MACHINE LEARNING FOR TOP QUARK PHYSICS

CERN-THESIS-2022-101
07/07/2022

EMIL SØRENSEN BOLS

PhD 2022

VRIJE
UNIVERSITEIT
BRUSSEL

Proefschrift ingediend met het oog op het behalen van de academische graad
Doctor in de Wetenschappen

MACHINE LEARNING FOR TOP QUARK PHYSICS

Emil Sørensen Bols

July 7, 2022

Promotor:	Prof. Dr. Jorgen D'Hondt
Jury:	Prof. Dr. Nick Van Eijndhoven (VUB), chairman
	Prof. Dr. Alberto Mariotti (VUB), secretary
	Prof. Dr. Wouter Verkerke (NIKHEF, UvA)
	Dr. Maria Aldaya Martin (DESY)
	Prof. Dr. Steven Lowette (VUB)
	Prof. Dr. Pieter Libin (VUB)

Faculteit Wetenschappen en Bio-ingenieurswetenschappen
Departement Natuurkunde

Printed by Crazy Copy Center Productions VUB Pleinlaan 2, 1050 Brussel
Tel : +32 2 629 33 44 crazycopy@vub.be www.crazycopy.be

ISBN: 9789464443349

NUR CODE: 926

THEMA: PHP

Abstract

In the last 5 years, machine learning algorithms, in particular the neural network, have proven to be a very powerful tool for high energy physics at the LHC. In the realm of top quark physics, machine learning has risen to prominence both in event selection and event reconstruction.

In this thesis a Deep Neural Network for jet flavour identification is presented. It is capable of identifying b jets, c jets, light quark jets and gluon jets. By utilizing a novel neural network architecture that can efficiently exploit the full jet information it achieves state of the art performance in each of the jet classification tasks. This neural network is extended further to estimate jet energy corrections. Since the jet energy response depends on the flavour of the jet, the architecture and inputs for jet flavour identification can be utilized for making a general jet energy correction, which leverages the flavour information going beyond the standard approach of jet energy corrections that has no jet flavour dependence.

Machine learning has not just found success for particle physics reconstruction, but it is also heavily used for particle physics event selection. In this case the machine learning methods are optimized to minimize the statistical uncertainty on the measurement by increasing selection efficiency and reducing the rate of background. However in modern particle physics analyses the systematic uncertainty is the dominant component of the total uncertainty. In this thesis a novel machine learning method is developed that makes an event selection that reduces the systematic uncertainty. A showcase of the method is done in the setting of a top quark mass measurement.

Samenvatting

In de afgelopen 5 jaar hebben machine learning-algoritmen, in het bijzonder het neurale netwerk, bewezen een zeer krachtig hulpmiddel te zijn voor hoge-energiefysica aan de LHC. Op het gebied van de fysica van top-quarks is machine learning uitgegroeid tot een veelbelovende applicatie zowel bij de selectie als bij de reconstructie van deeltjesbotsingen.

In dit proefschrift wordt een diep neuraal netwerk voor jet-smaak identificatie gepresenteerd. Het is in staat om b-jets, c-jets, lichte quark-jets en gluon-jets te identificeren. Gebruikmakend van een nieuwe netwerkarchitectuur die op een efficiënte manier de volledige verzameling jet-informatie kan benutten, bereikt het netwerk een technisch hoogtepunt wat betreft de prestaties in elk van de jet-classificatietaken. Dit neurale netwerk wordt verder uitgebreid om jet-energiecorrecties te schatten. Aangezien de respons van de jet-energie afhangt van de smaak van de jet, kunnen de architectuur en inputs voor jet-smaak identificatie ook benut worden voor het maken van een algemene jet-energiecorrectie, die gebruik maakt van de smaakinformatie en daardoor voorbij de standaardmethode voor jet-energiecorrecties gaat die geen jet-smaak afhankelijkheid in rekening brengt.

Machine learning heeft niet enkel succes gevonden bij de reconstructie van deeltjesbotsingen, maar het wordt ook veel gebruikt bij de selectie van deeltjesbotsingen. In dit geval zijn de machine learning-methoden geoptimaliseerd om de statistische onzekerheid op de metingen te minimaliseren door de selectie-efficiëntie te verhogen en de achtergrond te verminderen. In de analyse van een modern deeltjesfysica experiment vormt de systematische onzekerheid echter de dominante

component van de totale onzekerheid. In dit proefschrift is er een nieuwe machine learning-methode ontwikkeld die een selectie maakt van deeltjesbotsingen zodat deze systematische onzekerheid verminderd wordt. Een toepassing hiervan wordt gedemonstreerd in de context van een top-quark massa meting.

Acknowledgements

First I would like to thank my supervisor Jorgen D'Hondt for always being supportive and inspiring. I am very grateful for your great advice on physics and life, as well as for being a supervisor that gave me freedom throughout the development of my projects.

I would like to thank my collaborators on DeepJet, Mauro and Jan. When I got involved in the project, your guidance was invaluable, and it prompted my interest in machine learning. The general BTV group has been a wonderful working environment with countless of great people. Particularly I would like to thank Kirill, Ivan, and Seth for getting me involved in the group. I would also like to extend a thanks to my L3 co-convener Andrzej for the pleasant working relationship.

I would also like to thank Petra van Mulders for providing me with her original DELPHES ReSYST dataset, as well as advice for getting started with the method, while also being a great person to have as a coworker. Additionally, I would like to thank Hartmut Stadie for providing me with the CMS top quark mass analysis framework.

Thanks to Santiago Paredes, Laurent Thomas and Andrey Popov for the discussions on DeepJet JEC.

Thanks to all my colleagues from the IIHE for creating a wonderful working environment, and in particular Annemie, Alexandre, AR, Nordin, Denise, Diego, Jarne, Senne, Lieselotte, Douglas, Alexander, Inna, Denys, Simon, Christopher and Ali. Additionally, a big thanks goes to our secretaries Sophie and Marleen!

I am grateful to my parents, brother and sister for their continuous love and support despite me being away from home for so long. Finally I would like to thank Elsa! I am so grateful and happy to have you by my side!

Contents

1. Introduction **1**

2. Top Quark Physics at the LHC **5**

2.1. The Standard Model 5

2.1.1. The Top Quark 15

2.2. The LHC 20

2.3. The CMS experiment 23

2.3.1. The Silicon Tracker 25

2.3.2. Electromagnetic Calorimeter 28

2.3.3. Hadronic Calorimeter 30

2.3.4. Muon detectors 31

2.3.5. Trigger system 33

2.4. Reconstruction 33

2.4.1. Particle Flow 34

2.4.2. Vertexing 39

2.4.3. Jets 40

2.4.4. Missing Transverse Energy 43

3. Machine Learning **45**

3.1. Neural Networks 46

3.2. Architectures 54

3.2.1. Convolutional neural networks 55

3.2.2. Recurrent Neural Networks 57

3.2.3. DeepSet 59

3.2.4. Other architectures 60

4. Deep Neural Network for Jet Identification **63**

4.1. Training Samples and labeling 65

4.1.1. Input features and preprocessing 66

4.1.2. Neural Network Architecture . . . 69
4.1.3. Training procedure . . . 70
4.2. Jet flavour identification performance . . . 71
4.3. Calibration of performance . . . 82
4.3.1. Identification of b-jets . . . 82
4.3.2. Negative Tagger for light jet identification . . . 85
4.3.3. Identification of c-jets . . . 88
4.4. Physics Analyses using DeepJet . . . 88
4.5. Conclusion . . . 89

5. Deep Neural Network for Jet Energy Regression **91**
5.1. Neural Network model . . . 92
5.2. Performance of the Neural Network . . . 93
5.3. Jet Energy Correction Flavour uncertainties . . . 97
5.4. Conclusion . . . 102

6. Reducing the Top Quark Mass Systematic Uncertainty with Machine Learning **103**
6.1. Top quark mass measurement . . . 104
6.1.1. The $pp \to t\bar{t} \to bq\bar{q}b\mu\nu$ event selection . . . 105
6.1.2. Top Quark Reconstruction . . . 107
6.1.3. Ideogram Method . . . 110
6.1.4. Systematic Uncertainties . . . 116
6.1.5. Experimental uncertainties . . . 116
6.1.6. Modeling of hadronization . . . 117
6.1.7. Modeling of pertubative QCD . . . 117
6.1.8. Modeling of soft QCD . . . 118
6.1.9. Total systematic uncertainties . . . 119
6.2. Reducing the systematic uncertainty: ReSYST . . . 121
6.2.1. The ReSYST Method . . . 121
6.2.2. Mono-variable approach of ReSYST . . . 124
6.2.3. Caveats . . . 131
6.2.4. Multi-variable approach of ReSYST: Neural Network . . . 132
6.3. Conclusion and Outlook . . . 138

7. Conclusions and Outlook **141**

A. Input Variables: DeepJet and DeepJet JEC **147**
A.1. List of global variables 147
A.2. List of charged candidate variables 148
A.3. List of neutral candidate variables 149
A.4. List of secondary vertex variables 149

B. Input Variables: ReSYST **151**
B.1. Event Variables 151
B.2. Jet Variables 151

C. DeepJEC Herwig Pythia **153**

D. Thesis cover **157**

Author contributions **161**

Bibliography **163**

"An expert is a person who has found out
by his own painful experience all the mistakes
that one can make in a very narrow field."

— Niels Bohr

Chapter 1.

Introduction

When the LHC experiments at CERN discovered the Higgs boson in 2012 after 2 years of data collection [1, 2], it was a great achievement of the Standard Model of particle physics [3–6], as the Higgs boson was the missing piece of the original theory formulated in the 1970s. Several signs however indicate that the Standard Model is incomplete as it fails to explain phenomena such as the matter-antimatter asymmetry [7], the cosmological observations of dark matter and energy [8], as well as neutrino masses [9, 10]. However when probed at extreme energies at the LHC [11], the Standard Model has proven to be highly robust. During the 10 years since the discovery of the Higgs boson, the LHC has been collecting data, and every conducted measurement has been in agreement with the Standard Model prediction. If an observable discrepancy between theory and the collected data is present, it is clear that it must be a very small effect. The LHC is going to run for many years to come at a center of mass energy up to 14 TeV, collecting additional collision data, and the core goal will be to try to measure the Standard Model particle properties to the highest precision that can be achieved, as minute differences can illuminate physics beyond the Standard Model. However, hadron colliders, like the LHC, are historically perceived as discovery machines and not precision machines. Because of the low rate of synchotron radiation emitted by protons, they can be collided at very high center of mass energy, allowing physicists to push the boundaries of the energy frontier in order to discover new particles. However, since protons are composite particles interacting via the strong force, their collisions produce many particles, most of which are not of interest for the physics analysis. This means that performing precision measurements is very challenging. In fact, despite the LHC only having collected a small fraction of the projected data that will be obtained, many precision measurements already have a larger systematic uncertainty component compared to the statistical uncertainty

component. For the LHC program to continue to push the boundaries of particle physics in the next 20 years, it will be crucial to improve the methods, which are used to study the collected collision data. A physics analysis conducted with a particle physics detector like the CMS experiment [12], is entirely dependent on the collision event reconstruction, which consists of combining the measurements from millions of silicon pixels, silicon strips and calorimeters cells to estimate and reconstruct the thousands of final state particles that are produced in intervals of 25 ns. This is an extremely challenging procedure employing statistical methods that have been refined over the last 40 years of experience at particle physics experiments. However, in the last 5 years, deep learning has come to the forefront of the field. Due to the multivariate nature of modern particle physics reconstruction, deep neural networks have shown an incredible potential to outperform the traditional methods. Investigating, developing and adopting such methods for reconstruction can significantly increase the sensitivity for new physics at the LHC, and it is an essential ingredient for a successful physics program.

A particle which is crucial to study further at the LHC is the top quark [13–15]. When top quarks are produced they quickly decay into b quarks as well as additional particles. These decay products then hadronize and form showers of particles that are measured. To study the top quark properties it is essential to reconstruct the original top quark from all the decay products. That task encompasses identifying the type of quark that initiated a given particle jet as well as identifying the correct kinematics of the initiating quark. In this thesis both these tasks are being tackled using deep neural networks that perform jet identification and jet energy regression respectively. Both methods significantly outperform traditional methods. Given that the LHC will run for many more years, the dataset to perform physics analysis will become much larger, driving down statistical uncertainties. This enables us to impose more stringent selection criteria for physics analyses. Building on top of the ReSYST method [16], a neural network method is developed in the context of a top quark mass measurement, in order to find an optimal event selection for minimizing the systematic uncertainty on the measurement.

This thesis is organized in five main chapters. Chapter 2 functions to give a brief overview of top quark physics at the LHC, providing the needed context to understand the other chapters in thesis. It briefly introduces the Standard Model of particle physics, and it gives an overview of why and how the top quark is important. Then, the experimental apparatus needed to conduct top quark physics is described in terms of

the LHC and the CMS experiment. Finally, the reconstruction algorithms that are used to process the data collected by the CMS experiment are described. Chapter 3 briefly describes the basic methods of machine learning, and then provides an overview of how machine learning has been adopted in the particle physics toolbox. A description of the main machine learning methods used for the algorithms developed in this thesis is included. Chapter 4 describes a novel deep neural network that identifies what type of quark initiated a given jet. Chapter 5 describes a neural network developed with the aim of calculating an energy correction to the measured jets, which improves the jet energy resolution and response. Chapter 6 details, in the context of a top quark mass measurement study, a neural network method that tries to find an optimal event selection for reducing the systematic uncertainty on the measurement.

Chapter 2.

Top Quark Physics at the LHC

The top quark is a one of the fundamental particles of what is called the Standard Model of particle physics. Theorized in 1973 by Kobayashi and Maskawa [13], it was experimentally discovered in the Tevatron proton-antiproton collider in 1995 by the CDF and DØ experiments [14, 15]. The top quark plays an important role in the Standard Model warranting detailed study. However, due to its high mass and extremely short lifetime, large and powerful particle colliders and advanced particle detectors are needed to measure its properties. The state of the art for this is represented by the Large Hadron Collider and the CMS experiment at CERN.

This chapter functions to give a brief overview of the necessary elements for modern top quark physics studies, setting the stage for the following chapters. First the Standard Model will be introduced, followed by an overview of the role of the top quark in the Standard Model as well as a description of the phenomenology of the top quark. Then, the experimental apparatus for producing top quarks will be covered, the LHC. The CMS particle physics detector that is used for measuring the decay products of these top quarks will then be described. Finally, the techniques used to reconstruct the top quarks from the raw measurements of the CMS experiment will be described.

2.1. The Standard Model

The Standard Model of particle physics [3–5, 17–20] is a theory that classifies and describes the elementary particles of physics along with their interactions. The model was developed in the early 1970's and it has been highly successful, being able to

explain particle physics phenomena to a high precision. The Standard Model consists of 12 elementary particles of spin $\frac{1}{2}$ referred to as fermions. Additionally, there are 4 gauge bosons with spin 1 that mediate the three fundamental forces and thereby allow for interactions between the fermions. Finally, there is a single scalar boson with spin 0, the infamous Higgs boson that has its own unique role in binding together the theory. Each of these fundamental particles have their own unique properties, some are extremely similar except for having different masses, while some have wildly different interaction properties.

Fermions

The fermions can be divided into two groups called quarks and leptons. Additionally, they can be divided into three generations. Each generation consists of 4 particles with different properties as shown in Table 2.1.

Elementary Fermions - spin 1/2 [17]								
	Quarks Carry color charge				Leptons No color charge			
Generation	Name	Q [q_e]	m [GeV]	Year of discovery	Name	Q [q_e]	m [GeV]	Year of discovery
1st	up (u)	+2/3	$2.16 \cdot 10^{-3}$ $\pm 0.5 \cdot 10^{-3}$	1969	Electron (e^-)	-1	$5.1 \cdot 10^{-4}$ $\pm 1 \cdot 10^{-9}$	1897
	down (d)	-1/3	$4.67 \cdot 10^{-3}$ $\pm 0.5 \cdot 10^{-3}$	1969	Electron neutrino (ν_e)	0	$< 10^{-9}$	1956
2nd	charm (c)	+2/3	1.27 ± 0.02	1974	Muon (μ^-)	-1	$1.05 \cdot 10^{-1}$ $\pm 1 \cdot 10^{-6}$	1937
	strange (s)	-1/3	$9.3 \cdot 10^{-2}$ $\pm 1 \cdot 10^{-2}$	1969	Muon neutrino (ν_μ)	0	$< 10^{-9}$	1962
3rd	top (t)	+2/3	172.76 ± 0.3 [1]	1995	Tau (τ^-)	-1	1.78 $\pm 1 \cdot 10^{-4}$	1975
	bottom (b)	-1/3	4.18 ± 0.3	1977	Tau neutrino (ν_τ)	0	$< 10^{-9}$	2000

Table 2.1.: The properties of the Standard Model fermions. Q corresponds to the electromagnetic charge, and q_e indicates the electron charge.

[1] The definition of the top quark mass is ambiguous as will be discussed later. The value reported corresponds to the mass from direct measurements.

Each particle has a corresponding partner in the different generations that are nearly identical, but are only separated by the higher generation of particles having a larger mass as well as having different flavour quantum numbers. In fact, the third generation particles are in between 10^5 and 10^6 times heavier than the first generation. As they all have spin $\frac{1}{2}$, they obey Fermi-Dirac statistics [21].

The 1st generation of fermions consists of the particles that make up the world as we know it, as they are the only stable particles with mass in the Standard Model. The 1st generation of leptons consists of the electron with a charge of $-1 \cdot q_e = -1.602 \cdot 10^{-19}$ C and its neutral and nearly massless partner, the electron neutrino. The 1st generation of quarks consists of the up and down quark that make up protons and neutrons that together with the electrons form atoms. They have fractional electrical charge of $-\frac{1}{3}q_e$ and $+\frac{2}{3}q_e$ respectively. Additionally, they interact via the strong force as will be discussed further in the following section. The 2nd generation of leptons consists of the muon and the muon neutrino. As the muon is unstable, it almost always decays to the electron, an electron antineutrino and a muon neutrino with a life time of around $2.2 \cdot 10^{-6}$ seconds. The second generation of quarks consists of the strange and charm quark that have significantly larger masses than their 1st generation counterparts. Finally, the third generation of particles is considered the most exotic, as their large masses make them exceedingly rare to find outside of dedicated particle physics experiments. The leptons consist of the tau lepton and its neutrino counterpart. With a mass of 1.7 GeV the tau is able to decay into quarks unlike the muon and electron. The third generations of quarks consists of the top quark and the bottom quark, the two main objects of study in this thesis. The top quark is the heaviest know fundamental particle with a staggering mass of $\sim$173 GeV. This means that despite the top quark being pointlike, the mass is comparable to that of a gold atom. This enormous mass also causes the top quark to be very short lived with a life time $5 \cdot 10^{-25}$ seconds. It almost always decays into its third generation quark partner, the bottom quark. While being 40 times lighter than the top quark, it is still significantly heavier than the other fermions. Additionally, each of the fermions have a corresponding antiparticle, which has the same mass, but opposite charges with respect to the fundamental forces. For instance the electron has an antiparticle named the positron with $Q = +1q_e$.

Gauge bosons and the particle interactions

While the fermions are the constituents that make up the universe, the gauge bosons are the force mediators that allow them to interact with each other. Altogether there

exists four kinds of gauge bosons, which are mediating three distinct forces. These are the electromagnetic force, the strong force and the weak force. Clearly, a fourth force of gravity also exists in nature, but it is not described by the Standard Model. All the gauge bosons are spin 1 and obey Bose-Einstein statistics.

The most well known Standard Model force is the electromagnetic force, which is responsible for the electric and magnetic fields. The fundamental theory that describes this force is called Quantum Electrodynamics (QED). The mediator of the electromagnetic force is the photon, which in itself is electrically neutral. It is also massless allowing the electromagnetic force to be long ranged compared to the other forces described by the Standard Model. The force couples to particles that carry electric charge.

The weak force is another of the Standard Model forces. The force mediators of the weak force is the W^+, W^- and Z boson. The W bosons have a mass 80.379 ± 0.012 GeV and the Z boson has a mass of 91.1876 ± 0.0021 GeV. These large masses make the weak force extremely short range, which is why the force has appropriately been named weak. Beyond that the W boson carries electrical charge and therefore interacts via the electromagnetic force as well. The weak force couples to particles with different flavours, meaning that it is a mechanism that allows particle decays. Consequently, all the 12 fermions interact via the weak force. The weak and electromagnetic force are deeply interconnected as will be described later on.

The final force is the strong force, which is described by the theory of Quantum Chromodynamics (QCD). Its mediator is the massless particle, the gluon. The gluon contains color charge itself and it exists in eight different varieties that can couple to the quarks, which carries one of three distinct color charges. Correspondingly, the anti-quarks carries anti-color. The strong force has the strongest coupling strength by far. Despite the mediator being massless, it is in general limited by range. Because of the extremely strong coupling and because the mediator itself interacts via the force, a phenomenon called color confinement occurs that make the strong force effectively short ranged. Due to this nature of strong force interaction, particles with color charge are never observed, but are always found inside bound states referred to as hadrons. Hadrons mostly come in two varieties of baryons and mesons. The baryons consists of three quarks that each carry a separate color charge to achieve the net charge of zero. The proton (uud) and neutron (udd) are examples of such baryons. The mesons consist of a quark-antiquark pair that carry color and the corresponding anti-color to cancel each other achieving a neutral color state. The QCD theory does allow for more

exotic bound states such as tetraquarks (two pairs of quark-antiquark), pentaquarks (4 quarks and 1 antiquark) as well as glueballs (pure gluon state).

Gauge theory [18–20]

The mathematical formulation of the Standard Model relies on relativistic quantum field theory. The aim of this mathematical description is to model the dynamics and interactions of the particles. The described particles are each connected to a quantized field $\psi(x)$, where x refers to the coordinates in spacetime. Particles as observed at the particle physics colliders are described as excitations of these fields. Their dynamics and interactions can be found by evaluating the action S and imposing the principle of least action.

$$S = \int d^4x \mathcal{L}(\psi(x), \partial_\mu \psi(x)) \tag{2.1}$$

Here the $\mathcal{L}$ indicates the Lagrangian of the Standard Model, which therefore needs to be formulated. A non-interacting theory of fermions represented via a spinor field $\psi(x)$ is described by the Lagrangian in Equation 2.2.

$$\mathcal{L} = \overline{\psi}(x)(i\gamma^\mu \partial_\mu - m)\psi(x) \text{ with } \overline{\psi}(x) = \psi(x)^\dagger \gamma^0 \tag{2.2}$$

Here the indices μ run from 0 to 3, γ corresponds to the Dirac matrices and m is the mass of the fermion. Using the principle of least actions a relativistic wave equation can be found, and it is referred to as the aforementioned Dirac equation 2.3.

$$(i\gamma^\mu \partial_\mu - m)\psi(x) = 0 \tag{2.3}$$

Four plane wave solutions exist to solve this equation being combinations of spin up and down as well as positive and negative energy solutions. The negative energy solutions are reconciled by describing them as negative particles moving backwards in time. Since the time dependence of the wavefunction takes the form of e^{-iEt}, a positive energy can be recovered by performing the transformation $t \to -t$. These negative solutions yield the antiparticles related to the fermions. By examining the Equation of 2.2, it is clear that a global (i.e. no dependence on spacetime) U(1) symmetry is present, as the Lagrangian is invariant if the field ψ is rotated by a arbitrary phase θ that is constant through spacetime, i.e. $\psi \to e^{-i\theta}\psi$. Noether's theorem for global symmetries [22] states that if the Lagrangian of a physical system is invariant under a

global symmetry, then there must be a current and conserved charge associated with it. In the case of this U(1) symmetry it gives you the conservation of the electromagnetic current and charge. The interaction terms for the electromagnetic theory follow from imposing that the identified global symmetry, should also be a local symmetry. It is clear that due to the derivatives in Equation 2.2 if θ gets a spacetime dependence, the transformation $\psi \rightarrow e^{-i\theta}\psi$ changes the expression.

$$\mathcal{L} = \overline{\psi}(x)(i\gamma^\mu\partial_\mu - m)\psi(x) \rightarrow \tag{2.4}$$
$$\mathcal{L} = e^{i\theta(x)}\overline{\psi}(x)(i\gamma^\mu\partial_\mu - m)e^{-i\theta(x)}\psi(x) \tag{2.5}$$
$$= -\overline{\psi}(x)m\psi(x) + \overline{\psi}(x)i\gamma^\mu\partial_\mu\psi(x) + e^{i\theta(x)}\overline{\psi}(x)(-(i)^2\gamma^\mu\partial_\mu(\theta(x)))e^{-i\theta(x)}\psi(x) \tag{2.6}$$
$$= -\overline{\psi}(x)m\psi(x) + \overline{\psi}(x)i\gamma^\mu\partial_\mu\psi(x) + \overline{\psi}(x)(\gamma^\mu\partial_\mu(\theta(x)))\psi(x) \tag{2.7}$$
$$= \overline{\psi}(x)(i\gamma^\mu\partial_\mu - m + \gamma^\mu\partial_\mu(\theta(x)))\psi(x) \tag{2.8}$$

As can be seen from Equation 2.4, the local transformation of U(1) adds an additional term. The gauge principle states that a gauge field should be introduced that cancel this contribution. That can then be done by adding a field to the Lagrangian A_μ that under the local U(1) transformation becomes $A_\mu(x) \rightarrow A_\mu(x) - \partial_\mu(\theta(x))$. This field corresponds to the gauge boson and mediator of QED, the photon. The Lagrangian can then be written in the local U(1) invariant form as shown in Equation 2.9.

$$\mathcal{L} = \overline{\psi}(x)(i\gamma^\mu\partial_\mu - m + \gamma^\mu A_\mu(x)))\psi(x) \tag{2.9}$$
$$= \overline{\psi}(x)(i\gamma^\mu D_\mu - m))\psi(x) \tag{2.10}$$

Here the operator $D_\mu = \partial_\mu - iA_\mu(x)$ acting on ψ is also transforming as ψ itself under the symmetry. Finally, because a new gauge field A_μ was introduced, a kinetic term to describe the propagation of the associated particle is needed. Of course, this must be gauge invariant as well.

$$F_{\mu\nu} = \partial_\mu A_\nu + \partial_\nu A_\mu \tag{2.11}$$

Here $F_{\mu\nu}$ is referred to as the field strength tensor, and the complete Lagrangian is shown in Equation 2.12.

$$\mathcal{L} = \overline{\psi}(x)(i\gamma^\mu D_\mu - m))\psi(x) - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{2.12}$$

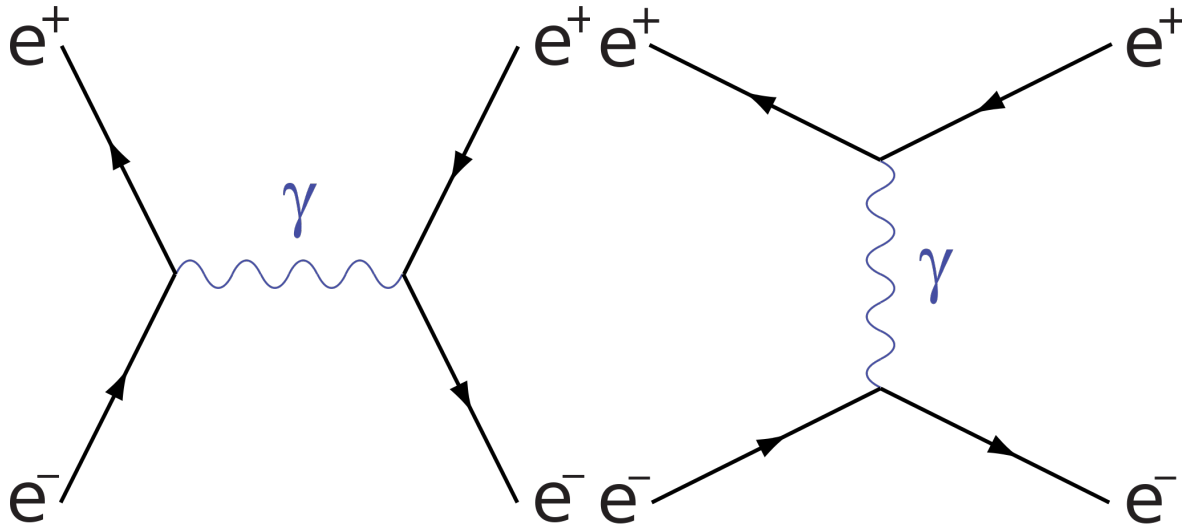
All the gauge bosons of the Standard Model follow from this gauge principle in an analogous way to this QED example. The full Standard Model Lagrangian's interaction terms follow from ensuring that the theory is locally invariant under the symmetry group shown in 2.13.

$$\mathrm{SU}(3)_C \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y \tag{2.13}$$

Here the $\mathrm{SU}(3)_C$ group is the symmetry group that yields the strong interactions, and it yields the 8 gluon gauge bosons. The $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ group yields the electroweak interactions, which in the "low" energy case breaks into both the electromagnetic and weak interaction generating the W^+, W^-, Z^0 and γ bosons.

