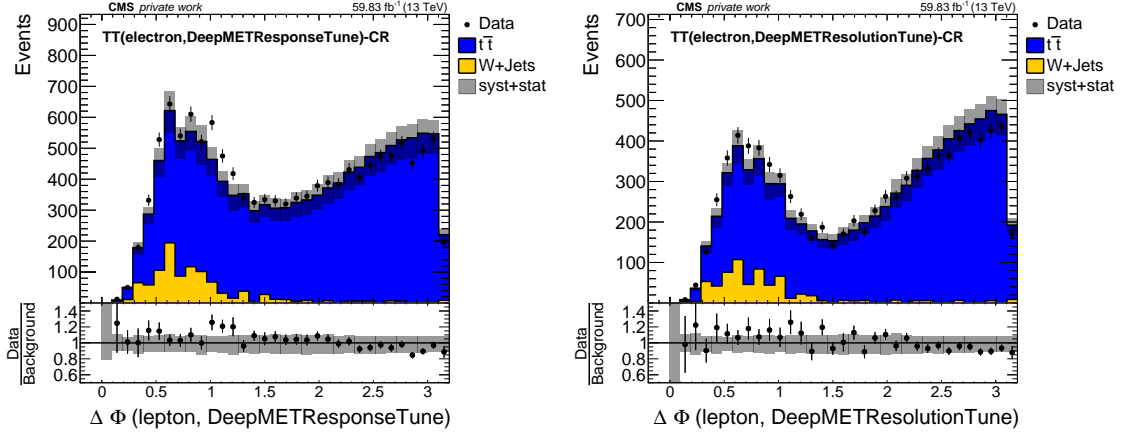


(c) GraphMET (left) and naively calibrated GraphMET (right).

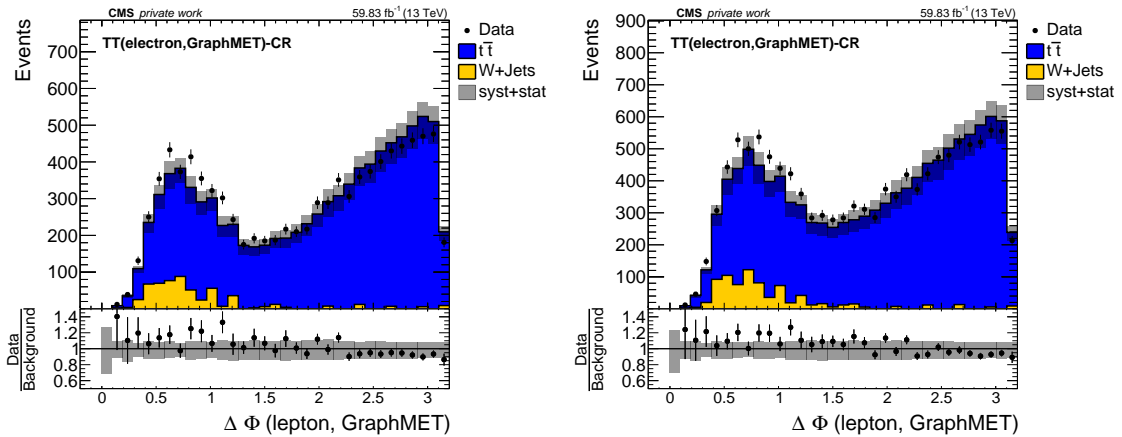
Figure D.38:  $\Delta\phi(\text{lepton}, \text{MET})$  distributions in the  $W$  CR with a muon in the final state. With respect to all events, the plots show an MC overshoot of roughly 10–20 % for all methods as the larger bins contain only few events. The event yields of GraphMET and DeepMETResolutionTune are smaller than those of the other methods. The distribution of MET-Type-I and PUPPI MET seem to be more smeared as a result of the worse resolution.

D.39  $t\bar{t}$  Electron CR:  $\Delta\phi(\text{lepton}, \text{MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).