Equipped with the formulation for the Lagrangian, the action can be evaluated. In quantum mechanics the probability of an event is given by the square of the probability amplitude, which can be found by taking the sum of all possible paths between the initial state and final state for the event, where the probability of a specific path is proportional to $e^{iS/\hbar}$. This then results in a probability amplitude that can be found by integrating over the probability of all possible paths. In quantum field theory these

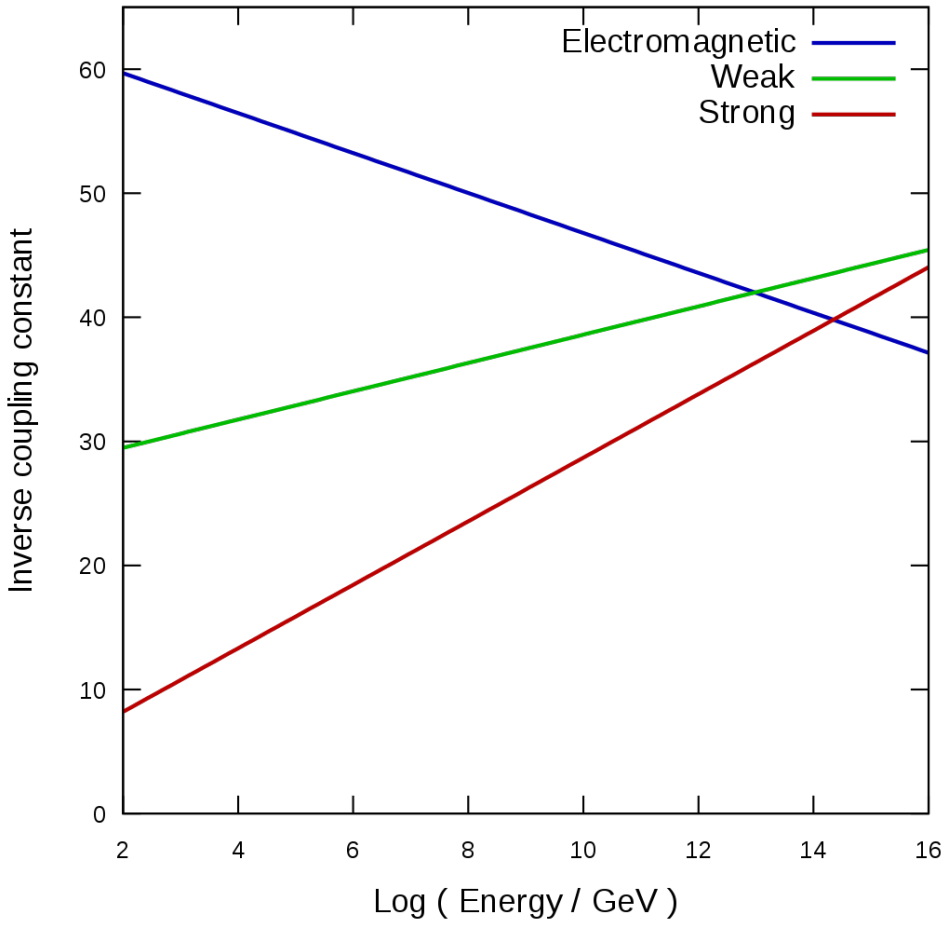
Figure 2.1.: The two Feynman diagrams of Bhabha scattering process $e^+e^- \to e^+e^-$. Due to each diagram having two QED vertices the amplitude is proportional to α^2.



probability amplitude calculations are based on approximate solution made using pertubation theory, by doing a series expansion around the coupling constant of the interaction terms. Terms of these infinite series for a given process can be represented

in a pictorial manner using what is called Feynman diagrams. Feynman diagrams of the leading order terms from the Bhabha scattering process is shown in Figure 2.1. For QED where the coupling constant is small $\alpha = 1/137$, this pertubative approach is very successful as higher order terms become small, and approximate QED solutions can be made and tested to extreme precision. However for QCD the situation is different.

Figure 2.2.: The coupling constant strength of the three SM forces are shown at different energy scales. From Ref. [23].



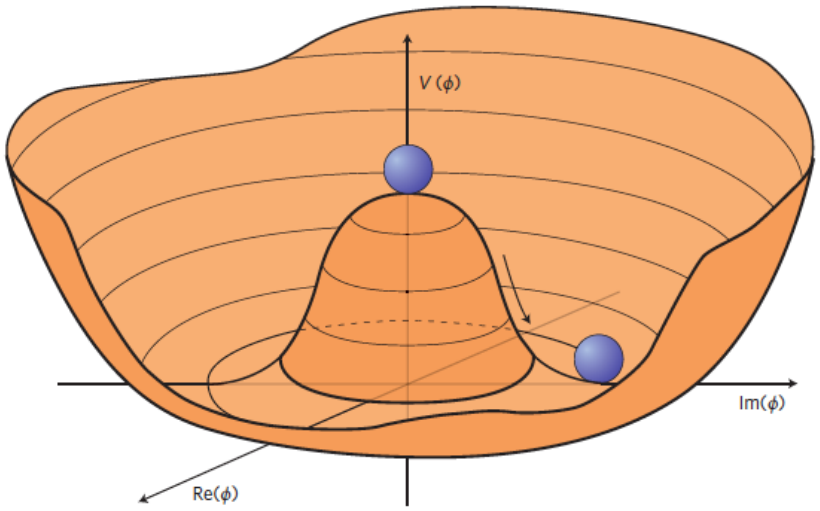
The strong force coupling constant α_S is close to 1 at low energy scale. However a phenomenon known as asymptotic freedom [24, 25] affects QCD, and it causes the interaction strength to decrease in an asymptotic manner as the energy scale of the interaction increases. Figure 2.2 shows an illustration of the behaviour of the couplings of the three forces at different energy scales. At large energy scales it turns out that the QCD coupling strength becomes weak enough that perturbation theory can be applied and approximate solutions can be found. However at larger distance scales,

such as the one to describe the process of color confinement, perturbation theory is not possible. Powerful computer simulations using the approach of lattice QCD have shown color confinement to emerge from the laws of QCD, and can in a very computational expensive manner calculate quantities such as the proton versus neutron mass difference [26]. Qualitatively color confinement can be understood as an effect where if two color charges are separated, a gluon "string" is created between them. Since the gluon has color charge itself, at some point the charges are far enough away from each other that forming a new quark-antiquark pair becomes favorable. Therefore, the free color charges are not observed as they always form neutral color states when isolated.

Brout–Englert–Higgs mechanism

The final piece of the Standard Model puzzle is the scalar Higgs boson. The particle is the crown achievement for illustrating the theoretical value of the Standard Model, as it was predicted already in 1964 by Brout, Englert and Higgs [6, 27–30], and then experimentally confirmed almost 50 years later at the LHC [1, 2] as a scalar boson with mass $\sim$125 GeV. The need for the Higgs field comes from identifying that the fermions and gauge bosons need to be massless in order to be gauge invariant under the symmetry shown in Equation 2.13. Of course measurements of the Standard Model particles showed that the particles did indeed have a non-zero mass indicating that some additional mechanism was needed to generate the mass term in a gauge invariant way. This can be resolved by adding a self-interacting complex scalar field ϕ with a potential term $V(\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2$ to the Standard Model Lagrangian. Figure 2.3 illustrates the Higgs potential. Unlike the other fields of the Standard Model, the vacuum expectation value of this Higgs field is non zero, and it takes value of $v \simeq 246$ GeV. Below the scale of the electroweak unification, the $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ symmetry of this Higgs field is spontaneously broken, which causes the W^+, W^- and Z^0 boson to get an effective mass. The fermions acquire a mass via a Yukawa interaction with the Higgs field. The Higgs bosons coupling strength to fermions is proportional to their mass, meaning that the strongest Higgs boson coupling to a fermion, is to the top quark. The coupling of the Higgs boson to the gauge bosons are proportional to the square of the gauge boson mass.

Figure 2.3.: The Higgs potential when $\mu^2 < 0$. From Ref. [31].



2.1.1. The Top Quark

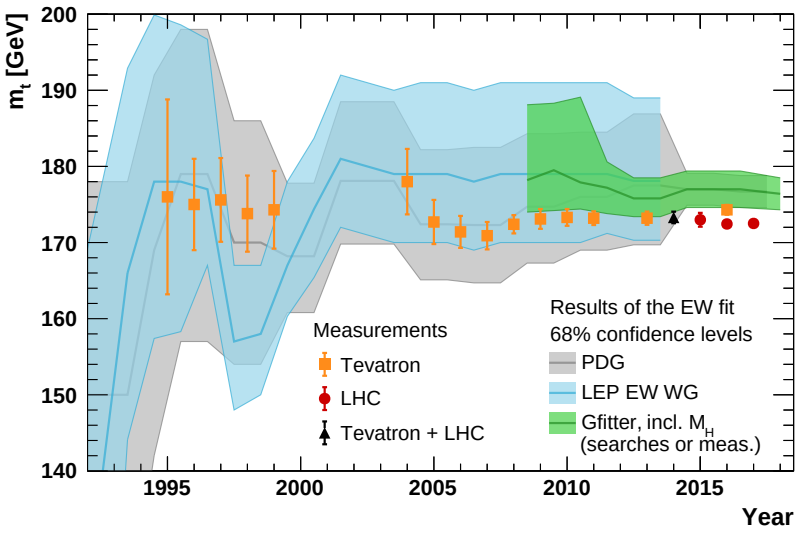
The top quark stands out in the Standard Model as both the fundamental particle with highest mass, and the fundamental particle with the smallest lifetime. The large mass creates a deep connection with the Higgs boson, and thereby as well with the Electroweak symmetry breaking. The decay of the top quark is almost exclusively to a bottom quark and a W boson with the other down type quark being extremely suppressed. Furthermore, because of the extremely short lifetime the top quark is the only quark not to hadronize, as the timescale of the hadronization process is longer than the $5 \cdot 10^{-25}$ seconds lifetime. These unique elements of the top quark make it an interesting subject for particle physics studies.

Electroweak fits and vacuum stability

The Standard Model consists of 19 free parameters [18] corresponding to the fermion masses and the Higgs boson mass, the mixing angles and CP violation phase of the CKM matrix [13] that dictates the weak decay rates of the three generation of quarks, the coupling strengths of the forces as well as the vacuum expectation value of the Higgs field. As these quantities are not given by the theory, their values are instead input based on experimental data. Of course while they are free parameters of the theory, they are all interconnected, and each of their values put constraints on the others. For instance, as the top quark appears as radiative corrections to the W, Z

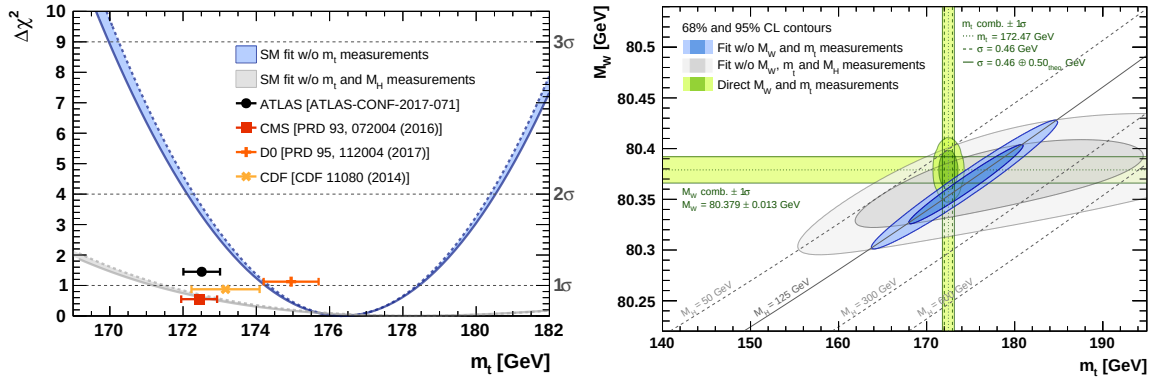
and Higgs boson propagators, its mass affect their masses. Historically electroweak fits [32] have been highly successful for constraining particle masses. In fact before the experimental discovery of the top quark, electroweak fits were used to predict the value of its mass. Figure 2.4 shows the electroweak fits of the top quark mass over time, which are in good agreement with the later measurements of top quark mass. Similarly, before the experimental discovery of the Higgs boson, it was determined that the Higgs boson was light from electroweak fits. The Standard Model can be fitted to see how well it describes the experimental data, and from this it can be inferred if new physics is needed to explain the measured parameters. Figure 2.5 shows the current best fitted values of the top quark mass in comparison with the experimentally measured values at the Tevatron and LHC. Similarly is shown a 2D fit of both the W boson mass and top quark masses compared with the experimental values. As of now, there is no indication that the Standard Model electroweak sector fails to describe the measured values, but increasing the measurement precision of the quantities might give further insight.

Figure 2.4.: The historical data of the predicted top quark mass through electroweak fits are shown by lines and shaded bands. The points are experimental measurements of the top quark mass from different years. From Ref. [32].



Because of the large radiative corrections of the top quark induced on the Higgs boson self coupling, the top quark mass plays an effect on the stability of the electroweak vacuum [33–35]. In fact the value of Higgs quartic coupling, λ, evolves at different energy scales. Depending on the specific values of particularly the strong cou-

Figure 2.5.: The predicted top quark mass through electroweak fits are shown on the left plot. A 2D fit of the m_W and m_t is shown on the right. From Ref. [32].

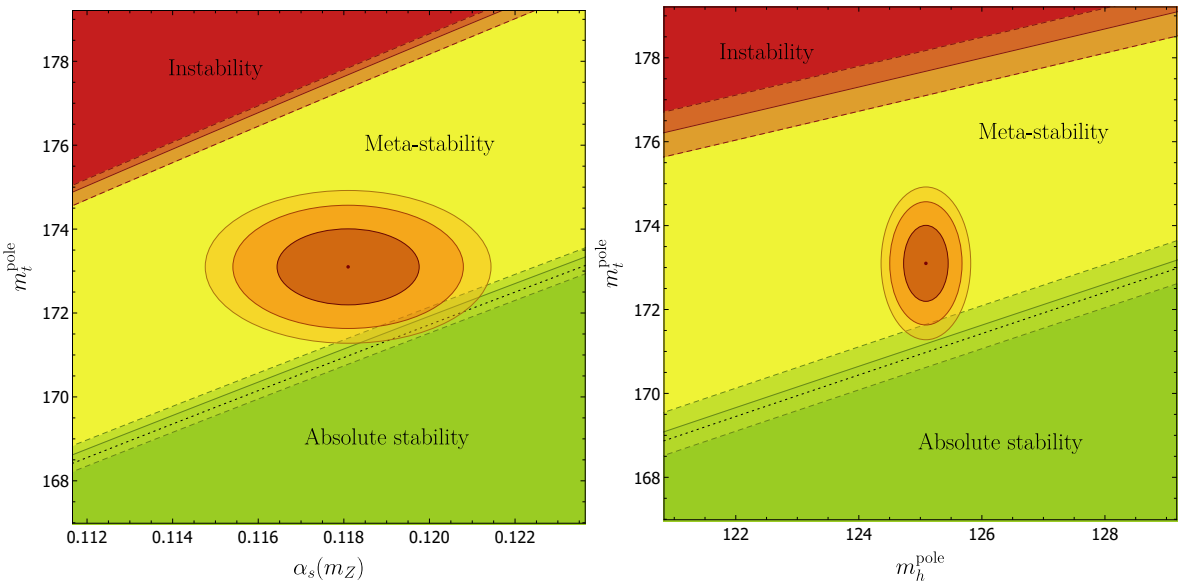


pling α_s and the top quark mass, λ can take negative values at very high energy scales. If this is the case, quantum tunneling effects could make the electroweak vacuum unstable. Figure 2.6 shows calculations based on the Higgs mass, α_s and the top quark mass of the vacuum stability. Absolute stability means that the electroweak vacuum will never decay, metastability means that it will decay, but the lifetime is larger than the lifetime of the universe, and instability implies that the universe should already have decayed. Interestingly the current measurement values happen to lie right on the boundary between metastability and absolute stability. According to Andreassen et al. [35], if the top quark mass was known to a precision below 250 MeV, the question of whether we are in the realm of metastability or absolute stability could be resolved.

Top Quark Mass definition in QFT and simulation [36]

The top quark mass as measured with direct measurements at hadron colliders is assigned an additional 0.5 GeV theoretical uncertainty when used in the electroweak fits [32]. Direct measurements of the top quark rely on using reconstructed kinematic observable of the top quark decay products, and then comparing these with the distributions obtained using Monte Carlo event generators using a given parameter for the top quark mass, m_t^{MC}. However, there is significant theoretical ambiguity in relating the measured top quark mass based on the m_t^{MC} with the top quark mass as it enters in quantum field theory. In quantum field theory the mass of a particle can only be defined within a given renormalization scheme [20]. These renormalization schemes allow for a way to absorb infinite terms from radiative corrections into the particle propagators. Several renormalization schemes can be chosen, as they have

Figure 2.6.: The stability of the universe is shown as a function of the top quark mass in units of GeV and the strong coupling α_s on the left. On the right it is illustrated as a function of the Higgs boson mass in units of GeV and the top quark mass in units of GeV. From Ref. [35].



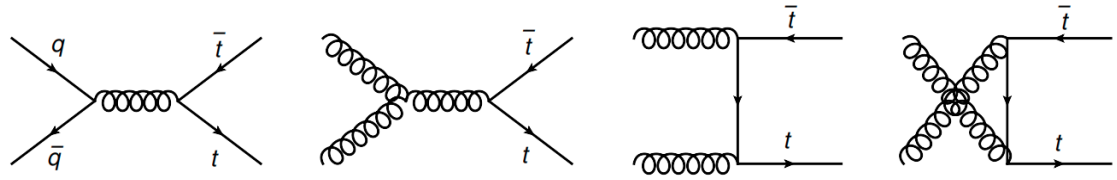
different utility depending on what observable quantities are to be calculated. In many calculations using pertubation theory, such as the electroweak fits and vacuum stability, the pole scheme is used for the top quark mass. This renormalization scheme absorbs all the radiative corrections from all energy scales. However, because the top quark is strongly interacting, this approach becomes nonphysical below a mass precision of ~ 0.5 GeV. This is because below a certain energy scale the strong force coupling becomes too large, and it can no longer be used in pertubation theory. A way to circumvent that problem is to use a renormalization scheme such as the MSR scheme [37], which only absorbs energy self corrections coming from scales larger than a fixed scale R. By setting R to a sufficiently large value, the low energy QCD corrections are not absorbed into the top quark mass. In the Monte Carlo event generators used for proton collisions several factorized steps are used for simulating events. Matrix elements are used for representing the hard scattering probabilities of the colliding particles, experimentally derived parton distributions functions are used for representing the internal structure of the protons, parton shower models are used for describing how the produced hard particles split into soft particles, and finally a hadronization model is used for dictating how the final particles will hadronize. The parton shower terminates at a energy value called the shower cut, which is often set to 1 GeV, in order to avoid the non-pertubative region of QCD. Therefore, there has

been much discussion as to whether the event generator mass is more analogous to the MSR scheme mass than the pole mass, as well as how close the MSR mass and the pole mass are to each other. Currently there is not a clear consensus as to how the direct top quark mass measurement should be interpreted. One attempt is made in reference [38], which compares e^+e^- collision event observables calculated analytically with simulated top quark events using the PYTHIA 8.205 MC event generator [39] to obtain a value of $m_t^{\text{MC}} - m_t^{\text{pole}} = 570 \pm 290$ MeV.

Top Quark Production

The dominant top quark production process at large energy hadron colliders is top quark pair production. At 13 TeV proton-proton collisions the production cross section is $\sigma_{tt} \simeq 780$ pb [40]. The leading order Feynman diagrams are shown in Figure 2.7. Roughly 10% of the production cross section comes from the quark-antiquark annihilation process. The remaining 90% are generated via the gluon fusion process. As

Figure 2.7.: The leading order Feynman diagrams for top quark pair production. The leading order quark annihilation process is shown as the leftmost diagram. The other three diagrams correspond to leading other gluon fusion processes.



the top quark has a very short lifetime it quickly decays to an on shell W boson and a down type quark. The down type quark corresponds to a bottom quark more than 95.5% of the time [17] and for most analysis it is estimated to be 100%. This means that the top quark pairs produced in a hadron collider will usually result in a process like:

$$t\bar{t} \rightarrow W^+W^-b\bar{b} \tag{2.14}$$

The produced on shell W bosons have a lifetime of $\sim 3.2 \cdot 10^{-25}$ s and they will therefore also quickly decay. The main decay modes of the W is $W^- \rightarrow q\bar{q}$ and $W^- \rightarrow l\bar{\nu}$, where l is one of the three leptons and $q\bar{q}$ refers to an up and down type quark-antiquark pair. Each possible decay occurs at the same rate of $\sim 1/9$. The quark pair decay occurs roughly $\sim 2/3$ of the time [17]. The decay rate of the W boson to a

quark flavour pair is proportional to $3|V_{ij}|$ where V_{ij} is the corresponding CKM matrix element. Therefore the hadronic decay should be dominated by the $u\bar{d}$ and $c\bar{s}$ pairs. This indicates three main decay channels exist for the top quark pair system.

$$t\bar{t} \rightarrow b\bar{b}W^+W^- \rightarrow b\bar{b}qq\bar{q}\bar{q} \tag{2.15}$$

$$t\bar{t} \rightarrow b\bar{b}W^+W^- \rightarrow b\bar{b}q\bar{q}l^-\bar{\nu}(l^+\nu) \tag{2.16}$$

$$t\bar{t} \rightarrow b\bar{b}W^+W^- \rightarrow b\bar{b}l^+\nu l^-\bar{\nu} \tag{2.17}$$

In the first decay mode both W boson decays to quarks. It is the most common decay mode and it occurs around 46% of the time [17]. Due to color confinement the quarks will hadronize into the collimated sprays of particles known as jets. This means that the experimental signature of such a process would be 6 jets, where two of them should have been initiated by the hadronization of a b-quark. This decay channel is therefore named the all-jets channel. In the second decay mode one of the W boson decays to a pair of quarks while the other decays to a lepton (antilepton) and antineutrino (neutrino). This corresponds to a rate of roughly 44%. This channel results in an experimental signature of four jets as well as a highly energetic lepton due to the large mass of the W boson. This channel is referred to as the semileptonic decay channel. Finally, both W bosons can decay to leptons which occurs around 10% of the time. Here two high energy leptons are produced.

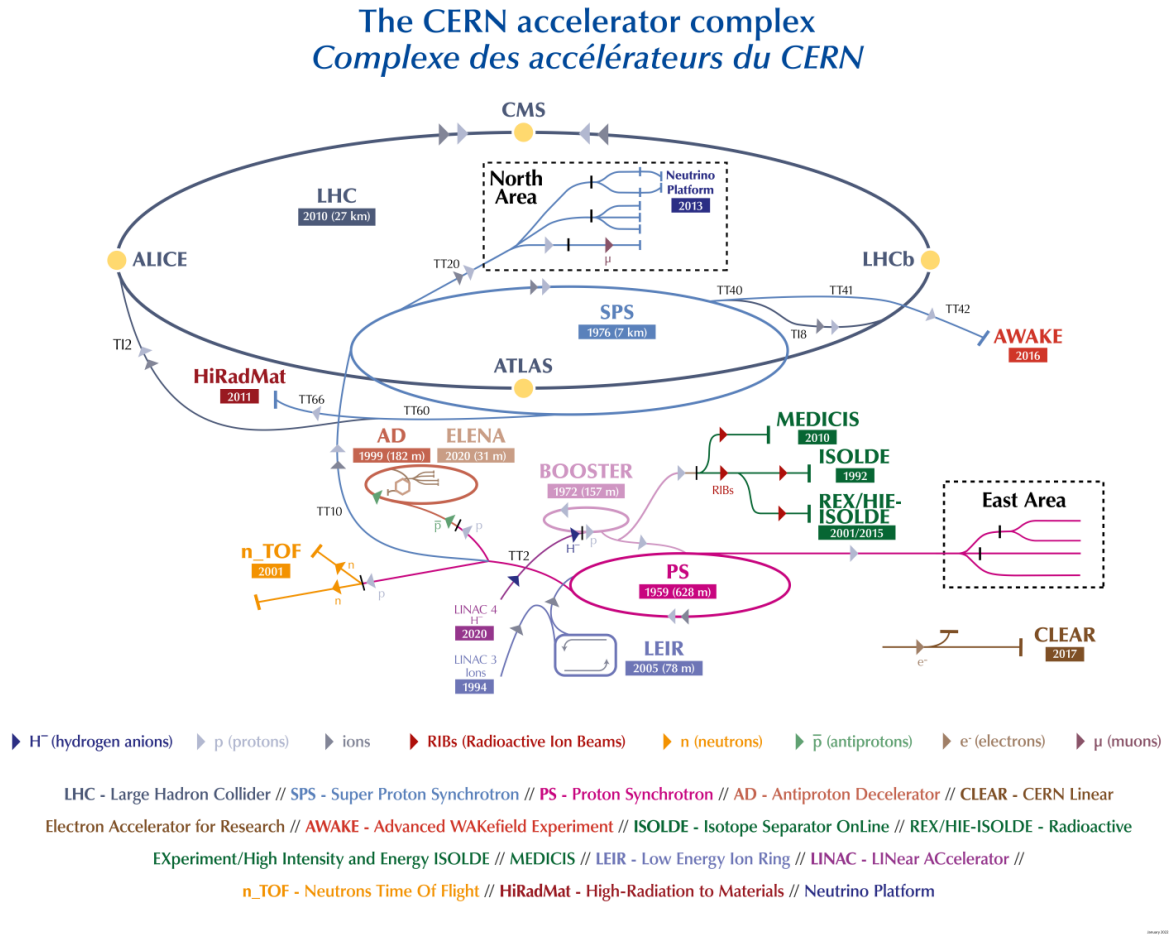
Therefore, a particle physics experiment that can probe the top quarks properties, first needs to be able to produce the top quark at a high rate. Since the mass of the top quark is so large, extremely high energy transfers are needed along with a high luminosity. To then measure the top quarks decay products, sophisticated detectors that can precisely resolve particle jets as well as high energetic leptons are needed. Additionally, it is crucial to identify the jets originating from b quarks to distinguish the quark coming from the top quark decay from the quarks produced in the W boson decay.

2.2. The LHC

The Large Hadron Collider (LHC) [11] is a circular proton-proton collider located at CERN in Geneva. Built in a massive underground tunnel with a circumference of 26.7 km, it is the largest particle accelerator ever constructed. It is capable of reaching the

highest particle collision energies in the world with a center of mass energy of 13 TeV with a design energy of 14 TeV. Since the top quark is heavy and interacting via the strong force, proton collisions at large energies produces top quarks in abundance, which makes the LHC the premier top quark factory today. It is part of the CERN accelerator complex pictured in Figure 2.8. The complex consists of several accelerators that bring the protons to a progressively increasing energy.

Figure 2.8.: The CERN accelerator complex is shown. The LHC ring is shown, as well as the different injector accelerators. From Ref. [41].



A bottle of hydrogen gas is used as the proton source and with electromagnetic fields the protons are stripped of their electrons. The initial accelerator is the LINAC4 [42], recently replacing the LINAC2 accelerator. The protons are accelerated to 160 MeV using radiofrequency (RF) cavities. These metal chambers produce an electromagnetic field that oscillates in a specific frequency. Protons are timed such that if they have energies below the target, the field will pull them as they approach the cavity and

subsequently push them as they exit. If a proton has a higher energy than designated it will arrive early to the cavity and instead be slowed down to the target energy. As protons are injected in bunches the oscillation also ensures that high energy protons in a given bunch get a smaller boost compared to the lower energy protons, making the momentum distribution in a given bunch more uniform. After the linear accelerator is a series of circular accelerators with progressively larger circumference. In circular colliders a larger circumferences enables a higher energy, as it limits the energy loss from synchrotron radiation, and it makes it feasible to curve the high energy protons enough with dipole magnets. The first is the Proton Synchrotron Booster (PSB) [43], which accelerates the protons to 2 GeV. This is followed by the Proton Synchrotron (PS), which accelerates the protons further to 26 GeV. The Super Proton Synchrotron (SPS) is then used to bring them to an energy of 450 GeV.

After this process the protons are ready to be injected into the LHC, which consists of two beam pipes with a diameter of $\sim$4 cm. Each pipe contains protons that are going around the ring clockwise and counter clockwise respectively. Throughout the system of accelerators, the beam pipes are kept in near perfect vacuum to avoid proton collisions with air molecules. Dipole magnets are used to curve the trajectories of the protons through the rings. As the LHC beam design energy is 7 TeV, these dipoles need to be very powerful to sufficiently curve the beam. The LHC dipoles are 15 meter long superconducting magnets that can generate a magnetic field of 8.3 T. Throughout the ring 1232 dipoles are used, each containing superconducting coils that need to be cooled to $\sim$2 K. Since the proton beam is electrically charged, it will disperse over time if left unchecked. Therefore quadrupole magnets are used to focus the beam throughout the accelerators. At four points the beam pipes merge to form collision points. Here the four LHC experiment (CMS, ATLAS, LHCb and ALICE) are located in order to measure the results of the collision. In order to achieve the best physics program at the LHC, it is needed to maximize the luminosity. For head on collisions with Gaussian distributed beams, it can be described as in Equation 2.18.

$$L = \frac{f N^2 n_b}{4\pi\sigma_x\sigma_y} \tag{2.18}$$

Here f is the frequency, N is the number of protons in a bunch, n_b is the number of bunches in the collider and the σ is the width of the beams in the transverse plane. At the LHC each of these factors are optimised to achieve a design instantaneous luminosity of $10^{34}\text{cm}^{-2}\text{s}^{-1}$. Each proton bunch consists of 10^{11} protons with each

ring containing $\sim$3000 bunches. Right before collision points, the beam is further "squeezed" using quadrupole magnets in order to reduce the transverse size of the beam.

2.3. The CMS experiment

The $pp \to t\bar{t}$ event has several different signatures, and specialized particle physics detectors are needed to identify them precisely. At the LHC there are two general purpose particle physics detectors that are ideal for measuring such events, the CMS and ATLAS experiment [12, 44]. The acronym CMS stands for Compact Muon Solenoid, which is meant to highlight the core difference in design with respect to the ATLAS detector. As can be seen from the Figure 2.9, which shows an illustration of the detector, it is cylindrical and contains several layers for measuring different particle properties. The whole detector is 15 meters high, 15 meters wide and 21 meters long, and it weighs 14000 tons. The CMS detector uses a powerful solenoid magnet to curve the trajectories of the charged particles produced at the LHC collision point as these trajectories can be used to measure the momentum of the particles. With 2169 turns and a current of 19.5 kA, a magnetic field of 3.8 T is produced over a length of 12.5 meters. The innermost part of the detector is the tracker, designed for measuring the curvature of the trajectory of charged particles. While both in the ATLAS and CMS detector the tracker is inside a solenoid magnet, the CMS detector is designed to be more compact in the sense that the electromagnetic and hadronic calorimeter are also enclosed by this magnet. This is done to reduce the amount of material the particles have to traverse before reaching the calorimeters in order to improve the measurement resolution. This compactness puts strict design constraints on the calorimeters.

The coordinate system used with the CMS detector is illustrated in Figure 2.10. The origin is defined at the nominal proton collision point at the center of the detector. The z-axis aligns with the direction of the beams, and it points towards the Jura mountains. The x-axis points towards the center of the ring, and the y-axis points vertically upwards. The azimuthal angle ϕ is defined in the transverse (x,y) plane, and it is measured from the x-axis. The polar angle θ is defined between the z-axis and the transverse plane, and it is measured from the z-axis. The pseudorapidity η is defined as $\eta = -\ln(\tan(\theta/2))$.

Figure 2.9.: A cutaway diagram of the CMS detector is shown. From Ref. [45].

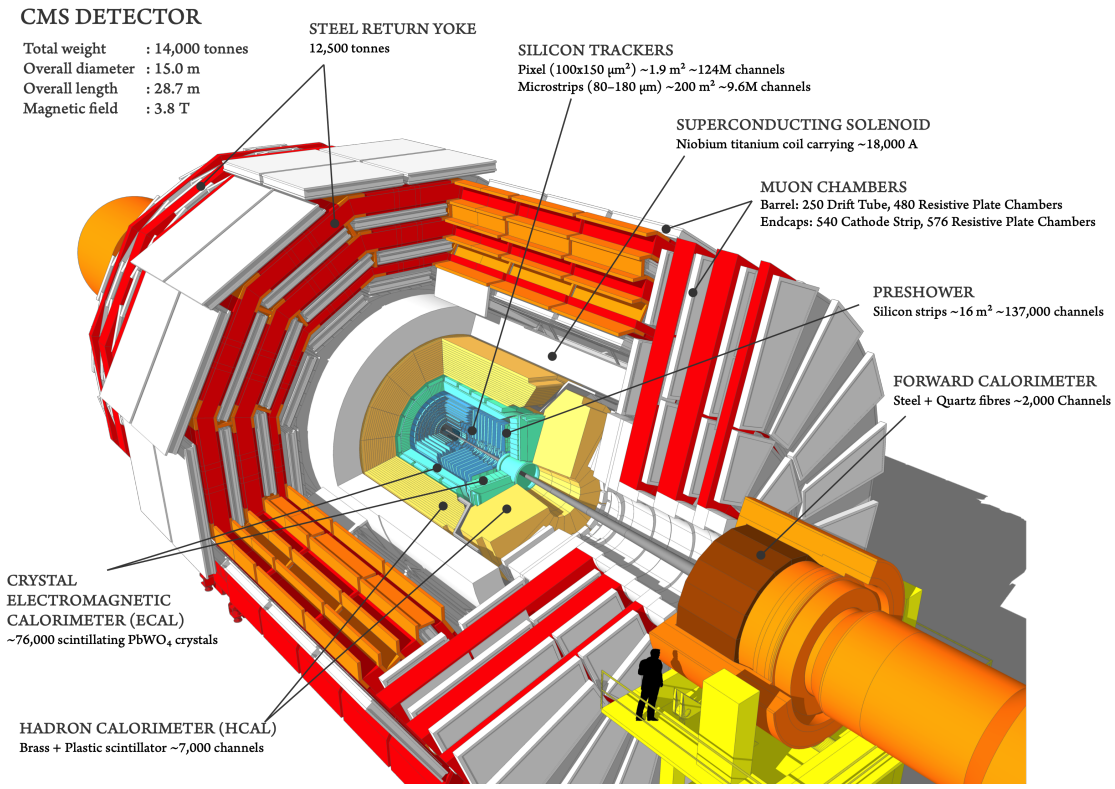
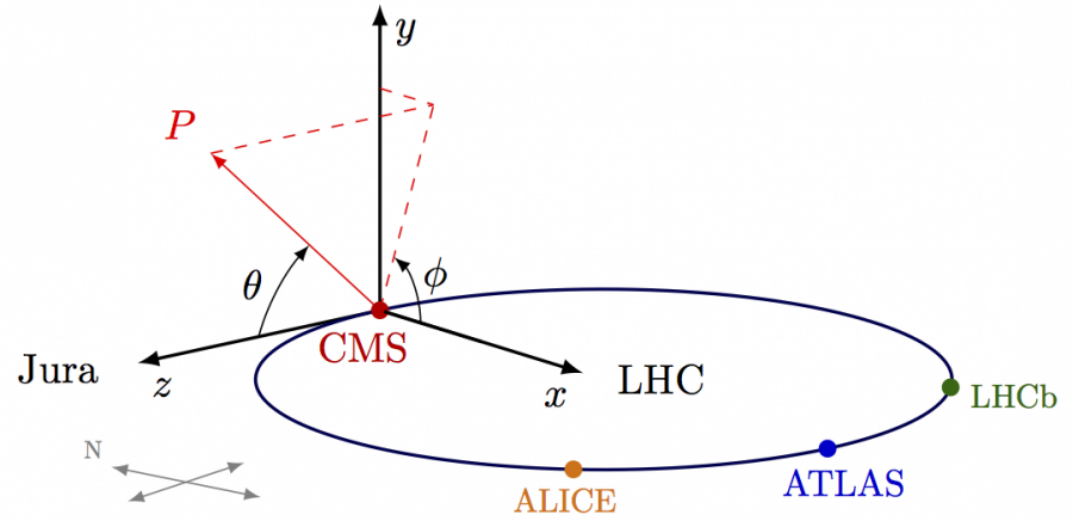


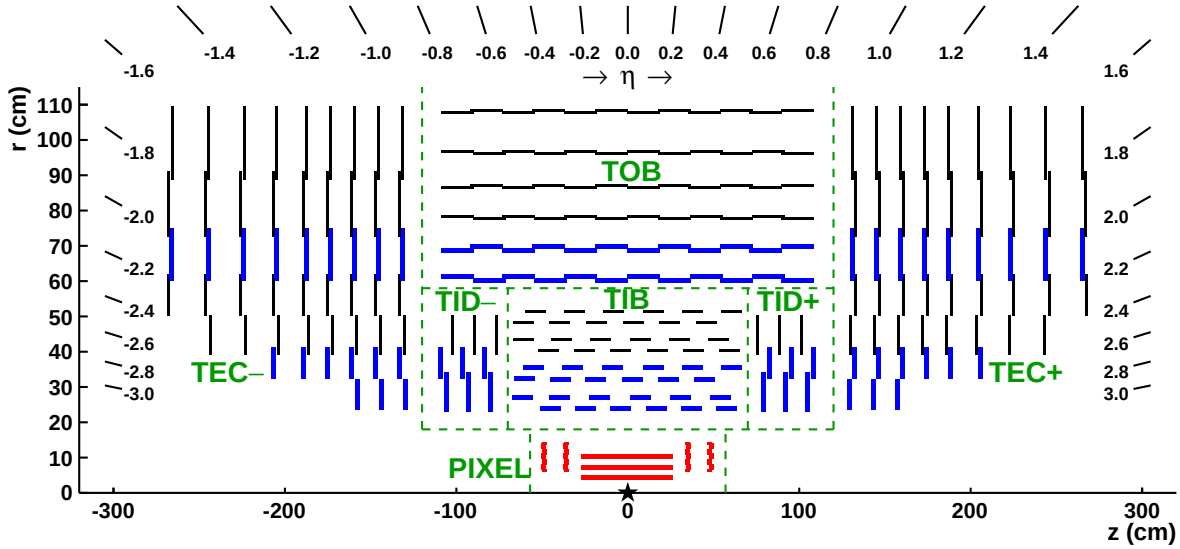
Figure 2.10.: The CMS coordinate system. From Ref. [46].



2.3.1. The Silicon Tracker

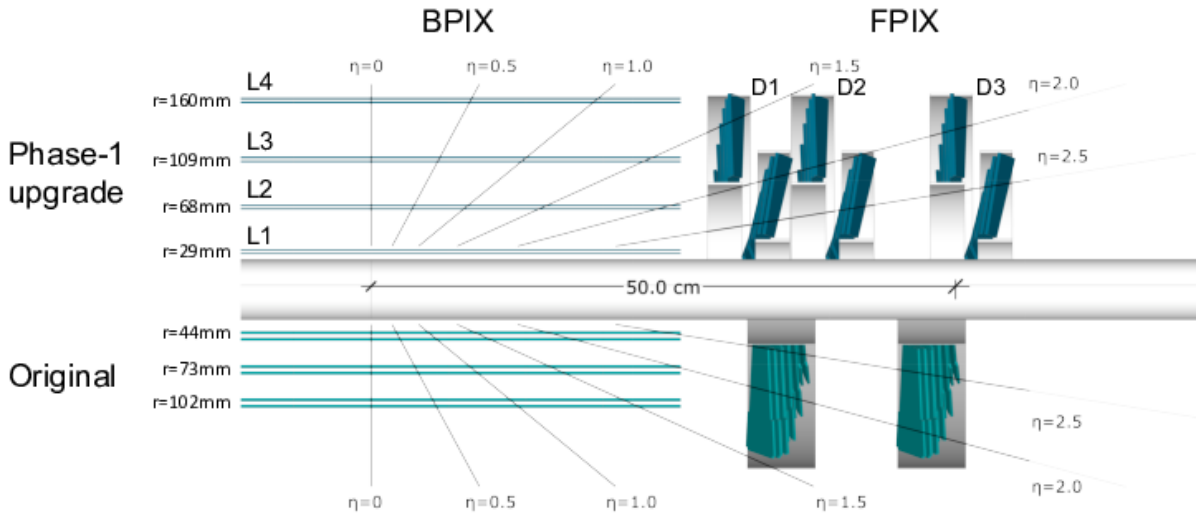
The CMS silicon tracker [47–50] is the detector closest to the beam pipe. Its function is to measure the trajectories of charged particles produced at the LHC. As a particle with charge q traveling with momentum p at right angles to a magnetic field with magnitude B moves in a circular path with radius r given by $p = Bqr$, a measurement of r therefore gives the momentum of the particle. The trajectories of the particles can also be used to perform vertexing to identify the primary collision point. Furthermore it can be used to reconstruct secondary vertices from particle decays which are crucial for heavy flavour identification. In order to measure the trajectories, $\sim$ 20000 silicon sensor modules are used to make up two endcaps and a barrel, which altogether provide a coverage up to $|\eta| < 2.5$. Figure 2.11 shows a cross section of the tracker in the (r,z) plane of the original CMS detector. In the years 2016 and 2017 the tracker was further upgraded by adding an additional pixel layer changing the design to what is illustrated in Figure 2.12.

Figure 2.11.: A schematic illustrating a cross section of the original CMS tracker in the (r,z) plane. As the tracker is symmetrical around r=0, only the top half is shown. From Ref. [51].



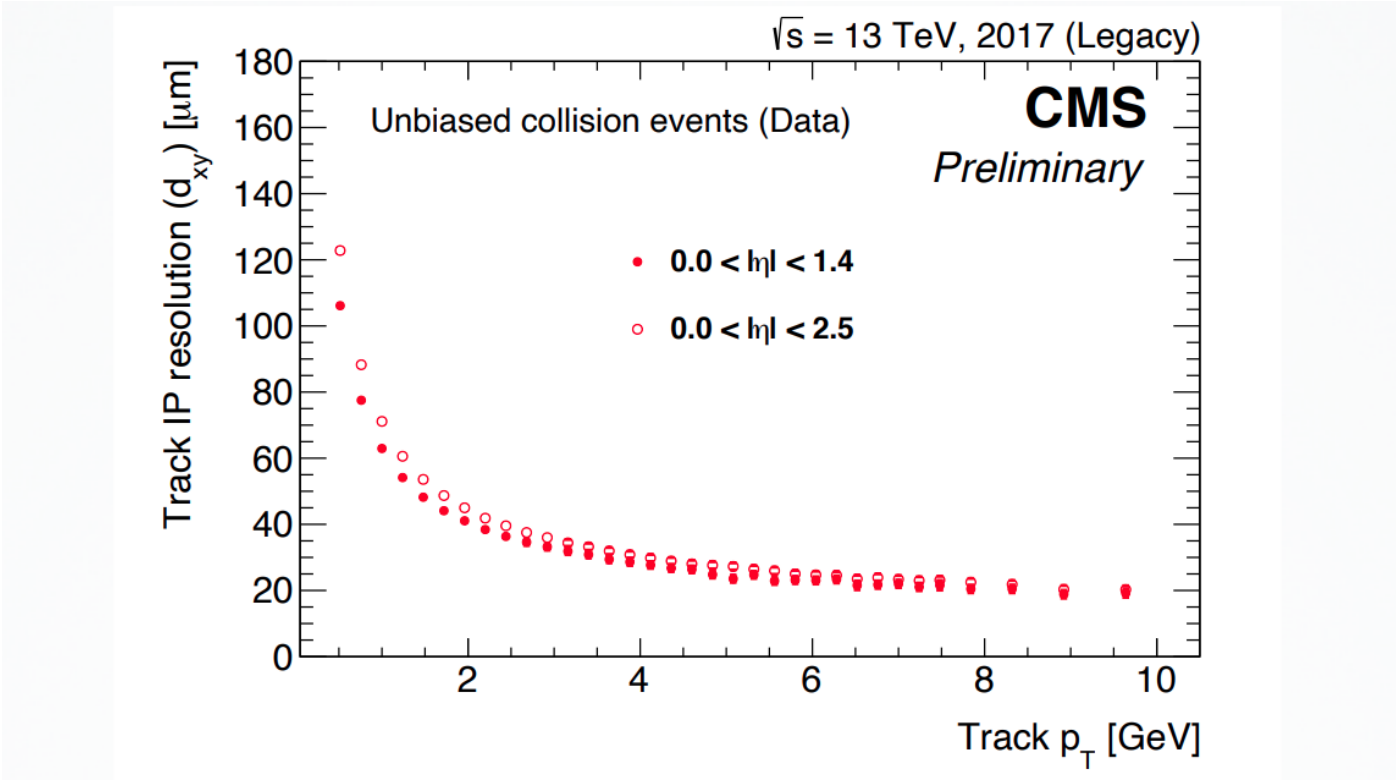
It has a detector radius of 1.2 m and a length of 5.6 m. Closest to the beam pipe at a distance of around 3 cm is the pixel detector. It is made up of high granularity silicon sensors with pixels of size 150x100 microns with a transverse (longitudinal) hit position resolution of 10 micron (20-40 micron). The barrel has four layers of pixel

Figure 2.12.: A schematic illustrating the difference between the original and upgraded CMS tracker in the (r,z) plane. As the tracker is symmetrical around r=0, only the top half is shown. From Ref. [50].



sensors and each endcap consists of three layers. Altogether the pixel detector has $\sim$120 million pixels and with a total active area of roughly 1.9 m^2. The outer tracker is made of silicon strip sensors with a lower granularity compared to the pixel modules, but the outer tracker allows a large coverage with an active area of roughly 198 m^2. The barrel consists of ten layers whereas there are 12 layers in each of the endcaps. A total of 9.8 million strips are used throughout. The strip pitch and width varies depending on the module location in the detector. The pitch can be between 80 and 205 μm, and a fixed width/pitch ratio of 0.25 is used. The CMS tracker is able to measure the transverse momentum, p_t, of charged particles at a resolution between $1-2\%$ depending on the particle p_t [52]. Primary vertices (PV) can be reconstructed at a resolution of around $10-12$ micron in the three coordinates. The transverse track impact parameter resolution strongly depend on the p_t of the track, and can be seen on Figure 2.13. The transverse track impact parameter is defined as the distance of closest approach of the track to the axis of beam.

Figure 2.13.: The transverse impact parameter resolution of tracks reconstructed with the CMS tracker is shown. From [52].

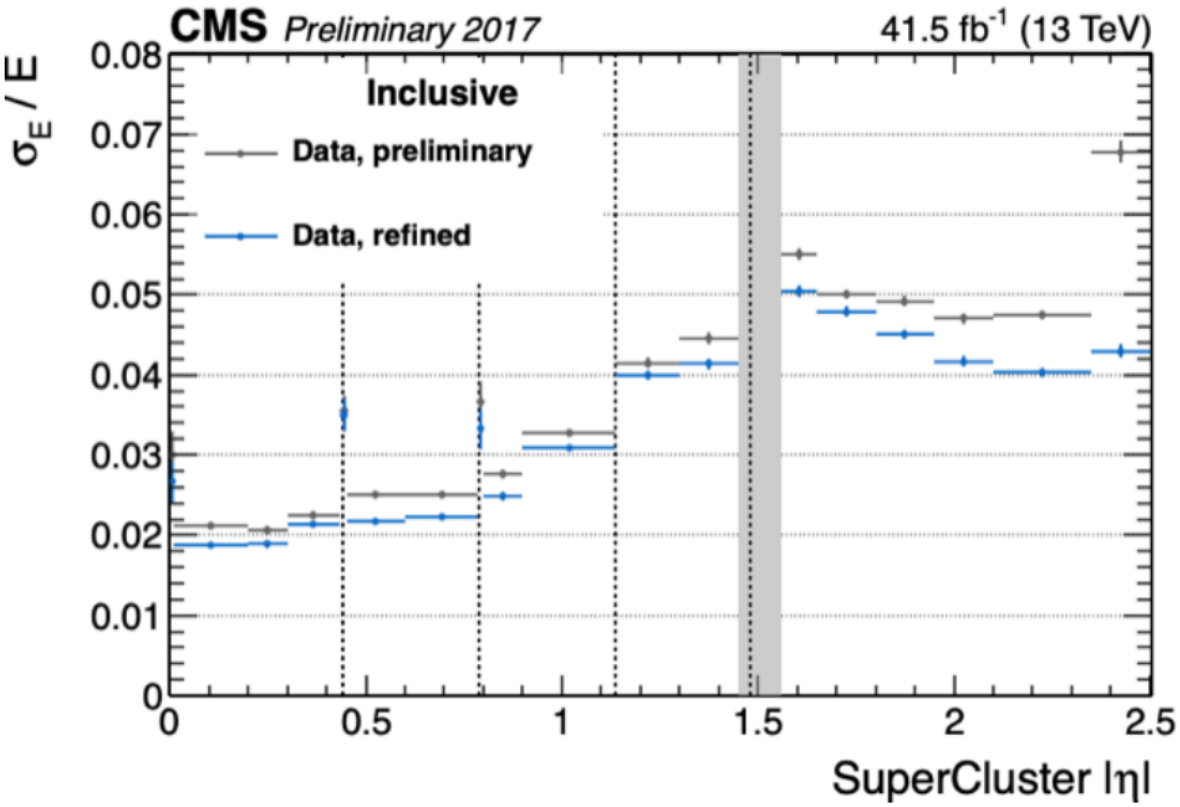


2.3.2. Electromagnetic Calorimeter

The CMS electromagnetic calorimeter (ECAL) [53, 54] is made to measure the energy of electrons and photons with high resolution. Since the ECAL of CMS has to fit inside the solenoid magnet, it is designed to be fairly small. The calorimeter is homogenous and the chosen medium is lead tungstate ($PbWO_4$) crystals. This material has a radiation length of ~ 0.89 cm, which means that the crystals only need to be 22-23 cm long in order to contain almost all of electromagnetic shower of high energy photons and electron. Like with the tracker, the ECAL consists of two endcaps covering $1.479 < |\eta| < 3.0$ and a barrel covering $|\eta| < 1.479$. As the lead tungstate has a Molière radius of 2.2 cm, the face of the crystals are chosen to be 2.2 x 2.2 cm^2 corresponding to $\Delta\eta \times \Delta\phi = 0.0174 \times 0.0174$. The endcap crystals are slightly larger with a face of 2.9 x 2.9 cm^2. Altogether there are 75848 crystal in the two endcaps and the barrel.

As can be seen from Figure 2.14 the energy resolution of electrons is $2 - 4\%$, and a better resolution is obtained in the barrel. Photons are similarly reconstructed with an energy resolution of roughly $2 - 4\%$.