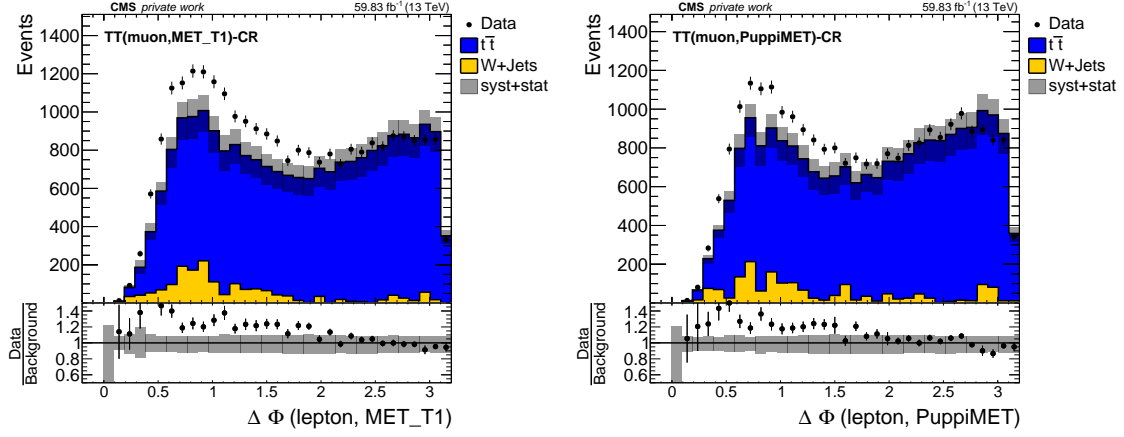


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

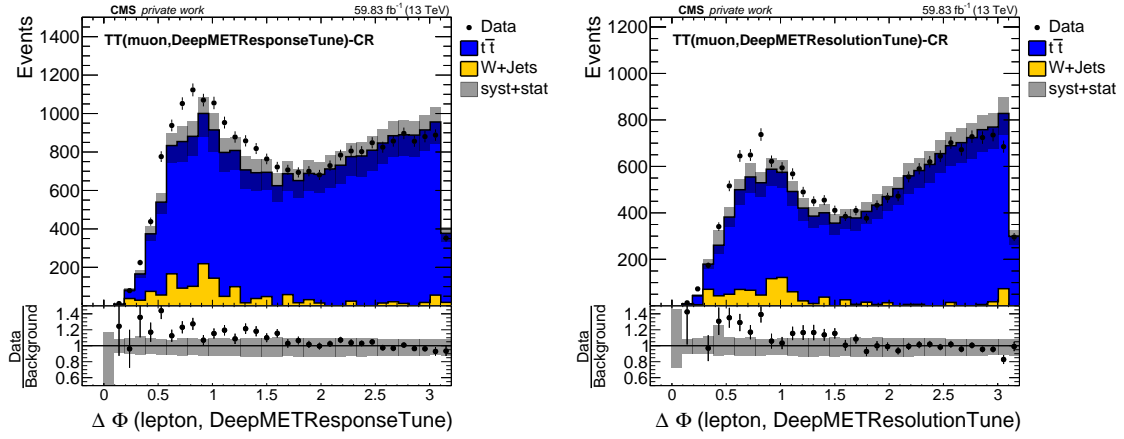


(c) GraphMET (left) and naively calibrated GraphMET (right).

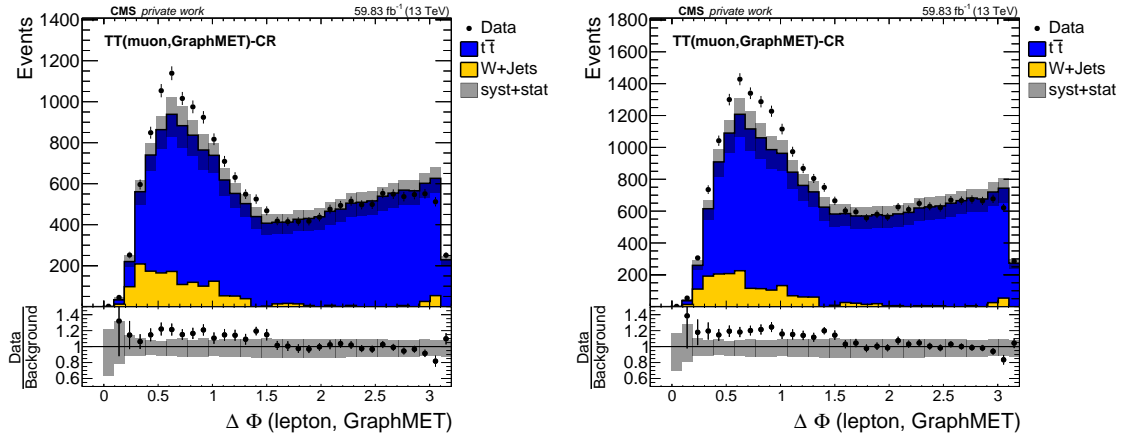
Figure D.39:  $\Delta\phi(\text{lepton}, \text{MET})$  distributions in the  $t\bar{t}$  CR with an electron in the final state. The larger angle differences are modeled well, whereas the smaller angle differences are challenging for the simulation, especially when using MET-Type-I. These smaller regions are less populated for GraphMET and DeepMETResolutiontune, which therefore have a smaller total event yield.