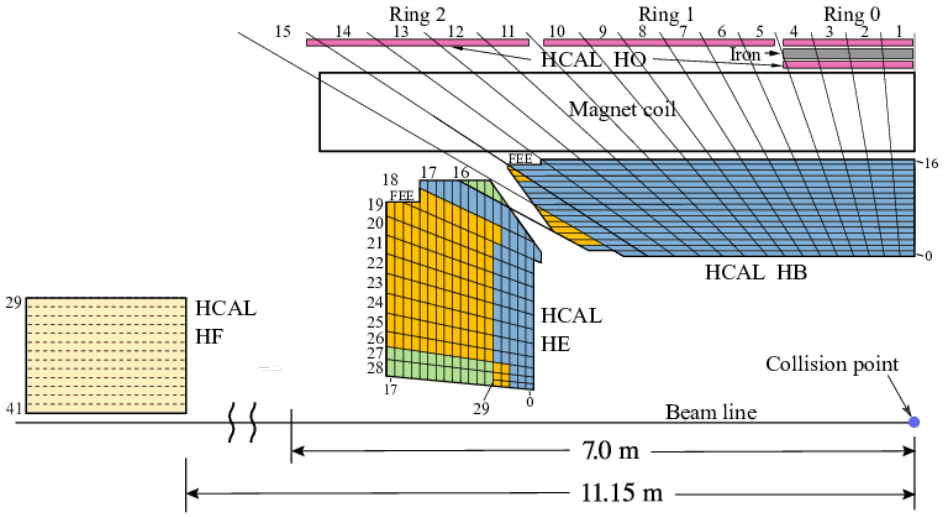
Figure 2.14.: The energy resolution of electrons as reconstructed with the CMS ECAL. The electrons were produced in $Z \to ee$. Refined indicates the obtained resolution after a data calibration procedure is performed. Superclusters refers to a collection of ECAL cells where the electron was assumed to deposit energy (see the later reconstruction section). From Ref. [55].



2.3.3. Hadronic Calorimeter

The CMS hadronic calorimeter (HCAL) [56–58] is a calorimeter designed to measure the energies of both charged and neutral hadrons produced in the LHC collisions. A schematic of the HCAL can be seen in Figure 2.15. Like the tracker and the ECAL it is composed of two endcaps (HE) and a barrel (HB). Additionally, it contains a forward calorimeter (HF) as well as a calorimeter outside the solenoid magnet (HO). It is a sampling calorimeter. The HB covers $|\eta| < 1.3$ and it uses brass as the passive medium and plastic scintillator tiles as the active medium. The brass layers are 5 cm thick and the plastic scintillator tiles are 3.7 mm thick. Altogether it consists of 17 layers, and it covers 5.8 nuclear interaction lengths for particles at $\eta = 0$. The HE uses slightly thicker brass absorbers of 8 cm and it covers $1.3 < |\eta| < 3.0$. It consists of 19 layers, which yields a depth covering roughly 10 nuclear interaction lengths. Both HE and HB are segmented into 2304 modules each covering $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$. The outer calorimeter HO consists of plastic scintillator layers and uses the CMS magnet coil as the absorber. Finally, the forward calorimeter HF, which is situated 11 m down the beam line from the collision point, covers $2.9 < |\eta| < 5.0$. It is a Cherenkov calorimeter with steel absorbers and quartz fibers for collection the Cherenkov radiation. The CMS hadron energy resolution [59] depends on the hadron pseudorapidity, but for single pions in the barrel it is of order $120\%/\sqrt{E}$.

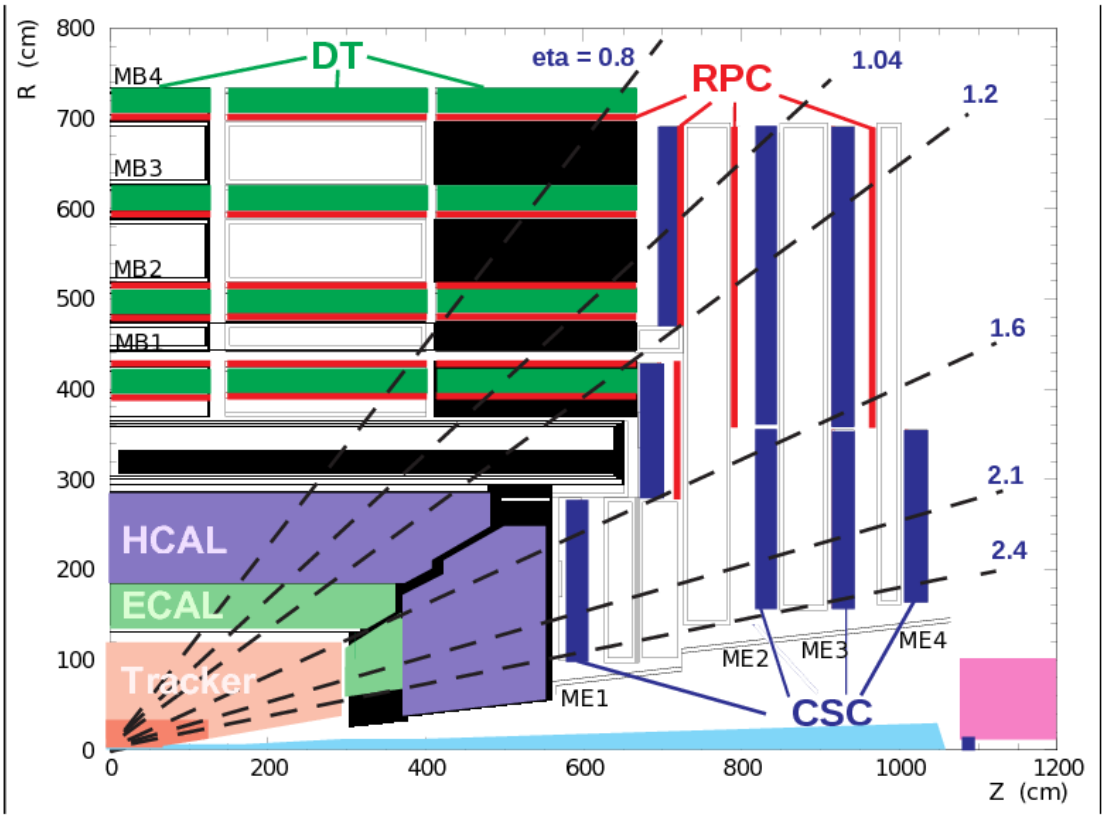
Figure 2.15.: A schematic illustration showing a cross section of a quarter of the CMS HCAL. From Ref. [58].



2.3.4. Muon detectors

Beyond the solenoid magnet and the outer calorimeter is the CMS muon system [60,61], which is used to identify and measure the momenta of muons. The magnet steel return yoke confines the magnetic field, and it is interleaved with the detector layers. Generally, only muons and neutrinos manages to reach this part of the CMS experiment. The detectors comprising the system are all gaseous detectors. A schematic illustrating the muon system layout can be seen on Figure 2.16. The muon system consists of a barrel that covers $|\eta| < 1.2$ and two endcaps which cover $0.9 < |\eta| < 2.4$. The barrel consists of four segmented layers of drift tube (DT) chambers. They can measure the position of muon hits with great spatial resolution, and they are supplemented further with resistive plate chambers (RPC) to obtain better timing resolution. Each endcap contains four disks of cathode strip chambers (CSC) which have a faster response time than the DTs but a more coarse spatial resolution. The CMS muon detectors have a typical muon reconstruction efficiency of $\sim 95\% - 99\%$.

Figure 2.16.: A schematic illustration showing a quarter of the CMS muon detector system. Longitudinal layout of one quadrant of the CMS detector. From Ref. [60].



2.3.5. Trigger system

The LHC produces bunch crossings every 25 nanoseconds yielding a rate of 40 MHz, and in every crossing roughly 20 proton collisions happen depending on beam conditions. The majority of the collisions are "glancing" inelastic collisions, where the momentum transfer is low, while the main interest of CMS is head on collision where the momentum exchange is high such that heavy particles like the top quark and the Higgs boson can be produced. Due to the enormous rate of proton collisions it is not feasible to store everything, which means that an extremely quick physics event selection needs to be applied to choose which events are of interest and therefore need to be stored. The CMS trigger system accomplishes this by employing a two level trigger system. The first part of this system is a level 1 (L1) trigger [62]. This system has to evaluate whether an event should be processed and kept in roughly 3 μs. Therefore there are significant constraints as to what event information can be used, as it needs to be reconstructed extremely quickly. These triggers rely on hit information from the muon system and energy measurements from the calorimeters, but it does not include tracking information. The trigger information is processed using dedicated hardware to meet the strict evaluation speed requirements, most commonly field-programmable gate arrays (FGPAs). The information is combined to determine whether an event should be kept, and the L1 trigger is able to reduce the selected event rate to 100 kHz. It is followed by a high level trigger (HLT) [63], which can perform a more refined selection as it has a couple of seconds to process the event. It uses simplified and faster versions of the offline reconstruction algorithms that will be covered in the following section. The HLT reduces the rate further to 1 kHz, which is manageable to process and store.

2.4. Reconstruction

The raw detector inputs consist of multiple silicon sensor and muon detector hits, as well as calorimeter clusters. In order to properly measure and identify a process like top quark pair production, it is essential to employ advance algorithms to identify which combination of hits and clusters are caused by which particles. This is further challenged by the many simultaneous proton collisions in an event and by electrical detector noise causing fake signals.

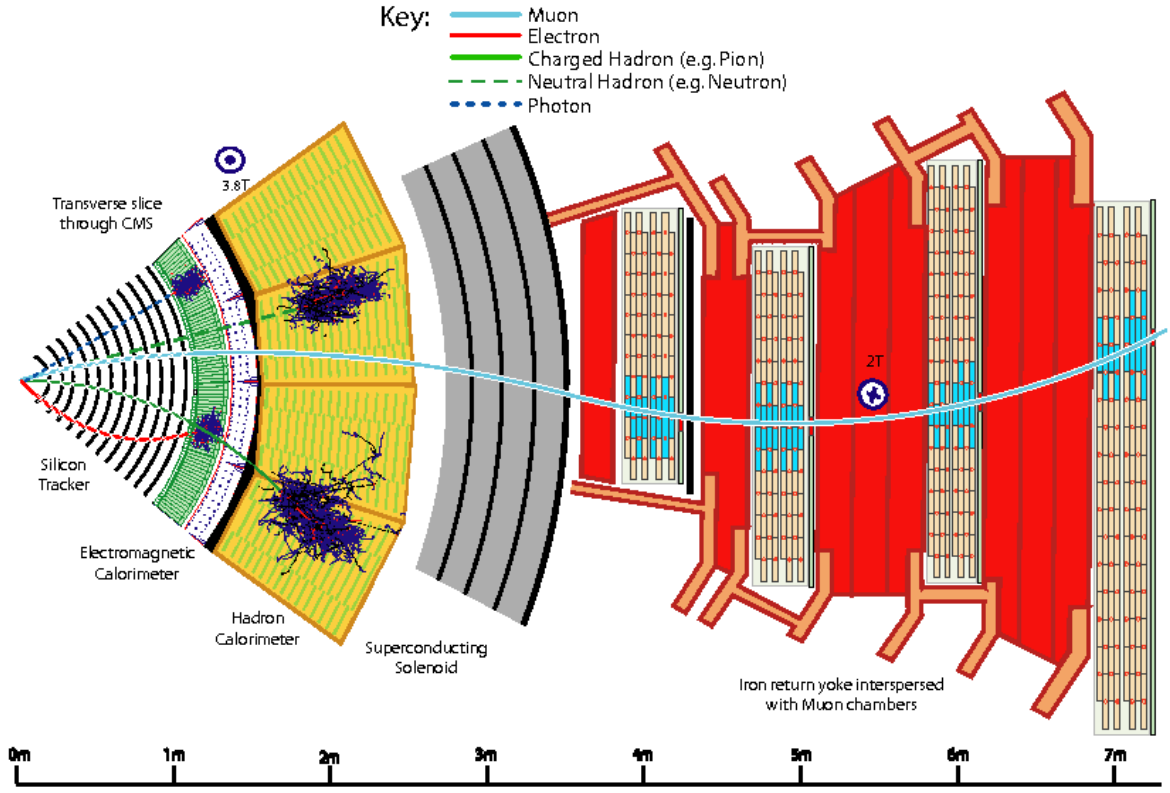
2.4.1. Particle Flow

In the CMS experiment the particle flow (PF) algorithm [64] is used to combine the information from all the separate detectors to create particle candidates. On Figure 2.17 the detector response of different particle types is illustrated. In top quark pair production the W boson can decay to quarks or leptons, resulting in the possible production of many types of particles. Neutral and charged hadrons are produced in the hadronization of the quarks from the top quark decay and the hadronic W boson decay. Neutral hadrons mostly leave a signal in the HCAL, but they can also start a hadronic shower already in the ECAL. The charged hadrons can leave a signal in both the HCAL, ECAL and tracker. Photons can also be produced in the process of the quark hadronization from the decay of produced hadrons or as radiation. They only leave a signal in the electromagnetic calorimeter. The W bosons leptonic decay can produce electrons, which leave a signal both in the ECAL and the tracker. Finally, a muon can be produced in the W bosons leptonic decay, and it can be measured in the tracker and the muon chamber. Since most particles are measured in multiple detectors, using the global event description allows the CMS experiment to achieve better reconstruction efficiency, compared to what can be done using individual detector systems separately. This is done using the PF algorithm. It relies on the reconstruction of PF elements using individual detector information, and then using a link algorithm to combine elements in order to form particle candidates.

PF Elements

The particle tracks left in the silicon tracker are a signature of most particles produced in proton collisions and are an essential element in the PF algorithm. The CMS tracking is done using a combinatorial track finder based on Kalman Filtering [52, 65] in an iterative manner. The algorithm aims to identify which of the many hits in the pixel and strip detectors are associated with single particle trajectories. An iteration of the algorithm functions in three stages. First, an initial track seed is generated to be compatible with a few hits caused by a single particle. The trajectory is further built by extrapolating into subsequent layers using the initial track seed, and then trying to identify hits that are compatible with this seed track. When a hit is found within the uncertainties of the trajectory extrapolation, it is added and the trajectory is updated. In the case that multiple hits are compatible with the same trajectory, multiple tracks are constructed, and the procedure continues for both. After this tracking building

Figure 2.17.: The different particle types are illustrated moving through the CMS detector. Each type have a unique detector response that can be used for identification and reconstruction. From Ref. [64].



step is done the ambiguities from multiple tracks that use the same hits or seeds are resolved, by keeping the tracks with the most total hits or best track fit χ^2. After this is done, the constructed tracks are fitted in order to extract the momentum from their curvature, as well as the direction and origin of the track. After one of these track iteration is done, the hits used to construct tracks are masked. A subsequent iteration is then done using less stringent track quality requirement with the remaining hits. The first iteration of the tracking specifically targets high energetic tracks identified by requiring good initial seeds with many pixel hits, as well as hits in nearly all detector layers. However, roughly $10-30\%$ of particles will undergo a significant nuclear interaction with the tracker material before being measured in all tracker layers. Therefore, subsequent iterations do not require the track to contain hits in all layers. The more loose requirements on the number of hits increase the fake track rate, so additional quality requirements are imposed on the track fit χ^2, the track seed and

primary vertex compatibility. Three iterations using triple pixel hits as seeds are done with progressively looser quality requirements. Following iterations are done using seeds formed of less pixel hits and eventually strip hits. Two final iterations are done using muon detectors hits as well as seeds.

Due to the low mass of the electron, it radiates significant amounts of bremsstrahlung in the CMS magnetic field, and it therefore requires a tracking procedure that can accommodate the energy loss. Additionally, it can also be measured in the ECAL and this information can be used to improve tracking. Seperate approaches are used for energetic well isolated electrons and softer electrons in jets. Due to the bremstrahlung, a larger fraction of the electrons energy is emitted as photons before reaching the ECAL. For energetic isolated electrons it is possible to collect most of this energy in a supercluster using a small η range with a wider ϕ range around the electron direction. The position and energy of these superclusters can then be used to extrapolate to the tracker and be matched with hits there. This is referred to as the ECAL based approach. For lower energy electrons, the tracks are curved by the magnetic field to such an extent that the bremsstrahlung is spread out in the ECAL, and cannot be fully gathered in the supercluster. This makes it difficult to extrapolate back to the tracker from the ECAL. The ECAL based approach is also inefficient for electrons inside a jet, as it is difficult to extrapolate from the ECAL to the tracker, since other particles in the jet also left hits in the tracker. Additionally the supercluster will contain energy deposits from the jet particles. Therefore a tracker based approach is instead used to identify these electron tracks. Here tracks are instead matched to clusters in the calorimeter. A track can be identified as an electron seed if the momentum of the track roughly corresponds to the energy deposit in the calorimeter. Since the electron loses energy through the tracker, a candidate electron track is refitted using a Gaussian-sum filter (GSF) [66] instead of a Kalman Filter. A final selection using a boosted-decision-tree is applied on the GSF tracks. The ECAL and tracker based electrons are combined and they are used as seeds for a higher component GSF tracking, obtaining the final electron tracks.

A different approach is used for muons as they also leave track segments in the muon chambers that can be helpful for tracking [67]. The muon tracks with the worst momentum resolution are the group referred to as standalone muons. They are composed of a fit consisting of only the segments measured in the muon chambers. These standalone muons can however be improved by matching them to tracks reconstructed

in the inner tracker, and then performing a new fit. These are referred to as global muons. These have the highest reconstruction efficiency for muons that have segments in several planes of the muon detector. However, for lower energy muons, segments in different planes might not be properly matched due to multiple scattering effects. Therefore muons are also reconstructed using another approach that relies on using tracks from the inner tracker and extrapolating them to the muon chambers to see if there is a match with a segment. These are referred to as tracker muons, and they often overlap with global muons, but they are better in low p_t cases. If a global muon and tracker muon share the same track reconstructed in the tracker, they are considered to be the same muon and are therefore merged.

Calorimeter clusters also functions as a PF element, and it is essential as it is the only method for measuring stable neutral particles. It also helps for measurements of the charged hadrons energy as well as for the electron reconstruction. In order to detect low energy particles in the calorimeters, and in order to properly separate energy deposits from different particles, an advanced clustering algorithm is needed. Since there are different detector geometries in the endcaps and barrel of both the HCAL and ECAL, the clustering algorithm is run separately in each of these. The algorithm functions by finding a seed consisting of a calorimeter cell with a large energy deposit exceeding a predefined threshold. To ensure it is the center of a cluster, it is required to have no cells with a larger energy deposit as a neighbour. From the cluster seed, a large cluster is then formed by adding cells that are adjacent, and have an energy larger than the twice the expected noise level. Using that the energy deposits roughly should have a Gaussian shape, larger clusters caused by multiple particles are separated into multiple clusters using an expectation-maximization algorithm based on a Gaussian-mixture model [64].

PF Reconstruction

Using the different particle flow elements constructed from the different subdectors, particle candidates can now be reconstructed by linking and combining these elements. A sophisticated algorithm is needed, since there are limitations in resolving individual particles due to detector granularity, and since trajectories can be affected by particle interactions with detector material. In order to establish links between the tracker and calorimeter, reconstructed tracks are extrapolated from its outermost hit in the tracker to the ECAL as well as to the HCAL. If a cluster is within the extrapolation

uncertainties in (η, ϕ) space, it is linked. The (η, ϕ) separation is recorded, and is used for arbitration if several tracks are linked to the same cluster or if several clusters are linked to the same track. For candidate electron tracks an additional procedure to link potential bremsstrahlung clusters is performed. The tangent position is extrapolated from the potential electron track and if a cluster is within the extrapolation uncertainties, it is linked to the track. Links are made between ECAL and HCAL clusters by checking if the cluster position in the ECAL in (η, ϕ) space is within the cluster position of the less granular HCAL. Again the same arbitration procedure is performed using the distance as a metric. Together linked PF elements form a PF block that is processed to reconstruct particles. Since muons are the simplest to identify and reconstruct, they are processed first, and the associated PF elements are removed from the block. Next, electrons and isolated photons are processed followed by charged hadrons, neutral hadrons and nonisolated photons.

Muons

The PF blocks that are considered as muon candidates, are the ones that contain muon track elements as described in the previous section. The simplest muons to reconstruct are isolated muons that are for instance produced via W boson decay. The track associated with a muon element in a PF block is checked to contain other tracks or calorimeter clusters within $\Delta R < 0.3$, where $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$. The total transverse energy of these nearby elements is required to be less than 10% of the muons transverse energy to be considered a muon. Of course muons produced in jets due to hadrons decaying leptonically cannot be identified via the isolation criteria. In the hadronic showers in the HCAL secondary muons can be produced and cause signals in the muon chambers that accidentally can be linked to a charged hadron track in the inner detector. To avoid these fake muons, quality cuts are applied on both the segments from the muon chamber and the inner detector tracks. For muons, the momentum is taken to be that of the associated inner track if the momentum is less than 200 GeV, since it provides the most accurate measurement. Above this threshold the curvature of the track is low, and the best track fit based on χ^2 is used.

Electrons and isolated photons

Electrons and isolated photons are processed in the same step, since electrons emit photons as bremsstrahlung and photons can convert to electron pairs making the

tasks intertwined. If a PF block contains a GSF track as described in the previous section it is considered as an electron candidate. Similarly, if the PF block consists of an ECAL cluster with no linked tracks it is considered as a photon candidate. Since both electrons and photons are expected to deposit almost all their energy in the ECAL, the HCAL is cross checked to confirm that there is no significant deposit near the ECAL cluster in (η, ϕ) space. For electron candidates the energy is calculated using both the linked ECAL clusters and the GSF track momentum. A boosted decision tree using different quality variables is used to evaluate these electron candidates as a final selection requirement.

Charged and neutral hadrons and nonisolated photons

The majority of the particles produced at the LHC comes in the form of hadrons that needs to be reconstructed and identified. Clusters and tracks not accounted for by electrons, muons and isolated photons are used. The ECAL clusters that are not linked to a track are considered to be a photon, and the photon energy is reconstructed as the deposited energy in the ECAL cluster. The HCAL clusters that are not linked to a track are considered neutral hadrons, and their energy is reconstructed based on the deposited energy in the HCAL cluster. The HCAL clusters that are linked to tracks can be used to construct charged hadron candidates. If an ECAL cluster is also linked to these elements, the sum of the HCAL and ECAL must be compatiable with the momentum of the track. If there is an excess this is identified as an additional photon or neutral hadron. If the measured energies are compatible in the sub-detectors, the hadron energy is estimated using a fit that is leveraging both the calorimeter and tracker information, as this obtains the most precise energy estimation. Any track that remains is identified as a charged hadron. No attempt is made to identify what type of charged hadron is produced, with the standard assumption being a charged pion. Outside the tracker acceptance of $|\eta| < 2.5$ no attempt is made to distinguish between neutral and charged hadrons.

2.4.2. Vertexing

In a given proton bunch crossing, several proton collisions can occur with an average number of around 20. As the chance of a collision with a larger momentum transfer is low, an event of interest usually only contain one high energy collision, with the

remaining low energy collisions being referred to as pileup. In order to reconstruct the proton collision points, vertexing using the tracks in the physics event is performed [52]. This task includes associating the tracks that correspond to the same vertex, but also fitting the position of the vertex. Several primary interaction vertices can be reconstructed in a single event. The vertex with the largest sum of particle p_t originating from it, is considered the main vertex, with the additional ones being referred to as pileup vertices. Potential vertices are clustered using a deterministic annealing algorithm based on the z coordinates of the point of closest approach of the tracks to the beam spot. This clustering constructs candidate vertices that are then fitted with an algorithm called the Adaptive Vertex Fitter (AVF) [68].

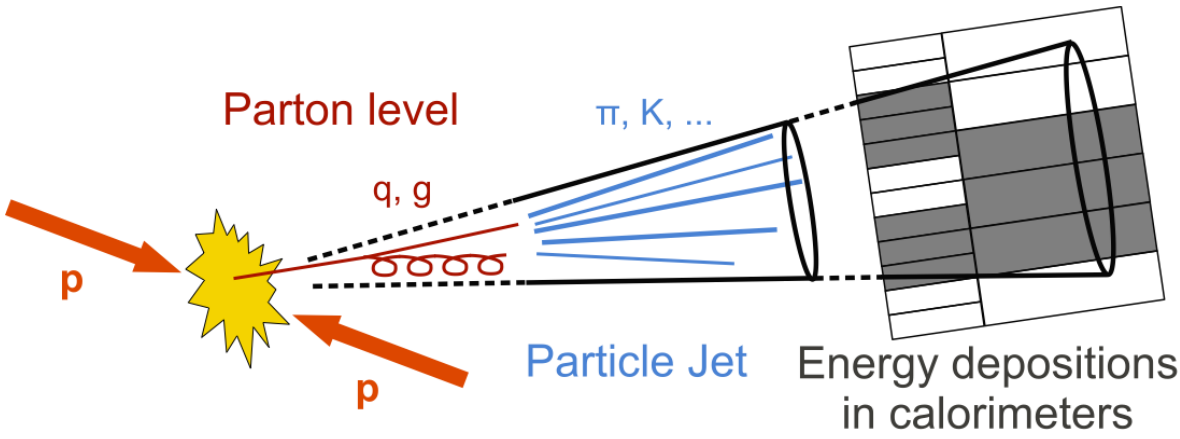
A different vertexing procedure is used for finding and reconstructing secondary vertices that are produced in decays of particles with longer lifetimes that originated from the primary interaction vertex. These are an essential ingredient for jet flavour identification algorithms that will be described in Chapter 4, since heavy flavour hadrons in general will undergo a second decay into light hadrons a short distance away (~ 0.5 mm) from the primary vertex. This is far enough away to be resolved, but it is close enough to still be in the beam pipe. The current secondary vertexing algorithm in CMS is the inclusive vertex finder (IVF) algorithm [69]. Secondary vertexing algorithms preceding it utilized just the tracks in the jet (see next section) connected to the secondary vertex, however the IVF algorithm uses all tracks in the event, and then the secondary vertices are mapped to the jets by ΔR matching of the vertex flight distance with the jet axis. This was observed to significantly increase the performance of the secondary vertexing, as tracks that were important for a vertex, might not have been clustered into the jet. The IVF algorithm works by identifying potential vertex seed tracks that are significantly displaced. Other tracks in the events are then clustered together with a seed based on the distance at the point of closest approach as well as their angles. During this process the distance between the track and the seed has to be lower than the distance to the primary vertex. Using the tracks associated to the seed, a vertex is fitted to obtain a flight distance and the vertex kinematics.