D.40  $t\bar{t}$  Muon CR:  $\Delta\phi(\text{lepton}, \text{MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).

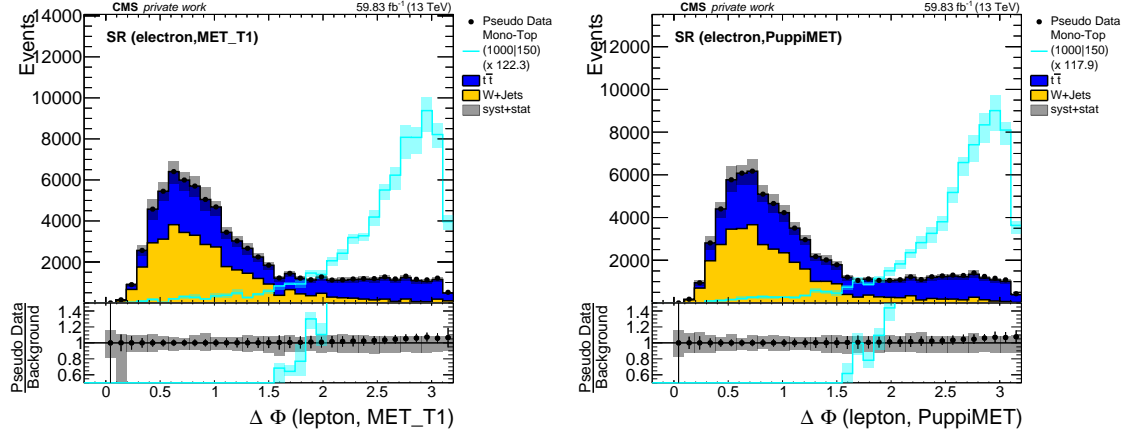


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

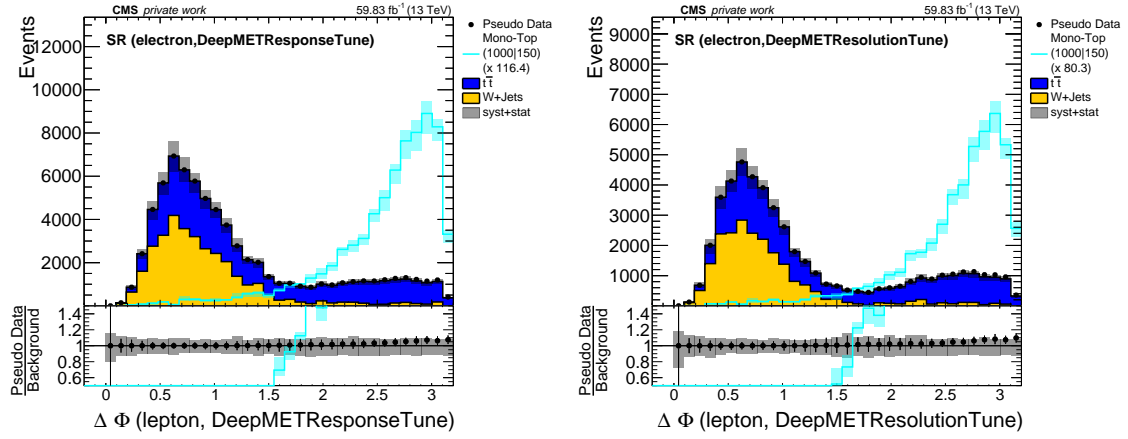


(c) GraphMET (left) and naively calibrated GraphMET (right).

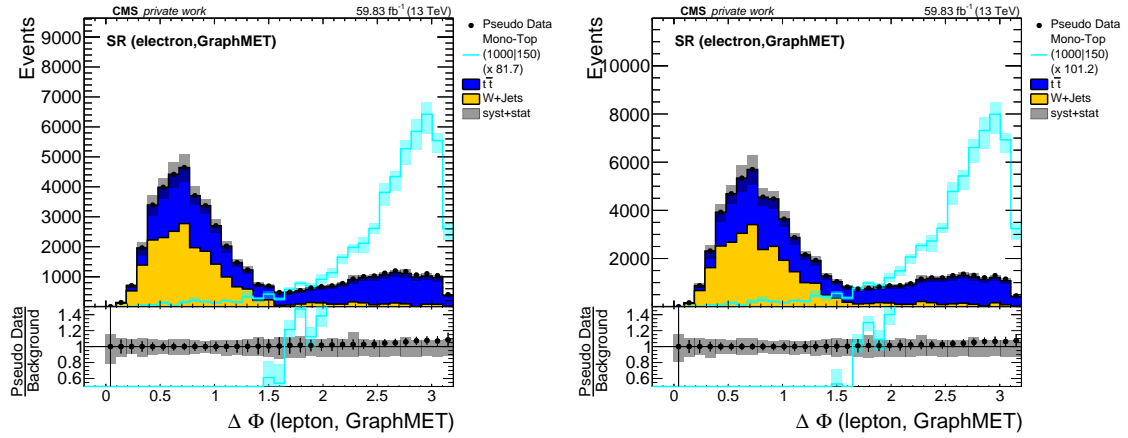
Figure D.40:  $\Delta\phi(\text{lepton}, \text{MET})$  distributions in the  $t\bar{t}$  CR with a muon in the final state. The larger angle differences are modeled well, whereas the smaller ones are challenging for the simulation, especially when using MET-Type-I or PUPPI MET. These smaller regions are less populated for DeepMETResolutiontune. The higher  $p_T$  regions are less populated for GraphMET. As a consequence, both distributions have a smaller total event yield than the others.

D.41 Electron SR:  $\Delta\phi(\text{lepton}, \text{MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).

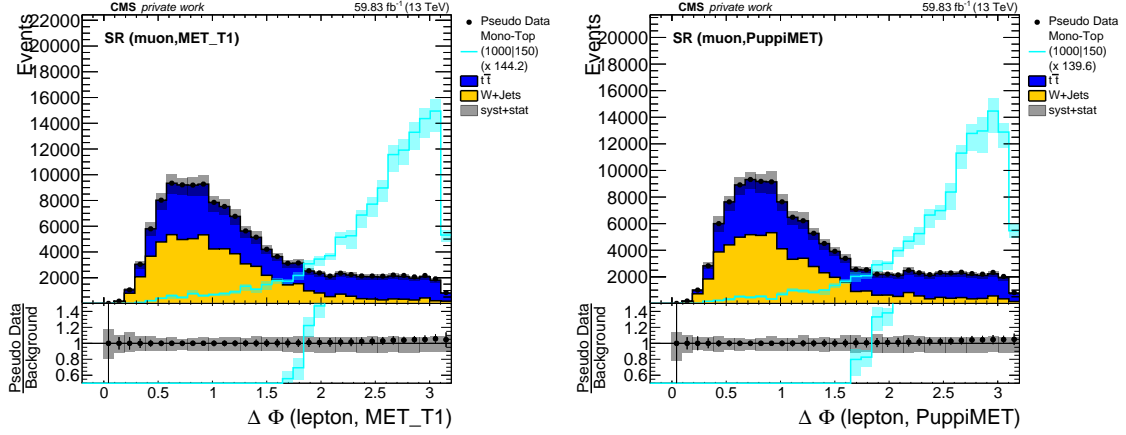


(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).