2.4.3. Jets

As mentioned in the previous section, the quarks produced at the LHC will hadronize into color neutral states and form a collimated spray of particles referred to as a jet.

These particles can be charged and neutral hadrons, but also photons, electrons and muons. Figure 2.18 shows an illustration of a jet being produced and measured. In general many particles are produced as a result of the hard scattering, and dedicated clustering algorithms are needed to identify which particles corresponds to a single quark. Once a jet is identified the energy of the jet needs to be reconstructed, and in order to get the best possible estimate, detector inefficiencies need to be accounted for and particles produced from pileup vertices need to be subtracted. Finally, it is of great value to identify the type of quark that initiated the jet. This task will be described in detail in Chapter 5 that includes an advanced algorithm developed for this.

Figure 2.18.: An illustration of a quark hadronizing into a jet that interacts with the CMS detector. From Ref. [70].



Jet clustering

The CMS experiment employs the anti-kT algorithm for jet clustering [71]. This algorithm is both infrared and collinear safe, meaning that if soft radiation or a collinear splitting from the original parton is added to the event, the clustered jet is mostly unaffected. The algorithm works by defining a radius parameter that indicates the average size of a jet in (η, ϕ) space. Picking this size parameter is non trivial as there is a trade off. A large size increases the chance of including all particles associated with the quark, while a small size reduces the amount of particles included from pile-up. The CMS experiment uses a radius parameter of $R = 0.4$ for the majority of the reconstructed jets (AK4 jets). The anti-kT algorithm works by calculating and

comparing two metrics shown in Equation 2.19 for particle pairs in the event.

$$d_{ij} = \min(k_{t,i}^{-2}, k_{t,j}^{-2}) \frac{\Delta\eta_{ij}^2 + \Delta\phi_{ij}^2}{R^2} \tag{2.19a}$$

$$d_{iB} = k_{t,i}^{-2} \tag{2.19b}$$

The indices i and j indicate particles being compared, and the transverse momentum of the particle is referred to as k_t. The separation of the particles in angular space is indicated as $\Delta\eta_{ij}^2 + \Delta\phi_{ij}^2$. For every possible particle pair in the event d_{ij} is calculated, and d_{iB} is calculated for all particles. Particle pairs are associated into a jet if the pair d_{ij} is found to be smaller than both d_{iB} and d_{jB}. Their momenta are combined into a pseudo-jet that replaces them for the continuation of the clustering. When no more particle pairs have d_{ij} smaller than both d_{iB} and d_{jB}, the algorithm stops. The intuition behind these metrics is to ensure that low-k_t particles are clustered around hard particles and to ensure infrared and collinear safety.

In the CMS experiment the reconstructed jets are clustered using the reconstructed particle flow candidates. However, the jets pileup is still clustered into the jets, and an additional pileup mitigation procedure is applied on the jets to reduce this further. The CMS experiment uses the charged hadron subtraction (CHS) algorithm [64] to reduce pile-up. It works by identifying reconstructed charged hadrons that are associated with reconstructed pile-up vertices, and then excluding them from the particle collection used in the clustering algorithm. Since neutral pile-up particles cannot be tracked, the effect of these are mitigated using a jet correction factor that depends on the total jet area [72].

Jet energy corrections

After the jet clustering of reconstructed PF particles is performed there is still a discrepancy with the momentum of the generated quark. This is due to a number of effects such as detector inefficiencies and response, as well as pile-up particles that might have been clustered into the jet. In order to correct for this a jet energy correction method is applied. The CMS experiment applies the jet energy corrections in three steps [72]. The first is referred to as the level 1 (L1) correction and it attempts to correct for the fake energy introduced in the jet from pileup particles that has not been accounted for by the already used pileup mitigation techniques. It is derived from

simulated dijets events, comparing the response with and without pileup. It is then parameterized in terms of four variables namely the jet area, the diffuse offset energy density as well as the jet p_t and η. The second and third energy corrections are called the level 2 and level 3 (L2L3) correction, and they aim at correcting for the detector response by comparing jets in simulated samples with a generator level jet that is ΔR matched, and deriving jet energy correction factors. They are parameterized only in p_t and η of the reconstructed jet. Finally, using both simulated events and data events from γ+jets, Z+jets and dijets process, residual corrections are derived that account for the difference between data and simulation. More details on this topic will be described in Chapter 5.

2.4.4. Missing Transverse Energy

The missing transverse energy (MET) is an important reconstruction quantity that can be used as an estimation of the momentum of particles that do not interact with the CMS detector such as neutrinos. It utilizes that the sum of the transverse momentum of all particles produced in the event should be zero as the colliding protons have no initial transverse momentum. To obtain the missing transverse energy a sum of negative transverse momentum vectors of all reconstructed particle flow candidates in the event is taken. Since the missing energy is also affected by jet energy corrections, these are propagated through in the MET calculation. Pile-up in the event also affects the MET and it is mitigated with dedicated algorithms [73].

Chapter 3.

Machine Learning

Machine learning (ML) refers to the science of creating computer algorithms without having to program explicit instructions. The algorithm produced with machine learning is a function that maps input data into a set of output values. Depending on how these values are represented the machine learning algorithm can perform different tasks such as classification or regression analysis. Usually these algorithms are defined into three groups of supervised, unsupervised and reinforcement learning. Supervised machine learning algorithms try to learn from training examples that consists of set of inputs as well as a training label that needs to be predicted. The algorithm is then improved by minimizing a defined loss function, which is based on the difference between the training label and the algorithm prediction. On the contrary, unsupervised learning tries to construct algorithms using unlabeled data, and therefore needs to be designed such that they can be optimized without having a target for every training example. Finally reinforcement learning deals with constructing intelligent agents that take action in an environment. Machine learning is highly useful for tasks that are difficult to solve with an algorithm defined with explicit instructions. Such tasks includes computer vision, natural language processing, speech recognition and self-driving cars. In particle physics experiments many such tasks exists like reconstruction of particle trajectories, jet identification and physics process event selection. An additional benefit for particle physics is that the machine learning algorithm can be trained on raw data without the need for extensive feature engineering by hand. In particle physics the raw data collected from the detector, often needs a lot of processing to be interpreted at the particle physics level. In every level of data processing some information can be lost that could be relevant for the next task in the analysis. Machine learning algorithms can be trained directly on data that has undergone minimal processing, referred to as low level data, skipping many levels of processing. The subject of curve fitting, which

is extensively used in particle physics, is very closely related to machine learning with the main difference being that in machine learning the structure of the function to be fitted does not need to be known. In the last 5 years, machine learning algorithms have proven to be a very powerful tool in high energy physics, and there has been a surge of interest in the topic. Particularly, the machine learning method of neural networks has become the golden standard in particle physics. In this chapter the machine learning methods and their adoption in the field of particle physics will be reviewed. First the basic building blocks of the neural network will be reviewed, followed by the state of the art methods used today. While many interesting possibilities for unsupervised learning and reinforcement learning in particle physics exists [74–76], the focus of the chapter will be on the supervised machine learning algorithms. In particle physics it is possible to obtain large datasets of high quality simulated training examples. These are easily labeled, and the datasets are therefore well suited for supervised learning. However this reliance on simulation does introduce some complications, since the machine learning algorithms have to be calibrated to data subsequently.

3.1. Neural Networks

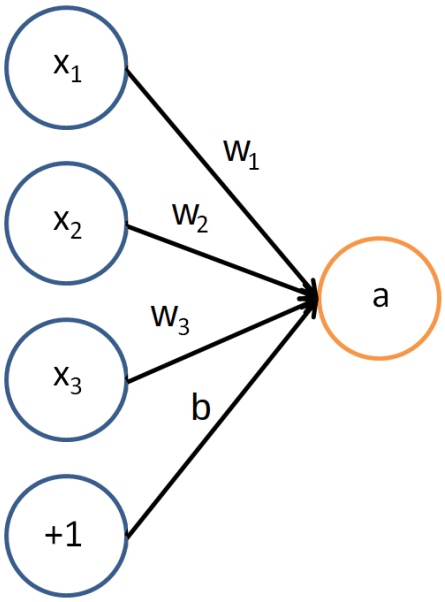
Machine learning as a subject dates back to the 1950s originally developed by neuroscientists and psychologists attempting to mimic the structure of the brain in the hope of constructing a generic algorithm that can learn a broad set of tasks and become a candidate for artificial intelligence [77]. One of the earliest and most important machine learning tools is the artificial neural network. The first neural network was the perceptron [78] that was tasked with identifying whether an object with a set of features, $\overline{x}$, was part of a specific class, known as binary classification. The perceptron consists of a weighted sum of the input variables, which is then passed to a threshold activation function as shown in Equation 3.1.

$$a(x) = f(b + \sum_{i=1}^{N} x_i w_i) \tag{3.1}$$

$$f(x) = \begin{cases} 1, & x > 0 \\ -1, & \text{otherwise} \end{cases}$$

Here N is the number of input features, x_i is the value of feature number i and w_i is a weight that is obtained by training the algorithm. The bias, b is also a trainable weight, but it is not multiplied with an input feature. The activation function, $f(x)$, is a threshold function that forces the output of the perceptron to be either -1 or 1. This allowed the perceptron to be used for binary classification with the output indicating the prediction category. An illustration of the perceptron architecture can be seen in Figure 3.1. The perceptron was a natural starting point for machine learning, as no

Figure 3.1.: An illustration is shown of the basic structure of a perceptron with three input variables. The connections between the nodes represent a dot product with the weight vector, $\overline{w}$. The connections originating from the +1 nodes are the bias b.



sophisticated learning algorithm was needed to choose a good set of weights. As a starting point the weights are set to zero. Then iterating through each training example, the weights of the perceptron can be updated with for instance Equation 3.2.

$$\overline{w}_{\text{updated}} = \overline{w} + r \cdot \overline{x_n} \cdot (y_n - a(\overline{x_n})) \tag{3.2}$$

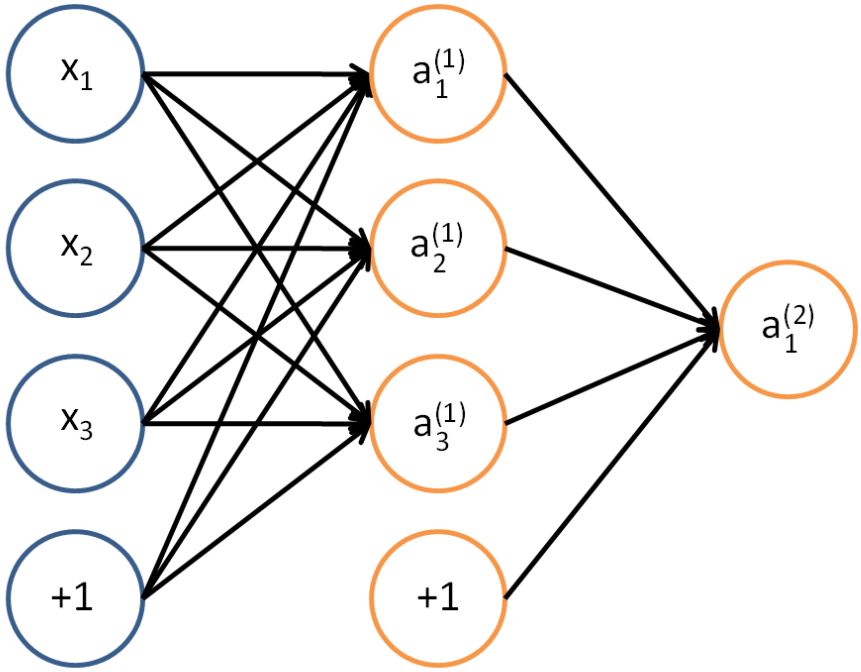
Here $\overline{x}_n$ refers to the input vector of training example n, $\overline{w}$ refers to the weight vector, y_n refers to the training label of training example n and $a(\overline{x_n})$ refers to the prediction of the perceptron using the previous set of weights. The parameter of the learning algorithm, r, is referred to as the learning rate. As can be seen from the equation, if the example was correctly predicted, then the weights are not updated, as $y_n - a(\overline{x_n})$ would be zero. If the prediction is incorrect the weights are updated to make the prediction closer to

the label. This correction term is motivated by the partial derivative of the weights for a loss function such as $L = \frac{1}{2}(\overline{y} - \overline{w} \cdot \overline{x})^2$. Shortly after the perceptron was discovered, it was proved mathematically that this optimisation procedure guarantees convergence after a finite number of error corrections [79]. However it was also realized that the perceptron was limited as a universal function approximator [80]. As it is only a linear model, it can for instance only be used for classification problems, where the classification categories can be separated with a hyperplane in the input feature space. One attempt to extend the number of problems that could be solved using the perceptron, is by mapping the original input features into a higher dimensional space, and then separating the categories with a hyperplane in this new feature space. The assumption is that the separation becomes possible or at least easier with additional dimensions. This approach evolved into what is known as kernel support vector machines [81] and was very successful. The second idea for extending the perceptron, was the realisation that by adding a hidden layer to the perceptron, a much more general function approximator could be constructed. This multilayer perceptron with a single hidden layer can be defined as shown in Equation 3.3 and an illustration can be seen in Figure 3.2.

$$\begin{aligned} a_i^{(1)}(\overline{x}) &= f^{(1)}(b_i^{(1)} + \overline{x} \cdot \overline{w_i^{(1)}}) \\ a^{(2)}(\overline{a^{(1)}}(\overline{x})) &= f^{(2)}(b^{(2)} + \overline{a^{(1)}}(\overline{x}) \cdot \overline{w^{(2)}}) \end{aligned} \tag{3.3}$$

The vector $\overline{a^{(1)}}$ represents the output values of the hidden layer. Each element of this vector corresponds to a neuron activation value of the hidden layer, where the number of neurons is a hyperparameter of the network. The vector $\overline{w_i^{(1)}}$ corresponds to the weights associated with neuron i in layer 1. This weight vector has the same dimension as the input vector $\overline{x}$. The value $b_i^{(1)}$ is the bias associated with neuron i. The function $f^{(1)}$ is the activation function of the hidden layer. The vector $\overline{w^{(1)}}$ corresponds to the weight associated with the output neuron and the value $b^{(2)}$ is the corresponding bias. The function $f^{(2)}$ is the activation function of the output layer. $a^{(2)}$ represents the output value. The activation function of the hidden layer, $f^{(1)}$, should be a non linear function. If a linear function is picked, simple linear algebra shows that the network simply reduces to a perceptron. The activation of the output layer can be a threshold function in case of classification problems, but it could also be a linear function, like the identity function, in case of regression analysis. The network defined by Equation 3.3 can easily be extended to have more hidden layers

Figure 3.2.: The basic structure of a multilayer perceptron with three input variables, one hidden layer with three neurons and one output node. The connections between the nodes represent dot products with the weight vectors associated with the layer. The connections originating from the +1 nodes are the biases $b_i^{(1)}$ and $b^{(2)}$.



by simply applying a new operation similar to $\overline{a^{(1)}}(\overline{x})$ on the output vector of the previous layer. It has mathematically been proven [82] that with a single hidden layer the multilayer perceptron is a universal function approximator. However depending on the complexity of the function to be estimated, the number of neurons needed in the hidden layer might be extremely large. The addition of the hidden layer does introduce additional complications, as it makes optimisation much more challenging, as the desired states of the neurons in the hidden layer are not directly specified by the task. For this reason, the multilayer perceptron was mostly ignored until 1986 when the back-propagation algorithm [83] was introduced. This algorithm made it possible to calculate an appropriate error correction term for the hidden units in a quick and efficient manner. Back-propagation is essentially an extension of the chain rule, and it relies on finding the partial derivatives of the neurons in network, with respect to a loss function that quantifies the difference between the prediction and the training example label. It is important that this loss function is differentiable for back-propagation to work. Before performing the back-propagation of error correction

terms, it is first needed to forward propagate through the network, calculating the value of the neurons and their activations, as these values are needed to calculate the partial gradients of the weights. Assuming the example of a binary classifier with 1 neuron in the final layer, referred to as layer n, the partial derivative of the loss function, $\frac{\partial L}{\partial a^{(n)}}$, with respect to the final activation can be calculated, since the loss function is differentiable. If the activation function is differentiable as well, the partial derivative of the neuron value can be calculated with Equation 3.4.

$$\frac{\partial L}{\partial x^{(n)}} = \frac{\partial L}{\partial a^{(n)}} \cdot \frac{\partial a^{(n)}}{\partial x^{(n)}} \tag{3.4}$$

$x^{(n)}$ is the neuron value prior to applying the activation function, i.e $x^{(n)} = \sum_i w_i^{(n)} \cdot a_i^{(n-1)}$, where $a_i^{(n-1)}$ is the activation of the neuron i in layer n-1, and $w_i^{(n)}$ is the weight that connects the neuron to the output node in layer n. After calculating the partial derivative of the output node, the chain rule can again be used to obtain the partial derivatives of the weights in the previous layer connecting to the output node as shown in Equation 3.5.

$$\frac{\partial L}{\partial w_i^{(n)}} = \frac{\partial L}{\partial x^{(n)}} \cdot \frac{\partial x^{(n)}}{\partial w_i^{(n)}} = \frac{\partial L}{\partial x^{(n)}} \cdot \frac{\partial}{\partial w_i^{(n)}} (\sum_j w_j^{(n)} \cdot a_j^{(n-1)}) = \frac{\partial L}{\partial x^{(n)}} \cdot a_i^{(n-1)} \tag{3.5}$$

These partial derivatives could then be used to optimize the weights connecting to the output layer, but in order to also optimize weights in the hidden layer, it is needed to obtain the partial derivatives for the neurons in the prior layers n-1, $\frac{\partial L}{\partial x^{(n-1)}}$, such that Equation 3.5 can be applied again. This can be found using the chain rule and the definition for the neuron value.

$$\frac{\partial L}{\partial x_i^{(n-1)}} = \frac{\partial L}{\partial a_i^{(n-1)}} \cdot \frac{\partial a_i^{(n-1)}}{\partial x_i^{(n-1)}} = \frac{\partial L}{\partial x^{(n)}} \cdot \frac{\partial x^{(n)}}{\partial a_i^{(n-1)}} \cdot \frac{\partial a_i^{(n-1)}}{\partial x_i^{(n-1)}} = \tag{3.6}$$

$$\frac{\partial L}{\partial x^{(n)}} \cdot \frac{\partial}{\partial a_i^{(n-1)}} (\sum_j w_j^{(n)} \cdot a_j^{(n-1)}) \cdot \frac{\partial a_i^{(n-1)}}{\partial x_i^{(n-1)}} = \frac{\partial L}{\partial x^{(n)}} \cdot w_i^{(n)} \cdot \frac{\partial a_i^{(n-1)}}{\partial x_i^{(n-1)}}$$

The back-propagation algorithm then starts with the output layer applying Equation 3.4 to calculate the partial derivative for every neuron in that layer $\frac{\partial L}{\partial x^{(n)}}$. With this quantity, the partial derivative for every weight connecting to the neurons in layer n can be calculated with Equation 3.5, and the partial derivative for every neuron in the previous layer n-1 can be calculated with Equation 3.6. After having obtained $\frac{\partial L}{\partial x_i^{(n-1)}}$,

the weights of the layer n-2 can then be calculated using Equation 3.5. Therefore the partial derivatives for the weights throughout the network can be calculated by starting from the output derivative, and going backwards through the network in a sequential manner. The weights of the network can then be updated with a number of optimisation algorithms, but a simple form of gradient descent could for instance be Equation 3.7.

$$\overline{w}_{\text{updated}} = \overline{w} - r \cdot \overline{\frac{\partial L}{\partial w}} \tag{3.7}$$

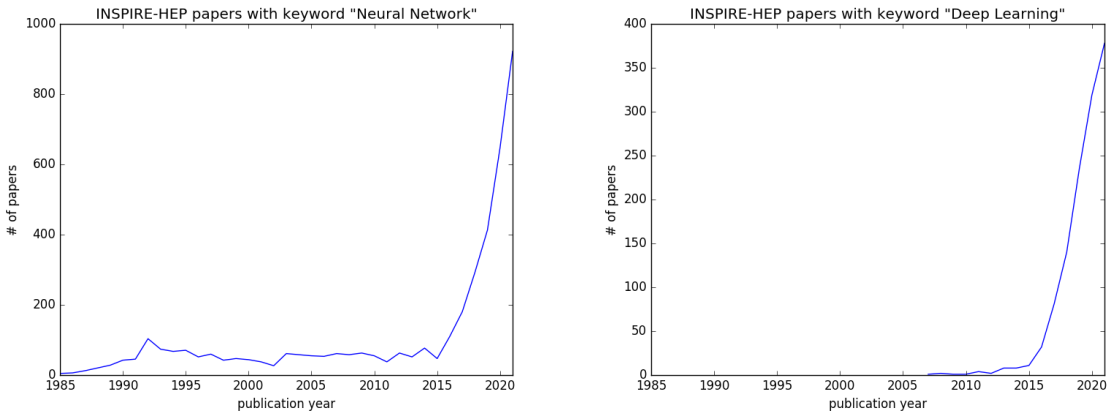
Back-propagation also meant that step functions, which had been used in the perceptron, stopped being used as activation functions, since the derivative of $\frac{\partial a}{\partial x}$ is zero, which means that all the gradients calculated with back-propagation are zero everywhere. Common activation functions used instead were the hyperbolic tangent and the sigmoid function. The weakness of these activation functions is that the derivative goes to zero as the argument goes to large values, and the derivative is bounded in the range from 0 to 1. If a network is made very deep the gradient can become very small in the early layers, since for every layer a small value is multiplied with to get the gradient. This is referred to as the vanishing gradient problem.

After the development of back-propagation, there was a surge in the use of neural networks to tackle problems of high complexity such as pattern recognition. Since tasks in particle physics like track reconstruction involve a degree of pattern recognition with large combinatorial complexity, neural networks were tried [84–86]. While showing initial promise, the neural networks were neither faster or better performing than traditional methods. In hindsight it is interesting that track finding was the first task to be attempted with neural networks, as even with modern methods track finding with neural networks remains a challenge. Jet identification was also tried with neural networks. The papers of Lönnblad et al. [87, 88] used a fully connected neural network to separate quark and gluon jets. They used the kinematics of the four leading particles in the jet as input and 10 neurons in a single hidden layer with three output nodes. This gave them better results than the traditional methods used at the time. It was also possible to identify b quarks to some extent as well, despite not using vertex information. Again despite promising intial results, likelihood based algorithms are still the main method of quark gluon tagging at CMS and ATLAS today [89]. The majority of papers in the 90s on neural networks in high energy physics were test of techniques and proof of concepts that showed promise, but they were never adopted in data

analysis. The first published physics application using a neural network was in 1992 by the DELPHI experiment at LEP, where a simple neural network with 25 hidden nodes were used for identifying the flavour of quark pairs in $Z \to q\bar{q}$, such that the partial widths could be calculated [90]. In 1997 at the D0 experiment at the Tevatron collider, a neural network was used for measuring the top quark mass [91]. The network had 4 inputs consisting of the missing transverse momentum, the event aplanarity as well as two engineered features. The hidden layer had 5 neurons, and the network was trained for separating top quark events from background events. While a successful approach at the time, later top quark mass analyses stopped using neural networks for classifying signal, reverting to more simple selection methods that where found to be more robust and better performing. The H1 experiment at HERA constructed and deployed a neural network L2 trigger [92] for separating physics processes. The CMS experiment also explored neural networks for triggering. A prototype perceptron L1 trigger for discriminating electrons and photons using calorimeter information was made, albeit this never ended up being implemented [93]. Indeed, while there was great excitement and interest for the prospects of neural networks in the 90s, at the end of the decade there was some disappointment as traditional techniques were often found to be either competitive or better [94]. This was mainly due to the networks that were used. They consisted of a single hidden layer with a small amount of neurons, and they were further limited by a small set of training examples, making it hard for them to learn complicated patterns using low level information. In 2004 the technique of boosted decision trees (BDT) was introduced to the particle physics community [95] as a competitor to the neural networks, and in many use cases the method outperformed typical neural networks used at the time, while being easier to tune and train. Neural networks remained part of the physics toolbox but they were not widely used. This was in parallel to computer science where interest in neural networks had slowed down significantly. In 2012 a groundbreaking development was made with a machine learning algorithm, AlexNet, that achieved state of the art performance for the task of image classification [96]. AlexNet was a deep learning algorithm, which is a machine learning algorithm that uses multiple layers. The network consisted of 8 layers where 5 were convolutional layers (see section 3.2.1) followed by 3 fully connected layers resulting in 60 million parameters, making it much bigger than past networks. This was made possible by a graphics processing unit (GPU) implementation for training and evaluating the network. This implementation was much faster than what was possible in the past. The vanishing gradient problem was overcome by replacing the hyperbolic tangent activation function, and instead using the ReLU activation

function $f(x) = \max(0, x)$ [97]. While the effectiveness of such an activation function is a bit surprising, since the gradient is either 1 or 0, vanishing gradients are avoided, making it easier to optimize a deep network. The network was trained on a dataset consisting of 1.2 million images making the number of free parameters much larger than the training set size. This can easily result in the phenomenon known as over-training, where the neural network learns very specific features of individual training examples, rather than learning generalizing trends. To avoid this a method known as dropout [98] was used. During training each neuron in the network has a fixed probability of not participating in the forward and back-propagation. This encourages the network to learn more general trends as specific neurons that have learned specific details are deactivated randomly during training. In validation and deployment all neurons of the network are used. After this paper showed the possibilities of deep neural networks (DNN), there was a surge in interest of deep learning and neural networks in computer science, but also in high energy physics. Shown in Figure 3.3 is the number of papers by publication year on INSPIRE-HEP with the keywords Neural Network and Deep Learning since 1985. While papers with neural networks has

Figure 3.3.: The number of papers released in a given year with the keywords "Neural Network" (left) and "Deep Learning" (right) are shown.



been around for many years, there has been an exponential growth in the number of publications using them since 2012. Deep Learning was barely mentioned before late 2000s, and similarly to neural networks the number of publications has increased exponentially. One of the main reasons machine learning has been so successful in particle physics is the availability of high quality simulation that can be used for training. This has allowed complex and deep architectures, because large training sets with accurate labels can be constructed easily. However it also introduces some novel

challenges, because despite the simulation being in good agreement with data, it is not perfect. This has created some hesitancy to create very large neural networks, as it is expected that there is a limit to how much the performance in simulation transfer to data. Deep neural networks in particle physics are therefore usually several orders of magnitude smaller than the ones used in state of the art computer science, despite having comparable sizes of training datasets.