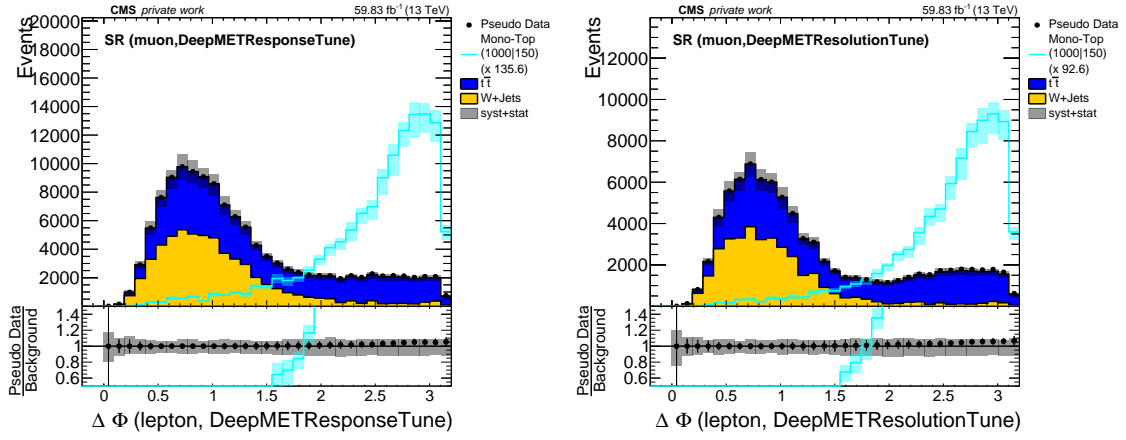


(c) GraphMET (left) and naively calibrated GraphMET (right).

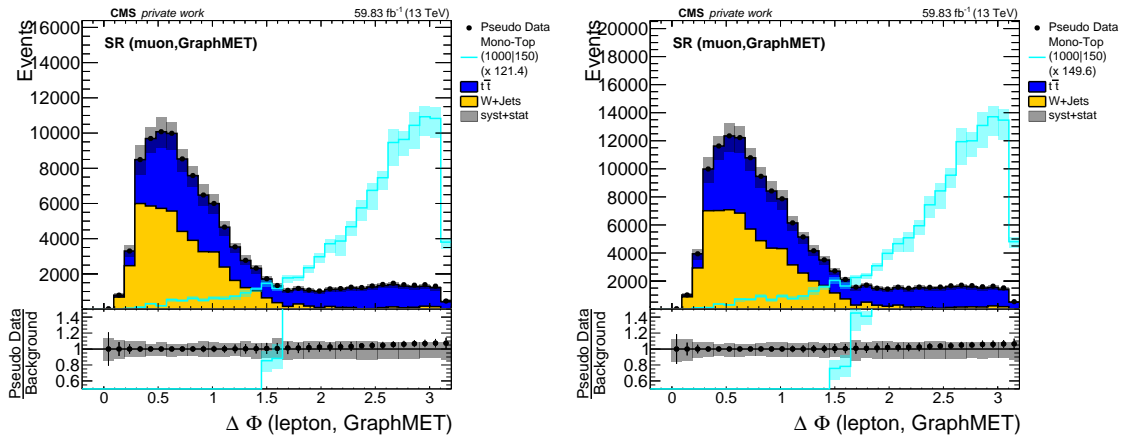
Figure D.41:  $\Delta\phi(\text{lepton}, \text{MET})$  distributions in the SR with an electron in the final state. The distributions appear to be quite similar. Apart from the smaller event yields for GraphMET and DeepMETResolutionTune, there are no major differences. The signal to background ratio is best for GraphMET and DeepMETResolutionTune as their signal multiplication factor needed to match the background event yield is the smallest.

D.42 Muon SR:  $\Delta\phi(\text{lepton}, \text{MET})$ 

(a) MET-Type I (left) and PUPPI MET (right).



(b) DeepMETResponseTune (left) and DeepMETResolutionTune (right).



(c) GraphMET (left) and naively calibrated GraphMET (right).

Figure D.42:  $\Delta\phi(\text{lepton}, \text{MET})$  distributions in the SR with a muon in the final state. The distributions appear to be quite similar. However, the discrimination power seems to be best for GraphMET as their share of background events is larger at smaller angular distances. The event yields are smaller when using GraphMET or DeepMETResolutionTune.

## E Significance Analysis with Naively Calibrated GraphMET

Using the naively calibrated GraphMET values, the expected median asymptotic limits are calculated and compared to the other MET reconstruction methods in order to investigate whether the improvement in the limit is caused by the constructed improvement in the signal to background ratio. For that, every GraphMET value is divided by the median response for large MET, that is 92%. Thus, the response is shifted to one while the resolutions are also increased by the inverse median response. Hence, the limits are expected to be a bit worse, as the signal to background ratio as well as the resolution are worsened.

Figure E.1 shows the upper limits for stat-only and including systematics. As GraphMET has still the lowest limits, there is reason to believe that the improvement with respect to the alternative methods is due to GraphMET's superior performance and not by construction.