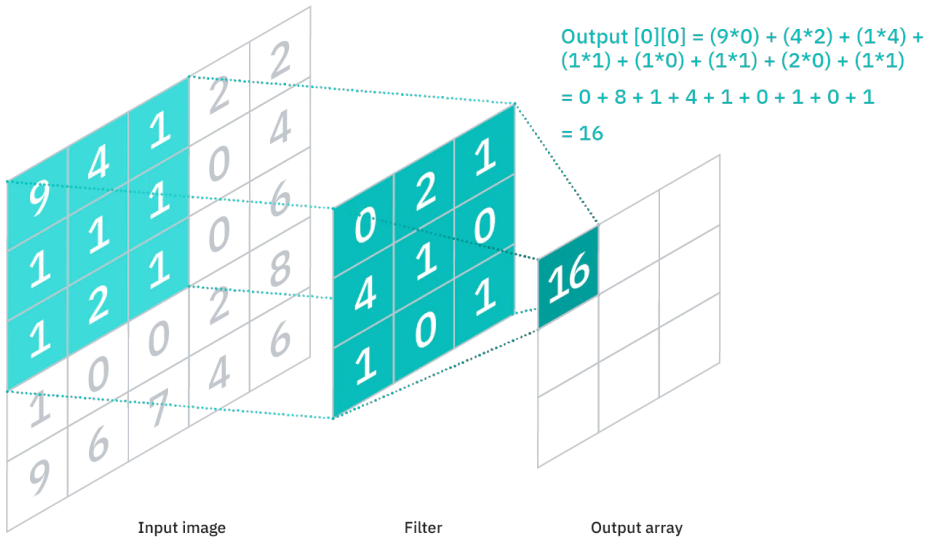
3.2. Architectures

Ideally a fully connected neural network with sufficient width and depth should be able to approximate any function, and therefore it should be possible to utilizes such a network for any given number of input data and problems. However, the AlexNet [96] paper made it clear that to properly train a deep neural network, the network architecture and layers needed to be well suited for the input data. Over the past 5 years the field of deep learning has seen the adoption of more complex and sophisticated network architectures, going much beyond the fully connected model. Rather than trying to apply the same general algorithm to every problem, a network is defined using prior knowledge about the specific task at hand. To really benefit from deep neural networks it is needed to take into account the symmetries and structure of the input data in the network architecture itself. The fully connected neural network should in theory be able to learn the exact same functions as these complex architectures, but in practice it turns out to be too difficult to train. Different network types have been constructed to handle different tasks such as image recognition on pictures and videos, object classification on point cloud data and natural language processing on text. In particle physics the research in machine learning has mainly been focused on developing new model architectures and designs that are specifically well suited to the particle physics data, as no machine learning model from traditional data science fits this task perfectly. This contains identifying the useful architectures that were developed for tasks that have data similar to what is used in particle physics, but also constructing architectures and layers specifically for particle physics.

3.2.1. Convolutional neural networks

The convolutional neural network (CNN) generally refers to a network which includes convolutional layers [99]. These layers were developed in the field of computer vision to be used instead of the multilayer perceptron. This was a needed development as image inputs in general have very high dimensionality. For instance a 1200x1200 pixel RGB color image would have 4.3 million inputs. If this is used as input to a fully connected layer with 100 nodes, there will be $100 + 100 \cdot 4.3 \cdot 10^6$ free parameters just in the first layer. In order to reduce this huge dimensionality, the convolutional layer seeks to exploit the symmetries of the images. In particular there is a local spatial structure in an image. Pixels adjacent to each other are more connected than two pixels in opposite sides of an image. Furthermore a group of pixels in one side of the image, is similar to a group of pixels in the other side of the image. Therefore the function applied should be invariant to the global location of pixels being considered. The convolutional layer is constructed by encoding these adjacency and invariance properties into a fully connected network. An illustration of the convolutional approach can be seen in Figure 3.4. This illustration shows a 2D CNN with a 3x3 kernel and a single

Figure 3.4.: The basic structure of the convolutional layer is displayed. From [100].



filter. The kernel is applied to a 3x3 sub-image by taking the dot product between the input pixels and the filter. It is called a 2D CNN because the kernel has two dimensions and therefore is encoding that there is adjacency in both x and y. In this case only one filter is shown, but in practice usually many more would be used. Each additional filter has a unique set of values for the 3x3 kernel, and it can therefore be trained to

capture a unique set of features from the image. After having computed the value for a sub-image the filter slides over to the next pixel to be computed. The number of pixels the filter moves after computation is called the stride. Often this value is just 1, but a higher values enables more dimensional reduction. In case of a stride of 1 and a 3x3 kernel size the example image shown in Figure 3.4 is reduced from 5x5 to 3x3. Each additional filter used in the CNN adds an additional output array. In the case of a color image where each pixel has three values (RGB), the layer can be extended to match this by adding three weights in each element of the filter, and then taking the dot product between these and the additional input layers. Adding color refers to having 2D CNN with a set of feature channels of three, rather than a 3D CNN, as there is no movement of the kernel in the color dimension.

A special subset of the convolutional layer, is the 1D convolutional layer with a kernel of size 1x1. Unlike 2D CNN developed for images, the 1x1 CNN does not incorporate the element of adjacency, but simply incorporates the properties of object invariance, since the weights of 1x1 kernel is shared throughout the input. In the case of particle physics, this is a useful tool, since in a lot of scenarios the input consists of several physics objects of a specific type, for instances jets or particles. Using a fully connected layer, it is possible to learn a completely different function for processing the first jet compared to the second jet, and the network has to learn itself that there is a similarity of the inputs. If you use a 1D convolutional layer to process such objects instead, it is ensured that a consistent function is used to process them, and the physics information is encoded directly into the network.

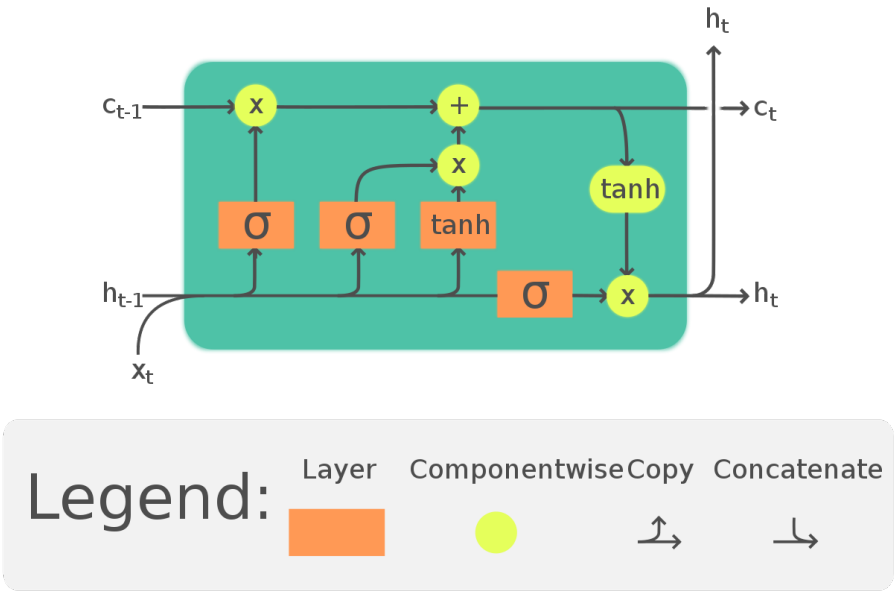
Convolutional based architectures were one of the first approaches beyond fully connected networks to be adopted in the realm of particle physics. The most simple approach is to try to represent particle physics data as an image, and then use conventional image recognition approaches. The ImageTop algorithm [101] tries to discriminate between boosted top quark jets and QCD jets by representing the jets as an image in ϕ and η space with the p_t of particle candidates in a given pixel yielding the pixel intensity. A series of 2D convolutional layers is then employed to perform the classification. The advantage of this, is that the 2D CNN naturally fits well to extract kinematic jet substructure, which is highly relevant in boosted jet tagging. However particle physics data is not naturally well suited for an image representation. The modern particle physics detector is represented by many different kinds of measured variables, for instance information about reconstructed tracks curvatures and impact

parameters. While it is easy to aggregate p_t information in pixels with a scalar sum, it is not possible to aggregate a variable like a tracks impact parameter. This loss of information could possibly be solved by making the image more granular, but this would make the image more sparse with more empty cells which causes several other issues with optimisation, performance and computational overhead of a 2D CNN. Furthermore, particle physics detectors can have irregular geometry that cannot be put into a perfect grid. Another approach for employing a CNN is the DeepAK8 [101] top quark tagging algorithm, which uses convolutional layers with kernel size 1x3, acting on a particle list rather than a jet image. This list is sorted by the particles p_t, and the convolutional layers acts on groups of three particles in descending order of p_t. This avoids the problem of having empty inputs, and it allows including every possible input features associated with particle candidates. The disadvantages of this approach is that it is not really clear how to sort the particle list in an optimal way, as it might not really be beneficial to convolve the three highest p_t particles with each other. It is also difficult to define the size of the kernel, there is no clear argument for why it should be optimal to convolve particles in a group of three.

3.2.2. Recurrent Neural Networks

The recurrent neural network (RNN) is a class of networks that are used to process sequential data. It has been heavily used in the domain of natural language processing, as it has the ability to encode the structure of the data, in this case the order of the words in a sentence. However, other tasks that have data that is sequential in nature can also be tackled by RNNs such as video analysis, speech recognition and time series prediction. Several types of RNN layers exist but one of the most common and powerful is the long short term memory (LSTM) architecture [102]. The main commonality of RNN layers, which also applies for the LSTM, is that they consist of some repeating modules, which are referred to as cells. The LSTM architecture consists of several LSTM cells, shown in Figure 3.5. These cells act on elements of the sequence to be considered, and they go beyond the fully connected network by containing a memory mechanism. The h refers to the output of LSTM cell with h_t being the current cell and h_{t-1} being the previous. x_t refers to element t of the input sequence. The memory value, c, is the unique aspect of the LSTM cells. This value is passed through each cell as they process elements of the sequence, and each cell can modify the memory value. The first operation of the LSTM cell is the forget gate, which is a fully connected

Figure 3.5.: The structure of a LSTM cell is shown. From [103].



layer with a sigmoid activation function that takes h_{t-1} and x_t as input. Due to the sigmoid function it outputs a number from 0 to 1, which is then multiplied onto the memory value c_{t-1} from the previous cells. This regulates how much of the previous memories of the LSTM chain should be forgotten. Next, new information based on the current element of the sequence is added to the memory. A fully connected layer with h_{t-1} and x_t as input and a tanh activation function is calculated to obtain a memory value for the current sequence element to be added to the global memory. This is multiplied with the output value of a new sigmoid layer, in order to regulate how much the current element of the sequence should modify the memory value. This obtained memory score is then added to the existing memory value. A tanh activation function is applied to this updated memory, which are multiplied together with the output of a final sigmoid layer that takes h_{t-1} and x_t as input. The cells new memory value and output value are then passed on to the next LSTM cell, until every element in the sequence x has been considered. Importantly each LSTM cell shares the parameters with every cell in the sequence encoding that the objects in the sequence can be treated in a similar manner. The LSTM layer can have varying types of output dimensionality. The number of neurons in the fully connected layers inside the LSTM cells should all be the same, and this size is equal to the size of the output vector h_t. The LSTM shown in Figure 3.5 outputs one vector for every element in the sequence, however LSTM layers can also be configured to only give a single output

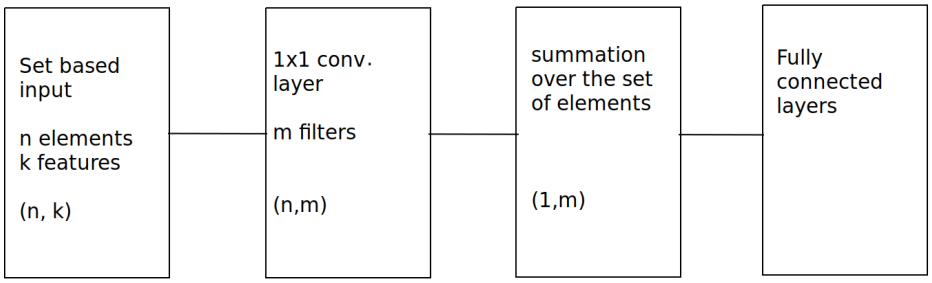
for a sequence, which is the output from the final cell in the chain. This is useful for particle physics, as you often want to consider a sequence of inputs, such as particles in an event, but the desired output often concerns a single event or jet, therefore requiring dimensional reduction. Another useful aspect of the LSTM cells for particle physics is that they can handle variable length input well, which is often the case in particle physics where there can be a different number of jets or particles in every event.

RNNs have been used as a technique for representing particles in a jet as it encodes that each element in the sequence is a similar object, and it is flexible at handling variable length input. It has been successfully employed for jet tagging, for instance boosted Higgs boson tagging [104]. Since RNNs are good at handling time series they have also been used for monitoring and fault protection of the LHC superconducting magnets using voltage distributions [105].

3.2.3. DeepSet

The DeepSet neural network [106] is an attempt to construct an architecture that is well suited for processing data which has a structure as a set such as point clouds, which are used for robotics and self driving cars. The important difference between sets based data and sequence based data, is that there is an inherent order for the sequence, whereas the set data is invariant under permutations. The DeepSet network incorporates this permutation invariance into the network architecture. It consists of a 1x1 convolution layer for transforming every element of the set into a latent space representation. Then a permutation invariant operation is performed to aggregate the elements of the set. A simple example of a network architecture can be seen in Figure 3.6. In the original paper, a simple sum is performed over every element of the set. This ensures that the network returns the same output no matter how the input is ordered. Many types of particle physics data is set based data, making the DeepSet network appealing. For instance the constituents of a jet have no inherent order, so a jet can be treated as a set. The DeepSet approach has been tried in an architecture for quark/gluon tagging named energy flow networks [107] with good results. However, the DeepSet approach is also somewhat lacking in that the aggregating function is usually quite limiting and inflexible, therefore becoming an information bottleneck.

Figure 3.6.: A simple example of a DeepSet network. The 1x1 convolutional layer transforms the set elements into a latent space representation where they are summed together. This can then be fed to a fully connected network leading to the output nodes.



3.2.4. Other architectures

Since neither LSTMs, CNNs and DeepSets are a "one-size-fits-all" architecture for particle physics, there is constant experimentation in trying to find new architectures for a given problem. There has been a lot of interest in graph neural networks for high energy physics. These are a category of networks that treats the input data as a graph, meaning a set of points that have neighbors connected in vertices. This type of network was popularised for molecule data, since molecules can easily be interpreted as a graph. They have also been used in connections with self driving cars and robotics, because these use point clouds in a physical 3D space as input. For these inputs each point has a clear set of neighbors based on the distance in Cartesian coordinates, making it graph like. Different types of graph neural networks exists, but in general they consist of some clustering algorithm, usually the k nearest neighbor (KNN) algorithm, that calculates a set of neighbors for each input element. A network layer then acts on this cluster of points. For particle physics these types of networks have been tried for several tasks such as jet tagging for highly boosted particles [108] and particle reconstruction in calorimeters [109,110]. The graph neural network fits particle physics experiments very well to represent data such a cells in a calorimeter, which can easily be represented as a point cloud in 3D coordinates, and it is more suited than CNNs since the detector might have irregular geometry that does not fit on a grid. For jet physics these networks have had great success, however it is unclear if this is really an optimal representation. Particle candidates in a jet are similar to a point cloud as they both are a set without a clear order. However for a point cloud there usually is a clear metric for the distance between the points, which is then encoded into the network with the KNN algorithm. For a jet there is no such

obvious metric for particle adjacency and different tasks can require different metrics. For a task such as AK8 jet top quark tagging where the jet substructure is highly important, a distance metric between particle candidates could be the distance in (ϕ, η) space, but it is unclear if this is optimal. For something like flavour tagging, where the tracking information provides the most important variables, using the distance in (ϕ, η) space is almost certainly not optimal. Since the KNN algorithm contains a sorting operation, it is not differentiable, which means that it is not possible to optimize the distance metric during training to automatically find an optimal metric.

An approach alternative to the graph neural networks is transformer networks. These networks were originally developed for language models, and are like RNNs designed for sequential data. However, they are more flexible in terms of how the data is ordered, as they utilize a mechanism called self attention to assign relevance between the elements of the input. An interesting application of transformer networks is jet to parton assignment in the reconstruction of top quark events [111]. They have also been employed for jet tagging of highly boosted particles [112], showing great promise. Another powerful tool of the transformer network, is that parts of the network can be pretrained. This refers to a process where the network is trained to predict its own input values. The pretraining is done using a dataset where some input variables of each training example are randomly masked, such that they are hidden from the network. Then the network is trained to predict these masked variables using the ones available. This allows the structure of the input data to be encoded into the network in a general way. For instance such a pretraining could be applied on jets, and this pretrained network could then be utilized for any number of jet physics use cases. The pretraining can be done on unlabeled training data, which would enable the use of experimental data rather than simulation and might be a powerful tool for mitigating data to simulation discrepancies.

Chapter 4.

Deep Neural Network for Jet Identification

This chapter mostly covers the content of the JINST publication "Jet flavour classification using DeepJet" of Bols et al. [113].

At the LHC experiments it is critical to identify which hard parton is at the origin of the measured jets, and to identify its flavour. For instance the top quark decays to a bottom quark more than 95.5% of the time, making it essential to identify jets that originated from the fragmentation of b quarks, if top quarks are to be identified. For other physics processes it is needed to identify jets originating from charm quarks such as for a Higgs boson decaying to a charm quark pair. The jets that originate from heavy flavour quarks, bottom and charm quarks, have several distinguishing characteristics that can be exploited for classification. Due to the fragmentation process of the heavy flavour quark, heavy flavour jets contain a charm or bottom hadron that often has a large fraction of the momentum of the quark. The bottom hadrons have a lifetime of $\tau = \sim 1.5$ ps, giving a range of typical decay lengths from a few millimeters up to a centimeter. Because of this, they usually decay before they reach the CMS inner tracker, yet their decay is far enough away from the primary collision vertex for the decay to create a secondary vertex that can be reconstructed from displaced tracks. The charm hadrons generally have a slightly lower lifetime of $\tau = \sim 1$ps, which also makes them identifiable from secondary vertices, albeit slightly more challenging as their signatures are in between that of jets originating from light (uds) quarks and b quarks. Beyond heavy flavour tagging, the task of separating light jets into quark and gluon jets, is another difficult challenge that is useful for α_s measurements, measurements

of parton distribution functions and new physics searches. The differences between quark and gluon initiated jets are subtle, but in general quark jets are more narrow than gluon jets. The quark jets have fewer particle constituents, and individual constituents are more likely to carry a larger fraction of the jet momentum. The classification of all jet flavours requires leveraging information of the lifetime, fragmentation and hadronization properties of the different flavours. From first principles it is not theoretically possible to define an all encompassing variable that can do this, and much less so from experimental observable variables. Due to this inherent multivariable nature of jet identification, machine learning has been employed extensively to solve this.

Historically several different machine learning methods have been used for b tagging in the CMS experiment. Initially algorithms were separated in track based taggers and secondary vertex based taggers that were then combined to perform discrimination [114, 115]. Eventually combined algorithms [116, 117] based on either neural networks with a single hidden layer or boosted decision trees were developed. These combined algorithms directly incorporated information of both input objects. The ATLAS b tagging algorithms generally have stuck with keeping vertex based algorithms and track based algorithms separate. These individual algorithm predictions are then combined in a final high level tagger [118, 119]. The ATLAS track based b-tagger utilizes a recurrent neural network [120]. This algorithm relies on a small number of reconstructed tracks associated with the jet. These tracks are required to be of high quality and strict selection requirements are imposed, for instance requiring 7 or more silicon hits for tracks to be considered. A secondary vertex based b-tagging algorithm is then used to supplement this track based tagger. The CMS b-tagger, DeepCSV [117], is a fully connected neural network. It uses 5 hidden layers with 100 neurons each. As input variables it uses several high-level engineered features as well as the jet p_t and η. It also relies on track features such as the 2D impact parameter significance. The six most displaced tracks associated with the jet are used as input to the algorithm. As with the ATLAS algorithm these tracks are required to pass strict selection requirements. DeepCSV differs from the ATLAS algorithms in that it directly incorporates the input features from a potential secondary vertex, making it a combined track and vertex tagger.

In general, traditional jet flavour identification algorithms rely on careful track and feature selection of the inputs to make the classifier learn easier and thereby perform better. However the selection of tracks and high-level features is done by hand, and

these tight selection criteria introduce loss of information that could be useful for the jet flavour identification. Additionally many jet identification algorithms rely on combining a separate track based and vertex based algorithm, which can hide important input correlations from the final algorithm. In this chapter the jet flavour identification algorithm DeepJet is described. This algorithm uses a novel neural network architecture that overcomes the input limitations past models had. It allows the algorithm to simultaneously use all jet constituents, secondary vertices as well as event-level variables. Secondarily, jet flavour tagging was previously done using separate specialized algorithms for identifying specific flavours. This algorithm unifies the different task of jet flavour identification, achieving state of the art performance in both b-tagging, c-tagging and quark-gluon tagging.

4.1. Training Samples and labeling

The DeepJet neural network is trained using a sample of AK4 [71] jets produced in simulated events. Two simulated physics processes are used. The first is fully hadronic top quark pair ($t\bar{t}$) events generated with POWHEGv2 [121–124] as they contain a high rate of heavy flavour jets. QCD multijet events generated with PYTHIA8 [39] are also included, as this sample enables large statistics of heavy flavour jets at very high transverse momentum, p_t. The top quark pair ($t\bar{t}$) b and c jet spectrum falls rapidly at high p_t, making it insufficient on its own. PYTHIA8 is used for the parton showering and hadronization in both cases. The GEANT4 [125] program is used to perform the detector simulation. The geometry of the CMS Phase 1 detector [49,126] is used. Pileup in the jets have been mitigated with the charged hadron subtraction method [127].

The b-tagger is supposed to be used in a general event topology beyond QCD multijet and hadronic top quark pair events. To avoid the network learning specific kinematic effects of these physics processes, it is needed to reshape the jet transverse momentum and pseudorapidity distributions to be the same for all jet flavours. In order to do this jets from each flavour category are randomly removed until the given flavour transverse momentum and pseudorapidity distributions are identical to the b jet distributions. After this process the sample consists of $\sim$130 million jets. The flavour of the jets are assigned through the process of ghost association [117]. Copies of the generated b and c hadrons are made before they are allowed to decay. These copies, referred to as ghosts, have their momentum scaled to an infinitesimal value. The

ghosts are then included in the reconstructed jet clustering, and if a jet contains at least one ghost b hadron, it is labeled as a b jet. If a jet contains at least one ghost c hadron, and no ghost b hadrons, it is labelled a c jet. If neither ghost b or c hadrons are present, the jet is labelled as a light jet. To further separate these light jets into quark and gluon jets, parton level information is needed. This is done by matching light jets with the closest quark or gluon parton in ΔR space. A similar ghost association process is used for this. If no ghost parton is within $\Delta R < 0.4$ of a light jet, the jet flavour is considered undefined, and it is not used for training. The b jets are also further labelled into three categories consisting of b jets where the b hadron is leptonically decaying (b_{lept}), b jets where the b hadron is decaying hadronically (b), and jets where there are two b hadrons (bb). The addition of the bb jet definition was done since $g \to b\bar{b}$ is a common physics analysis background, and it might be useful to exclude this. The separation of hadronically and leptonically decaying b jets was done to allow for calibration cross checks, i.e. the calibration of the performance obtained with simulated events to data. This is because some b jet calibration procedures rely on using leptonically decaying b jet, but are then applied on hadronically decaying b jets as well. In order to monitor overfitting, the method of cross-validation is used. A data set, which the algorithm is trained on, is made from 76.5% of the total collection of jets. A fraction of 10% of the jets is kept as a testing sample that is used to monitor that the algorithm performance generalizes beyond the training set. Consequently all performance plots are made on this sample. Finally, a validation set consisting of 13.5% of the total number of jets is used to monitor the performance of the algorithm during training on a sample separate from the training sample. This can help tuning the algorithm as well as identifying overfitting, as it is occurring, during training.

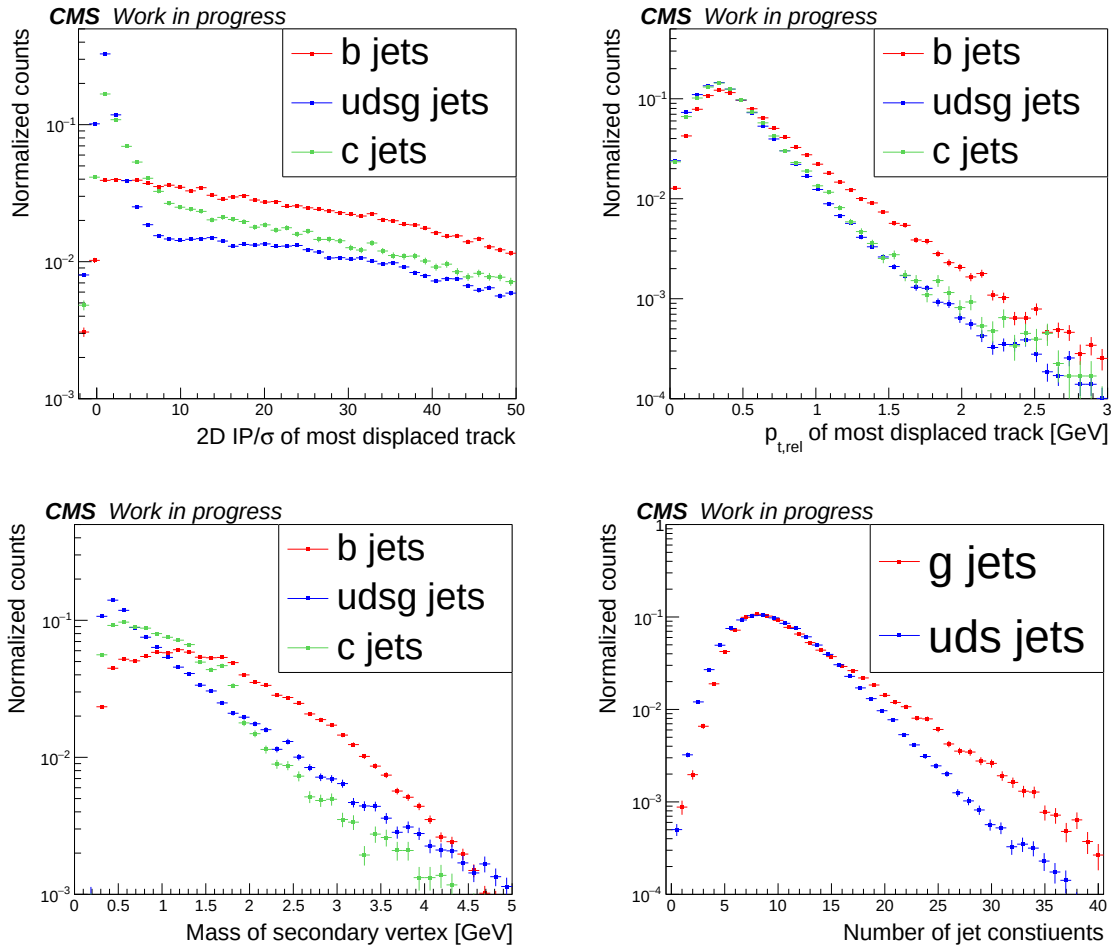
4.1.1. Input features and preprocessing

The DeepJet algorithm uses ~ 650 input features. These are separated into four categories. The first category groups global variables, which consist of jet level variables such as the jet p_t and η, the jet constituent multiplicity, and the number of secondary vertices in the jets. One event variable is also included in terms of the number of reconstructed primary vertices (N_{PV}) in the event. Usually, event level variables are hidden for jet identification algorithms to avoid the algorithm inferring the jet flavour based on the event topology, as this can ruin the algorithms ability to generalize to different physics processes. However, the N_{PV} variable doesn't give information on

the event topology on its own. It is included, since additional primary vertices produce additional particles in the event that can affect the jet reconstruction. By including N_{PV}, the network can identify which jets are most affected by this and take this into account in the jet identification. The most important features for DeepJet are the charged particle flow (CPF) candidates. Up to 25 charged particle flow candidates are considered, and they are sorted by the track impact parameter significance. This variable is the track impact parameter displacement with respect to the primary vertex, divided by the uncertainty on the measured impact parameter value. This variable is used for the sorting as it is assumed to be the most important variable. If more than 25 charged candidates are associated with the jet, the 25 candidates with the largest track displacement significance are picked. Each charged particle flow candidate has 16 features consisting of different kinematic variables as well as tracking variables, most notably the track impact parameters. Neutral particle flow candidates are also included in DeepJet. Traditionally, this information was not used for heavy flavour tagging, as it has less discrimination power compared to the charged candidates. It is however important for quark gluon tagging. Up to 25 neutral particle flow candidates are included, and they are sorted by decreasing angular distance to a secondary vertex. If no secondary vertex is in the jet, they are sorted by decreasing p_t. Each neutral particle flow candidate has 6 features, mostly kinematic. Up to four secondary vertices are considered, and they are sorted by displacement significance. They each have 12 features containing information on the kinematics and displacement. The full variable list can be found in Appendix A.