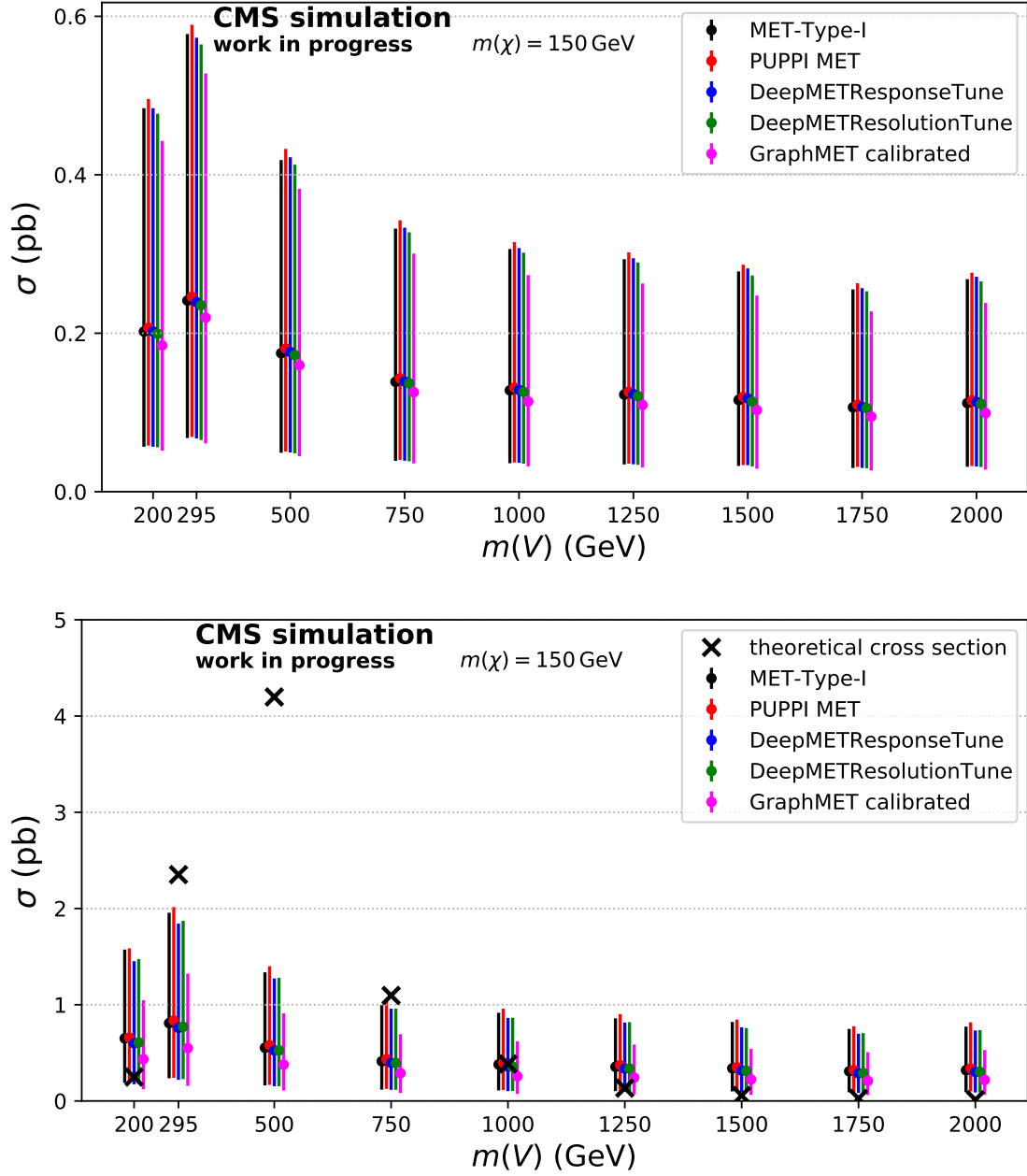


Figure E.1: Expected asymptotic upper limits on the DM production cross section for different mediator masses and a DM mass of  $m_\chi = 150 \text{ GeV}$  using stat-only (at the top) and systematics-included (at the bottom) combined limits in the SR. The dots are the expected asymptotic upper limits and the error bars indicate their 16% and 84% quantiles. The investigated mass points indicate a stronger limit to the cross section when using the naively calibrated GraphMET version with respect to the other reconstruction methods.