Figure 4.1 shows some selected input variable distributions. The top left distribution shows the transverse impact parameter significance of the most displaced charged particle flow candidate. As can be seen, b and c jets are more likely to have displaced tracks compared to light jets due to the lifetime properties of heavy flavour hadrons. The top right plot shows the p_t relative to the jet axis of the most displaced CPF candidate. It is seen that, on average, the b hadron decay products have larger $p_{t,\mathrm{rel}}$, illustrating the harder fragmentation of b quarks compared to light quarks. The bottom left plot shows the invariant mass of the tracks associated with a secondary vertex that was matched to the jet. The light jet distributions is described by a falling spectrum, whereas bumps at higher masses are observed in the c jet and b jet distributions, since the HF hadrons have large masses. The differences between quark and gluon jets are more subtle. The bottom right plot shows the total number of jet constituents in quark and gluon jets, where it can be seen that gluon jets generally have more constituents than quark jets.

Figure 4.1.: Four different input variables are shown for selected jet flavours. The 2D IP significance of the most displaced track (top left), the p_t relative to the jet axis of the most displaced track (top right), the secondary vertex invariant mass (bottom left) and the number of jet constituents (bottom right).



Reconstructed objects in the events can in rare case have features that are either not available or have infinite values. In that case the value is replaced with a value that is purposefully picked to be outside of the physical range of values for the specific feature, without being too large to cause numerical issues in the training. No normalization of the features is done at this stage, as this is built into the architecture itself. Several input variables have values that are also bounded within a physically well defined range.

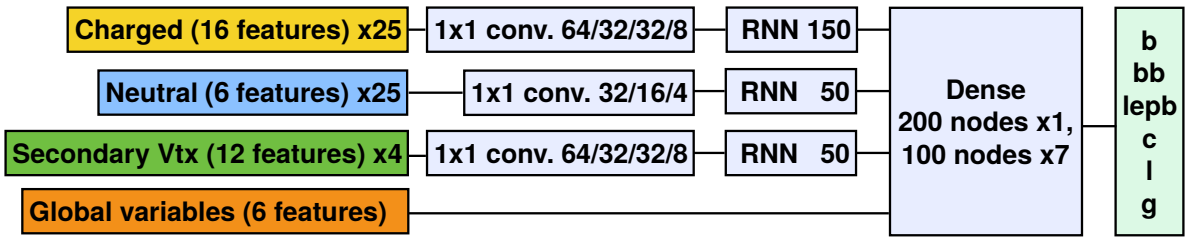


Figure 4.2.: The architecture of the DeepJet algorithm is shown.

4.1.2. Neural Network Architecture

The architecture of DeepJet is shown in Figure 4.2. It consists of a series of convolutional 1x1 layers, processing each type of physics object. This exploits the object symmetry, forcing the network to learn the same transformation for each physics object regardless of their input order. The charged particle flow candidates are input to a series of four convolutional layers with 64, 32, 32 and 8 filters respectively. Since the number of filters is decreased at every step, this can be seen as an automatic feature engineering, condensing the original 16 input features per charged particle flow candidate into 8 optimised jet tagging features per charged particle flow candidate. The secondary vertices are processed with an identical series of convolutional layers. The neutral particle flow candidates are processed by three 1x1 convolutional layers in sequence with filters of size 32, 16 and 4. Fewer layers with smaller filters were chosen as it is assumed that the neutral candidates were less important, and because there are fewer input features compared to the charged particle flow candidates and secondary vertices. All the convolutional layers use the ReLU activation function [128]. After this step, the objects with the learned features are input to LSTM layers. The LSTM layers ensure that a similar transformation is learned for every physics object, since the parameters are shared across the LSTM cells. The order of the inputs does however matter, since the LSTM memory value fades as objects are processed, making the last object in the sequence most influential. It is therefore ensured that the object considered most important, based on the ordering from the previous section, are shown to the LSTM last. It was tried to add a sorting layer before the LSTM, which would order the objects based on one of the engineered features. However, since the gradient doesn't propagate through the indices of the sorting, the network wasn't able to optimize the convolutional layers to create a better order than what was done by hand. The charged particle flow candidates are processed by an LSTM of size 150, whereas the neutral particle flow candidates and the secondary vertices are processed

by an LSTM of size 50. To combine the information from the groups of physics objects, the outputs of the LSTM layers are concatenated and together serves as input to a series of fully connected layers. The first has a size of 200 nodes. This is followed by seven layers with 100 nodes each. The number of layers were determined by adding layers until performance stopped improving. The size of the layers were found as a balance of memory consumption and performance. These fully connected layers all use the ReLU activation function. The final layer is an output layer with 6 nodes, representing the six target classes for the training. The output layer uses the softmax activation function, which ensures each neuron activation has a value between 0 and 1, and that sum of the nodes is 1. In between every layer of the network, there is a batch normalization layer [129]. This layer normalizes the features in a batch by subtracting with its mean and dividing by its standard deviation calculated on the batch. It then applies a transformation $y = \gamma x + \beta$ where γ and β are trainable parameters of the network. This allows the network to rescale the features in a manner that is optimal for convergence. Batch normalization is applied to avoid sudden shifts in the output distributions of neurons, when the parameters are being updated in the network during training. A batch normalization layer is also applied at the start of the network. Additionally dropout of value 0.1 is applied in between every layer.

4.1.3. Training procedure

The training is done on batches of 10000 jets. The full data set is iterated through 65 times (65 training epochs). The loss function used is the categorical cross entropy function. During the training, the loss function is evaluated on the validation data set, and its value is monitored to ensure there is no overtraining. If the validation loss stops improving during training for more than 10 epochs, the learning rate is halved. The ADAM [130] optimizer is used with a learning rate of 0.0001. The beginning batch normalization layer applied on the input features is only trained for the first epoch, as the distributions of the input features of course do not change during training. The training is performed on an Nvidia GEFORCE GTX 1080 Ti GPU, which takes three days on the training sample of 130 million jets with a total size of $\sim$ 130 GB. The model was implemented and trained using a dedicated machine learning framework for particle physics based on KERAS+Tensorflow [131–133].

4.2. Jet flavour identification performance

While the DeepJet network is trained to identify the b jet sub-labels (b,b_{lept} and bb) described earlier, they are not used as separate classes for discrimination. The final b jet discriminator is instead the sum of each of these nodes. This was done, since the discrimination power between these groups is low, and because they are hard to calibrate separately. It could however still be an interesting prospect to try and remove the bb node from the b jet discriminator in the future, since it might remove gluon splitting $g \to b\bar{b}$ that is a background to most physics analyses that use b jets.

The heavy flavour identification performance of DeepJet is compared to the previous CMS heavy flavour identification algorithm, DeepCSV. In general, heavy flavour identification performance varies a lot with the p_t of the jet. At very high p_t the tracks are more collimated and are therefore more difficult to resolve and reconstruct properly. Additional more b hadrons will be so energetic that they make it past the innermost pixel layer before decaying. Very low p_t jets can also be difficult due to multiple scattering effects and overall lower track quality. In general, the best b tagging performance is usually observed in the region of 90 GeV $< p_t <$ 150 GeV for jets. To quantify the performance improvement in these different kinematic regions as well as in different sample flavour compositions, different physics event topologies with varying jet selections are examined. The main plot used to identify performance gain is the receiver operating characteristic (ROC) curve. For b-tagging, for different discriminator cut values, the efficiency of identifying a b-jet labelled jet as a b-jet is plotted against the efficiency of identifying jets with another label as a b-jet. The latter efficiency is called the misidentification probability. Since it is more difficult to separate b-jets from c-jets than it is with light jets, separate ROC curves are made for b vs. c and b vs. udsg discrimination. A comparison is made on a fully hadronic $t\bar{t}$ sample with two different p_t cuts, shown in Figure 4.3. It is observed that DeepJet performs significantly better than DeepCSV in both separating b from light jets, as well as b from c jets. With the inclusive selection of $p_t > 30$ GeV an absolute improvement of $\sim$13% at 10^{-3} misidentification probability for b vs. light jets is observed. This corresponds to a relative improvement of $\sim$25%. For the more high p_t selection of $p_t > 90$ GeV, the improvement is slightly bigger with an absolute efficiency increase of almost 20% at 10^{-3} misidentification probability corresponding to a relative improvement of $\sim$30%.

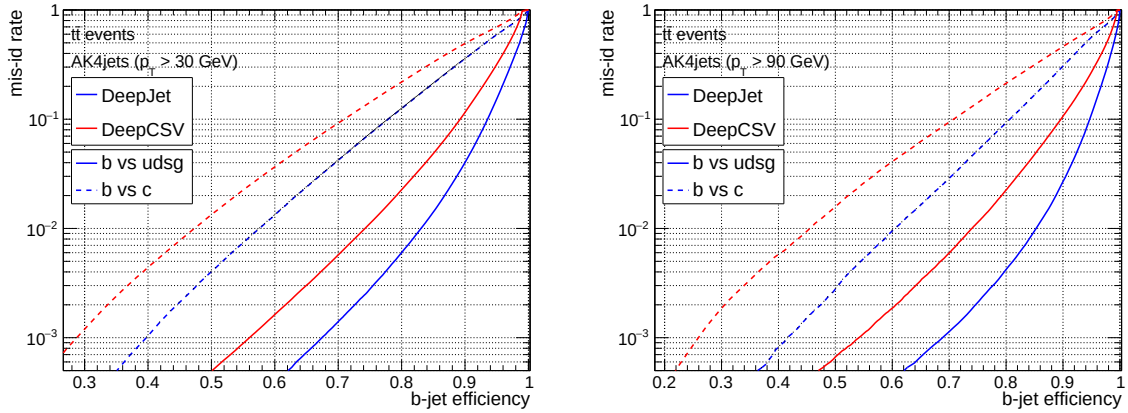


Figure 4.3.: A comparison of the b-jet identification performance of the DeepJet and DeepCSV algorithms is shown for $t\bar{t}$ events where the top quarks decay hadronically. A selection criterion of $p_t > 30$ GeV (left) and $p_t > 90$ GeV (right) is imposed. The performance is shown for b vs. light (solid lines) and for b vs. c classification (dashed lines).

Since there are few jets with p_t greater than 150 GeV in the fully hadronic $t\bar{t}$ samples, a QCD multijet sample is used to measure the performance in the high p_t regions. In Figure 4.4 ROC curves calculated using jet selections of 150 GeV $< p_t <$ 300 GeV, 300 GeV $< p_t <$ 600 GeV and 600 GeV $< p_t <$ 1000 GeV are shown. Large performance gains are seen in each case with an absolute (relative) efficiency improvement at 10^{-2} misidentification probability of $\sim$12% ($\sim$15%), $\sim$19% ($\sim$33%) and $\sim$20% ($\sim$40%) respectively. This improvement is assumed to be caused in part by the removal of the track selection requirements used in the previous b tagging algorithms. These were designed and optimised for jets in the 30 GeV to 130 GeV p_t range, and this selection remove a lot of important information in high p_t jets. Since the DeepJet algorithm can make a track selection itself, it gets a significant advantage over DeepCSV in this region of the phase space. The b jet efficiency as function of the jet p_t is shown for fixed light jet misidentification rates of 10%, 1% and 0.1% in Figure 4.5, and the same trends seen in Figure 4.4 are observed.

The DeepJet algorithm achieves state of the art performance in c jet identification as well. It is again compared with the DeepCSV model. The c vs light jet discriminator is constructed from the DeepJet output nodes for light quarks ($P(\text{uds})$), gluons ($P(\text{g})$) and c quarks ($P(\text{c})$) as $\frac{P(\text{c})}{P(\text{c})+P(\text{uds})+P(\text{g})}$. A comparison of the c jet identification capabilities of DeepJet and DeepCSV is shown in Figure 4.6. A fully hadronically decaying $t\bar{t}$ sample with a jet selection of $p_t > 30$ GeV is used. An absolute (relative) increase

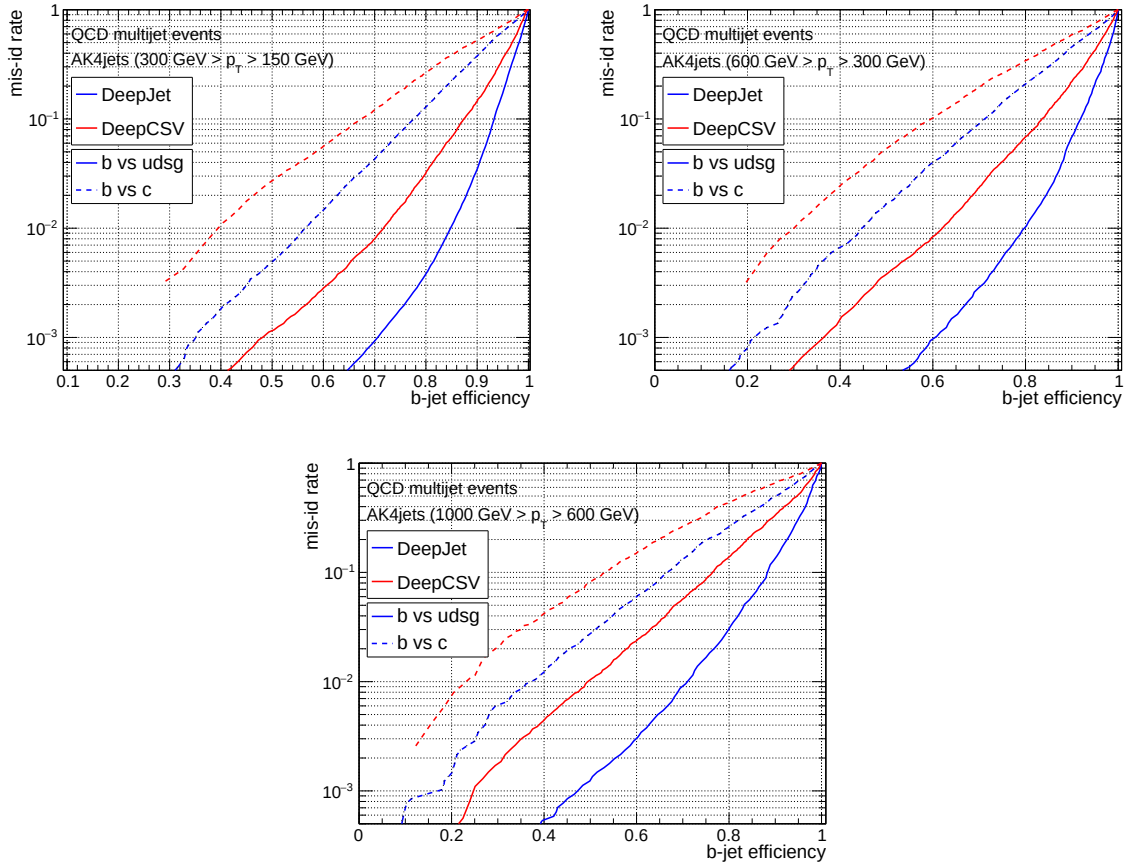


Figure 4.4.: The performance of DeepJet and DeepCSV is shown for QCD multijet events using three different jet p_t selections for both b vs. c (dashed lines), and b vs. light (solid lines).

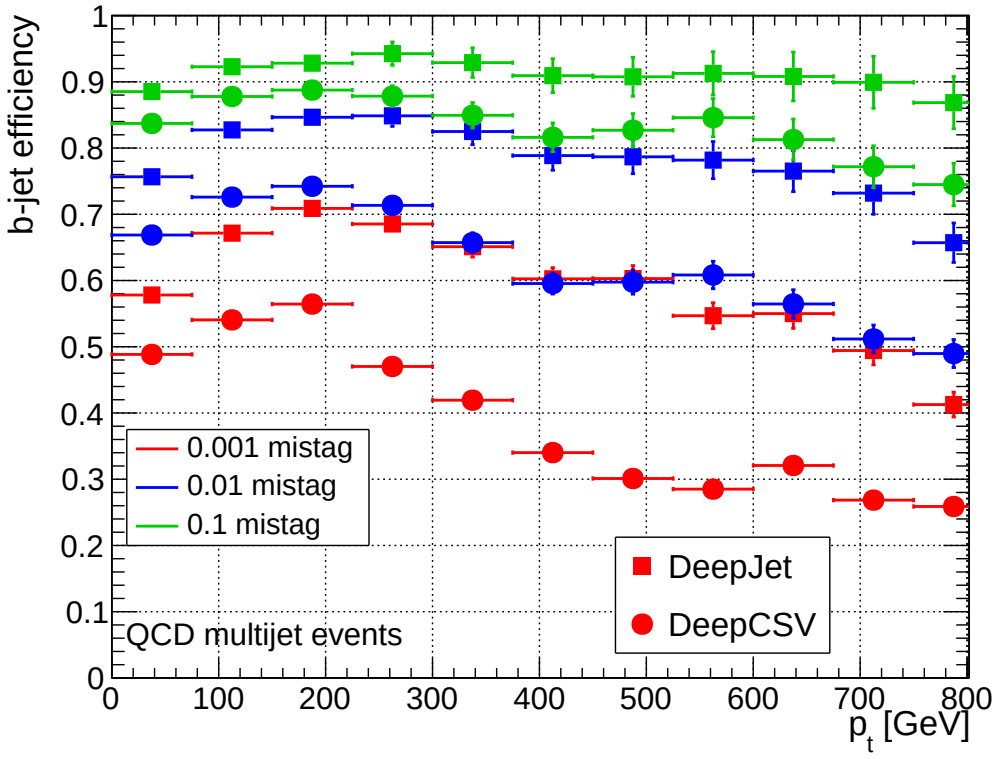


Figure 4.5.: The b jet efficiency of the DeepJet and DeepCSV algorithms are shown for three different fixed light jet mistag rates as a function of jet p_t. The physics topology used is QCD multijet events.

in performance of $\sim 6\%$ ($\sim 20\%$) at a light jet misidentification rate of 1% is observed.

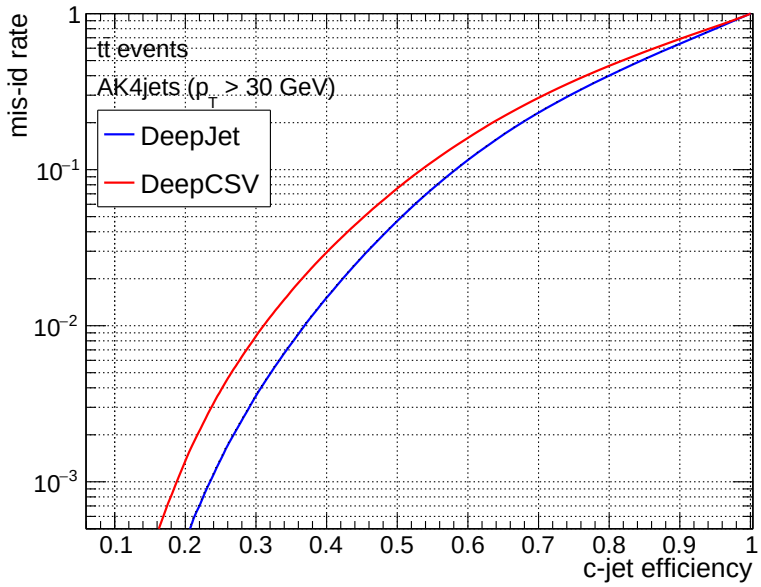


Figure 4.6.: The performance of DeepJet and DeepCSV is shown for c vs light jet identification for $t\bar{t}$ events where both top quarks are decaying hadronically. A jet selection of $p_t > 30$ GeV is applied.

The addition of quark and gluon nodes in the DeepJet network allows for separating quark and gluon jets as well. A discriminator is constructed as $\frac{P(\text{uds})}{P(\text{uds})+P(\text{g})}$. The current CMS quark gluon discriminator is a quark/gluon likelihood [134], which is built from the product of three variable probability density functions. These variables are picked based on physics knowledge of differences in quark and gluon jets, and consists of the jet constituent multiplicity, the jet minor angular opening and the jet fragmentation distribution variable. As can be seen in Figure 4.7, the DeepJet algorithm outperforms the quark gluon likelihood. An absolute (relative) improvement of $\sim 10\%$ ($\sim 20\%$) quark efficiency at 20% gluon misidentification probability is observed.

As the DeepJet algorithm uses more jet constituents with less selection requirements, it is important to verify that the performance is robust to varying degrees of pileup. In Figure 4.8 the b-tagging efficiency is shown for different light-jet misidentification rates as a function of number of reconstructed primary vertices in the event. For comparison, the DeepCSV algorithm is shown as well. When there is more pileup, the b jet identification performance decreases for both algorithms. The rate at which it

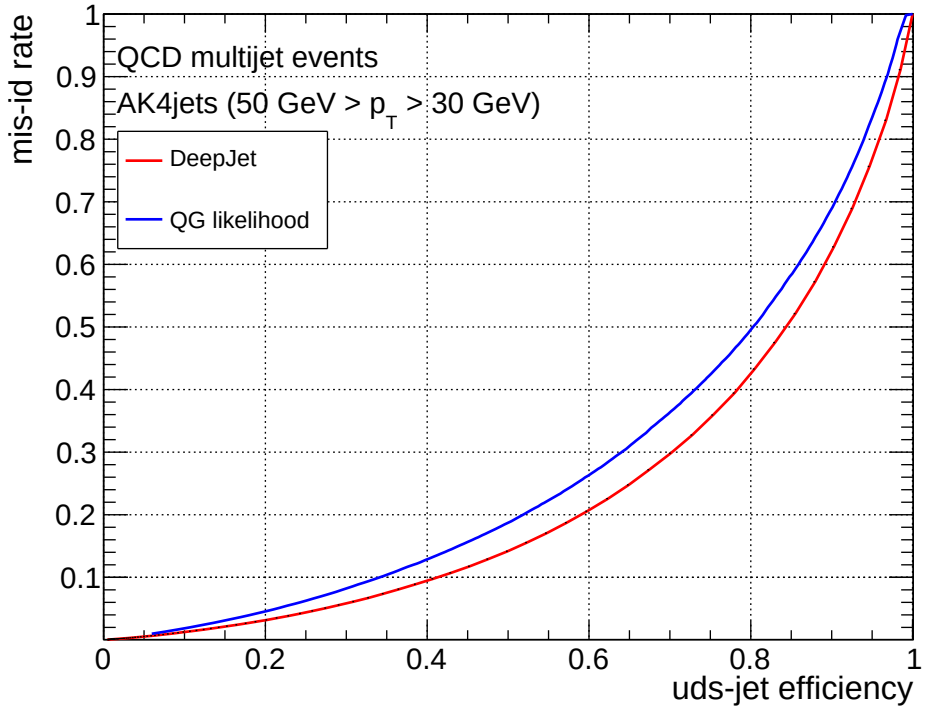


Figure 4.7.: The quark vs gluon identification performance of DeepJet is compared to the quark-gluon likelihood discriminator. A QCD multijet sample is used.

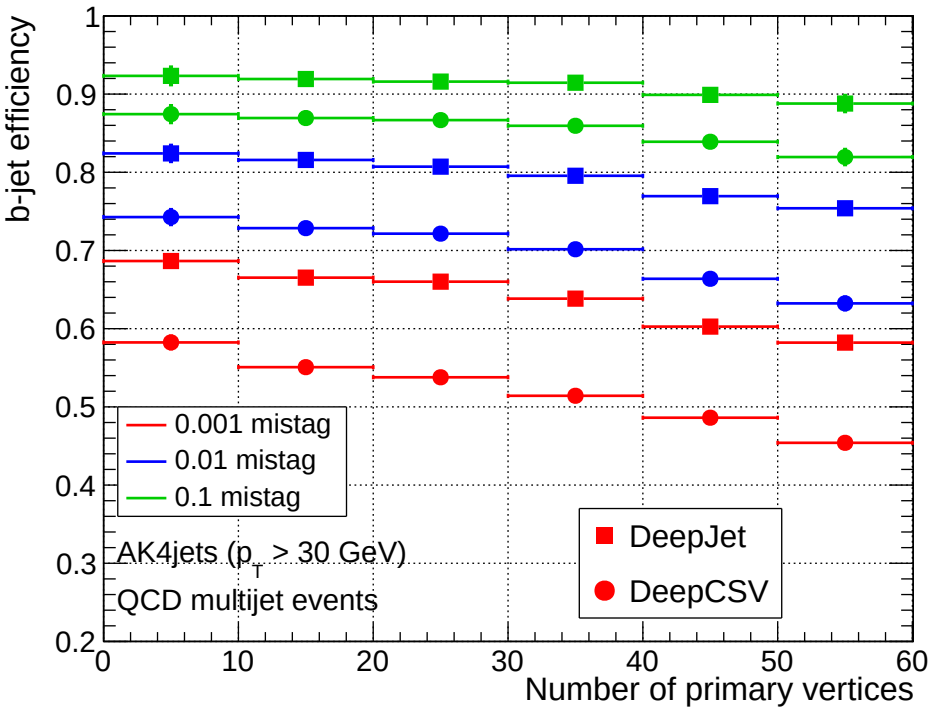


Figure 4.8.: The b jet identification performance of DeepJet and DeepCSV as a function of the number of reconstructed primary vertices in the event is shown. The b jet efficiency is shown for fixed light jet misidentification probabilities (mistag rates) of 10%, 1%, 0.1% using jets from QCD multijet events.

decreases is very comparable between DeepCSV and DeepJet, indicating that DeepJet is able to internally learn a track selection to some degree.

DeepCSV and DeepJet differ in both the network architecture and the number of input variables used. It is therefore useful to identify how each of these difference contribute to the performance improvement. To that end, some additional neural networks were trained. One model uses a fully connected architecture as DeepCSV but uses the same input as DeepJet. Similarly, a model was trained using the DeepCSV inputs, but it uses a series of convolutional layers as well as an LSTM to process the DeepCSV tracks. Their performance is shown in Figure 4.9. Comparing the perfor-

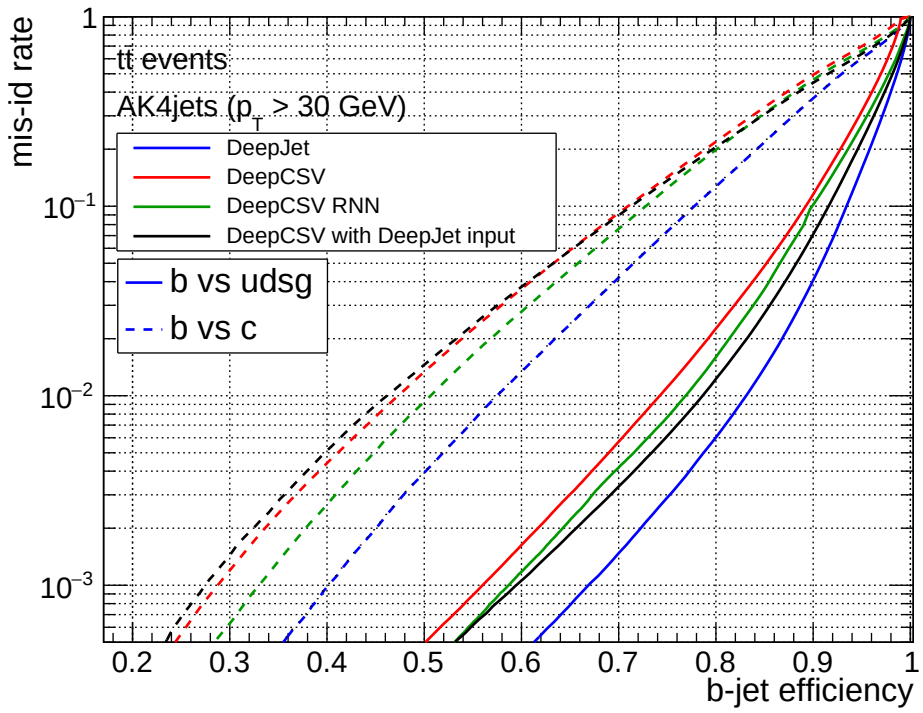


Figure 4.9.: The performance of DeepCSV and DeepJet is shown on a fully hadronic $t\bar{t}$ sample. Two models interpolating between DeepCSV and DeepJet are also shown.

mance of these models with DeepCSV and DeepJet, it can be seen that both perform better than DeepCSV, but neither reaches the performance of DeepJet. The DeepJet architecture performs a bit better than DeepCSV using the same inputs, but without removing the track selection and increasing the number of physics objects considered, this gain is very limited. Similarly, it is clear that increasing the number of inputs used has limited gain, if the neural networks architecture is not well designed to process

the large sets of input features.

The DeepJet algorithm consists of roughly 200 thousand free parameters, while the training dataset is of order $\sim$100 million jets. To understand to what extent this large dataset is useful for the neural network, the same model was trained on training datasets with different sizes. The performance of these trainings can be seen in Figure 4.10. A performance improvement similar to $\sqrt{N}$ is, as the training dataset is increased, validating the utility of the large dataset. To produce this plot separate hyperparameter tunings were not done for the dataset. If the number of training epochs were increased and the learning rate scheme was adjusted, it is possible that a slightly more performing training could be achieved with the smaller datasets.

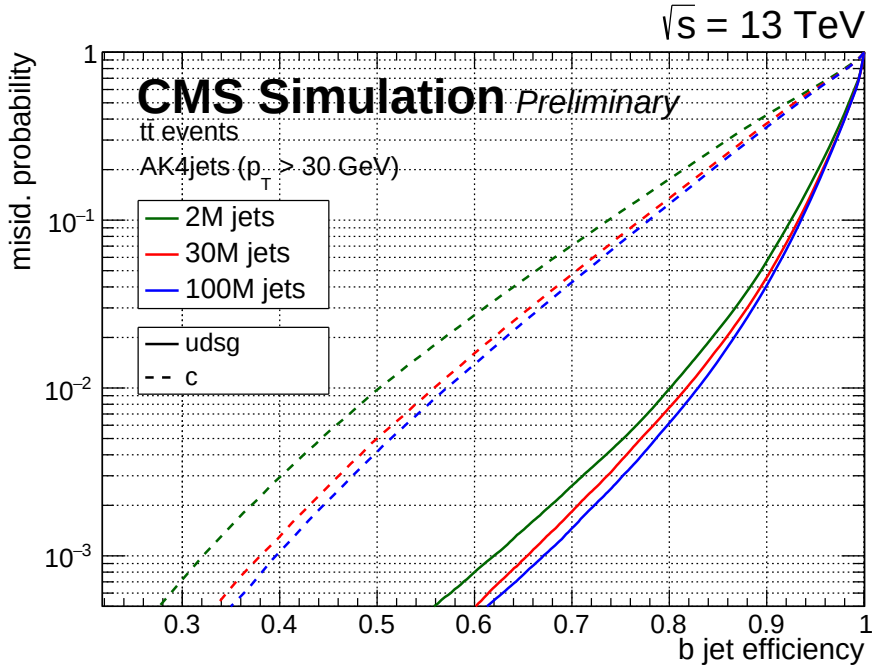


Figure 4.10.: The performance of the DeepJet algorithm trained on datasets of varying size is shown.

The DeepSet approach, described in Chapter 3, is very similar to DeepJet from a technical point of view. The DeepJet architecture becomes a DeepSet model, if the LSTM is replaced by a permutation invariant operation over the physics objects. An example of such an operation would be a scalar sum. Therefore it is interesting to compare the performance of the models to understand how important DeepSets imposing of object permutation invariance is compared to the flexibility of object aggregation

of DeepJet. If the filter sizes of the DeepJet convolutional layers are kept exactly as is, and the LSTM is just replaced with the scalar sum, the DeepSet approach is not competitive. However, this is to be expected as there simply is not enough information being passed on to the fully connected layers. A natural modification to fix this information bottleneck, is to enlarge the convolutional layers of DeepJet. Therefore the 1x1 convolutional branches are extended to contain 100 filters each, increasing the number of free parameters. The final layer of each branch is enlarged further, with the charged particle branch being set to 256 filters, whereas the vertex and the neutral particle branch are set to 128 filters. The LSTM of DeepJet is then replaced with a scalar sum over the physics objects to make a DeepSet model. For comparison a DeepJet model with this new convolutional structure while keeping the LSTM is also trained. Figure 4.11 shows the performance of the original DeepJet model, the DeepSet model and the DeepJet model with the new convolutional structure. The DeepSet model performs very well highlighting the utility of imposing the permutation invariance of objects in a jet. It does however require enlarging the convolutional layers significantly compared to DeepJet. It is also seen that the DeepJet model using the same convolutional layers outperforms DeepSet. This means that the LSTM still provides information, and it probably indicates that the DeepSet approach of using a scalar sum to aggregate objects is a bit too limiting.

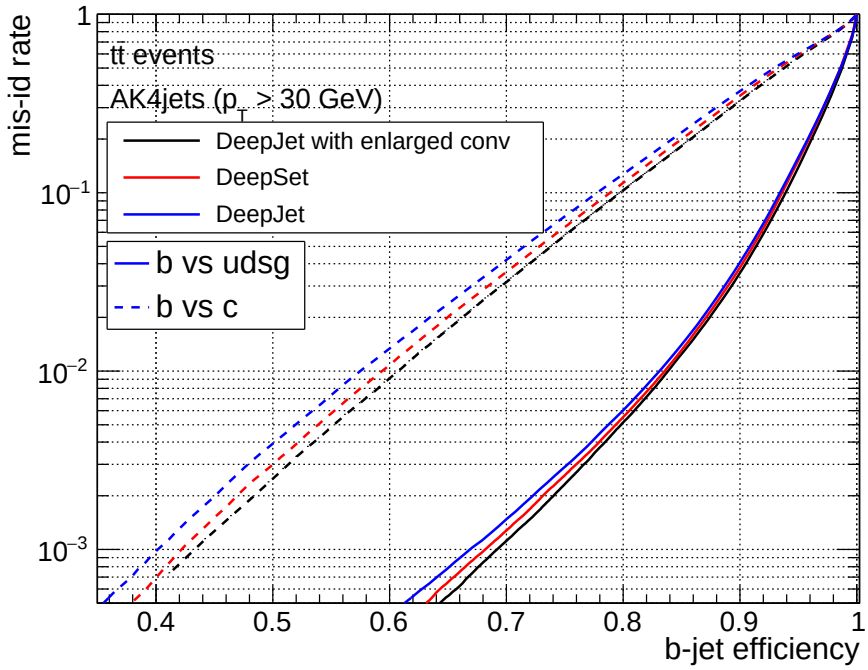


Figure 4.11.: The b jet identification performance is shown for the DeepJet architecture, a DeepSet architecture and a DeepJet model that uses a larger number of convolutional filters.

4.3. Calibration of performance

Since the DeepJet algorithm is trained on simulation, it is needed to validate the performance in real data. There have always been some discrepancies between jet flavour identification in simulation and data, even using simple taggers. In the case of heavy flavour identification at the CMS experiment, there is a significant difference between the track impact parameter distributions in data and simulation due to modelling difficulties in the CMS tracker. Since all heavy flavour identification algorithms necessarily use this information, there will be some difference in performance. The term scale factor indicates the efficiency in data divided by the efficiency in simulation to identify a specific jet flavour. A scale factor of 1 therefore means similar performance in data and simulation, below 1 means lower efficiency in data than simulation and larger than 1 means higher efficiency in data than simulation. However, the performance of a jet identification algorithm cannot be quantified solely by the efficiency. The misidentification rate can also change, and it will need to be measured as well in data. The heavy flavour identification performance changes significantly as a function of p_t and η and accordingly the scale factors are measured in bins of these jet properties.

It is important to note that the quark-gluon discrimination performance of DeepJet has not been measured in data yet. As the difference between quark and gluon jets are theoretically hard to model, it might be that the performance gain compared to the simple quark-gluon likelihood method is smaller in data.

4.3.1. Identification of b-jets

In CMS, b-jet scale factors are calculated for fixed discriminator working points. These working points are defined as the discriminator cut that on a QCD multijet sample gives a light-jet misidentification rate of 10%, 1% or 0.1%. They are referred to as the Loose (L), Medium (M) and Tight (T) working point respectively. Several different methods exist to measure b-jet scale factors described in reference [117], but a brief overview is made here. In general, the scale factor methods require a selection that is enriched in b jets, and they either use $t\bar{t}$ events or QCD multijet events. The $t\bar{t}$ events are naturally enriched in b-jets due to the top quark decay. One of the top quark methods, referred to as the "kin" method, uses events from the dileptonic $t\bar{t}$ decay channel. In this decay channel it is expected that only b jets are produced,

although pileup jets and jets from ISR and FSR can still occur. In order to remove background from these additional jets, a BDT is trained to identify events with only b jets. In order not to bias the scale factor measurement this BDT only use event level kinematic variables. When the b-jet enriched selection has been obtained, the rates in data and simulation are compared to obtain the scale factors. Another top quark pair method is the tag and probe method, based on the semileptonic $t\bar{t}$ decay channel. Here, one of the b jets from either the hadronic or leptonic top quark decay is tagged, and then the other b jet in the event is used as probe to measure the b-tagging efficiency. The methods using QCD multijet events are not naturally enriched in b-jets. Jets are selected from this sample by requiring a soft (low p_t) muon inside the jet. This targets the b hadrons that are decaying leptonically and enriches the sample in b-jets. To avoid biasing these methods the DeepJet algorithm does not use information about leptons in the jets, which otherwise could marginally improve the performance. Two of the QCD multijet scale factor methods rely on constructing templates from simulation of variables that are sensitive to the b jets and then fitting these distributions to data. The lifetime secondary vertex method uses a 2D template of the probability that the jet was produced from the primary vertex and the mass of the secondary vertex. The p_t-rel method uses the p_t of the muon relative to the jet axis.

Since the DeepJet algorithm uses significantly more input variables than DeepCSV, a worry was that the performance gain in data would be smaller than in simulation. Therefore, it is interesting to compare the scale factors of DeepCSV and DeepJet in order to see if this really is the case. In Figure 4.12, the CMS 2018 combined b-jet scale factors are shown for DeepCSV and DeepJet. Since the scale factors depend on the p_t and η, this inclusive scale factor value depend on the sample. A fully hadronically decaying $t\bar{t}$ sample is used with a minimal jet p_t selection of 30 GeV. As can be seen there is essentially no difference between scale factors from DeepCSV and DeepJet, implying that despite the larger set of inputs and the more extensive architecture, the DeepJet model is robust to simulation to data differences. However the performance in data cannot be fully quantified without the misidentification scale factor.

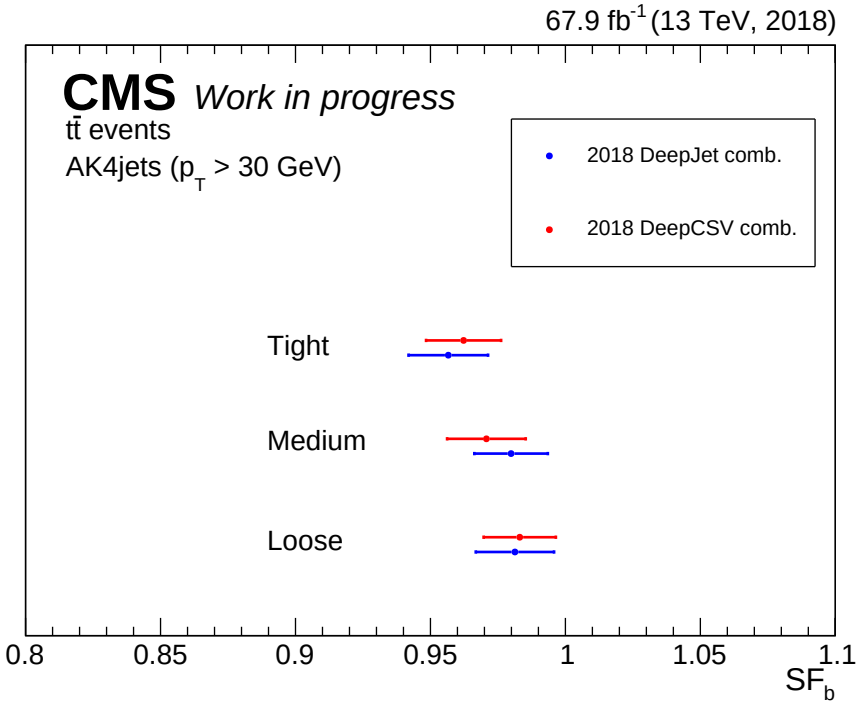


Figure 4.12.: The measured scale factors for b jets. This inclusive scale factor value (SF_b) is evaluated using the p_t and η distribution from a fully hadronically decaying $t\bar{t}$ sample.

4.3.2. Negative Tagger for light jet identification

In order to measure the light jet scale factors for DeepJet, what is referred to as a negative tagger must be constructed. The main reason why light jets are identified as b-jets, are due to the light jet tracks being reconstructed as displaced, when they really originated from the primary vertex. The track impact parameter resolution of the CMS detector is roughly symmetrical around zero as can be seen from Figure 4.13.

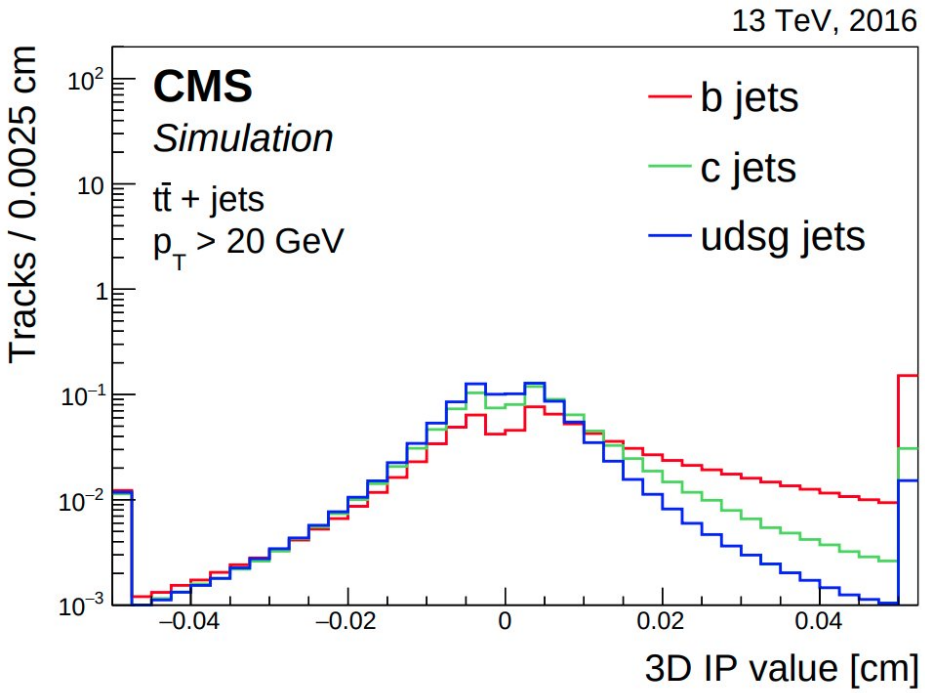


Figure 4.13.: The 3D impact parameter value distribution for b jets, c jets and light jets. From Ref. [117].

Since this means that the impact parameter distribution of tracks originating from the primary vertex is invariant under sign inversion, the light-jet discriminator distribution should also be mostly invariant under this operation. Of course, the b-jets track impact parameter distribution have mostly positive track impact parameters. If the sign inversion is performed, it causes a large fraction of b-jets to obtain very low discriminator values. This allows measuring the light-jet misidentification rate, as the light-jet distribution of the discriminator is mostly unchanged, while the b-jet distribution is shifted to low values. A negative tagger is then a variation of the b-tagging algorithm where all impact parameters are sign inverted before being input to the network. In general, the negative tagger method has had increasing uncertainties

on the measured scale factors as the b-taggers started to use more input information beyond track impact parameters. Still, it remains the only method for estimating light-jet scale factors in CMS to date.

Different designs where explored for the DeepJet negative tagger, as it was unclear what is the best way to treat the additional inputs. The method inverts all the charged particle flow candidates track impact parameters. In addition, charged particle flow candidates with a track impact parameter significance larger than 10 are removed from the inputs. This reduced the b-jet contamination in the negative tagger, presumably as the network was using kinematic variables to identify these tracks originating from b jets. In order to process the vertices, it was needed to define a custom secondary vertexing sequence. Similarly to the IVF vertexing algorithm, the secondary vertices are clustered using the full event track collection. However, the vertices are associated to a jet by requiring the secondary vertex momentum to be within $\Delta R < 0.4$ of the inverted jet direction. A vertex therefore has to be in a cone opposite to the jet direction. Furthermore, the vertex is only saved if the direction of the sum of the vertex tracks has an angle of $\cos(\theta) < -0.95$ with the secondary vertex flight direction. This selection ensures that vertices with negative flight distances are included in the inputs.

Figure 4.14 shows the discriminator distribution of the DeepJet tagger and the DeepJet negative tagger on a $t\bar{t}$ sample. The distribution from -1 to 0 is the negative tagger, whereas the normal tagger is shown from 0 to 1. The light-jet distribution is fairly symmetrical, however the b-jet distribution is significantly changed, with the peak at 1 being completely removed in the negative tag distribution.

The negative tagger is expected to be less symmetrical for DeepJet compared to DeepCSV due to the increase of inputs, yielding slightly larger uncertainties. The Figure 4.15 shows the CMS 2018 udsg-jet scale factors for DeepCSV and DeepJet. As can be seen, the nominal values are very similar for the two taggers. Together with the observations made from the b-jet scale factors, this shows that the performance degradation of DeepJet when using data is comparable to DeepCSV.

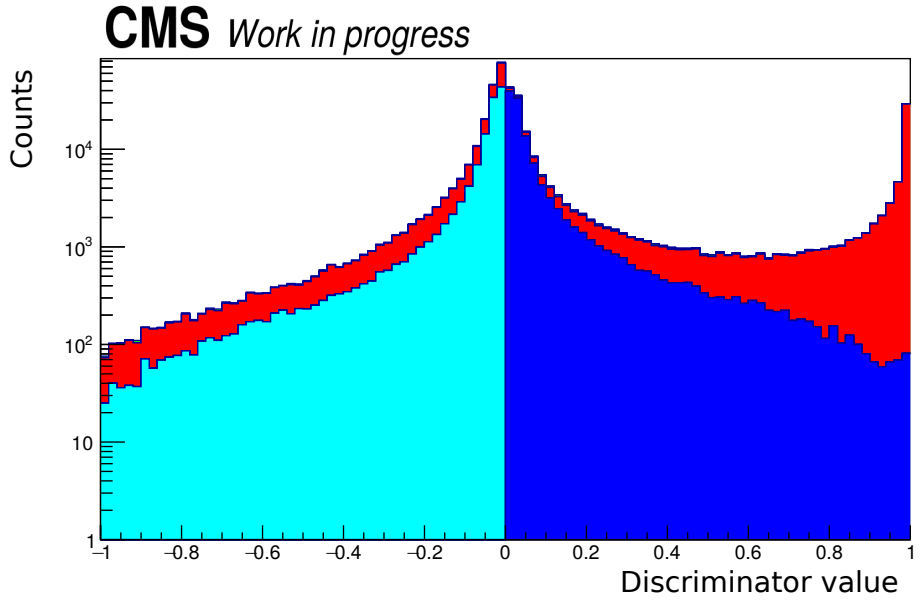


Figure 4.14.: The distribution of the DeepJet negative tagger evaluated on simulated $t\bar{t}$ events. The negative tagger discriminator distribution is plotted from -1 to 0, whereas the regular tagger is plotted from 0 to 1. The dark and light blue histograms refer to light jets, the red histogram indicates b-jets, and the barely visible green histogram indicates charm jets.

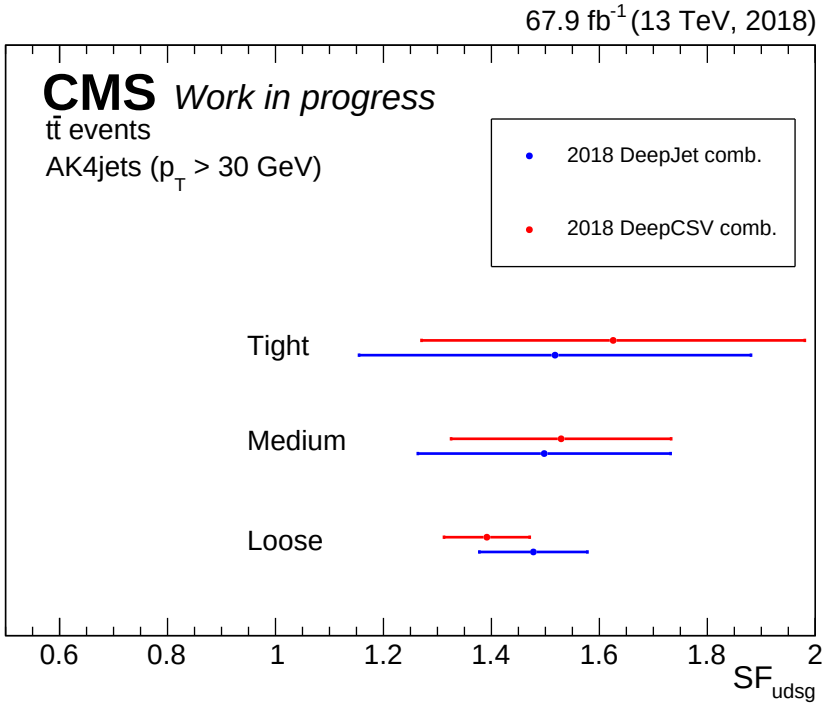


Figure 4.15.: The measured scale factors for udsg-jets (SF_{udsg}). This inclusive scale factor value is evaluated using the p_t and η distribution from a fully hadronically decaying $t\bar{t}$ sample.

4.3.3. Identification of c-jets

Scale factor measurements are done for c-jets as well [117]. In the case of a misidentification source for the b-tagger, simply the b-jet scale factors with inflated uncertainties are used, as c-jets rarely are a major background source for b-jets in physics analyses. However, for analysis dedicated to c-jet identification scale factors are needed for both c vs. b tagging and c vs. udsg tagging, as many analyses using c jets have both as background. Fixed working points are defined by a simultaneous cut on both the c vs. udsg and the c vs. b discriminator. The scale factors for these can be derived in using the $W + c$ channel as well as in semileptonic $t\bar{t}$ events where the hadronic W decays to a c quark and a down-type quark. Methods for deriving scale factors for the full c-tagging discriminator distributions have also recently been developed in reference [135]. Comparing DeepCSV and DeepJet for c-tagging, similar conclusions are reached as for b-tagging.

4.4. Physics Analyses using DeepJet

Since DeepJet is the current best performing heavy flavour jet identification algorithm in CMS, it is a useful tool for CMS physics analyses. Several published CMS analyses use DeepJet, and many more are in the pipeline. For many top quark analyses usually one or two b tags are imposed, so the improvement from DeepCSV increases the selection efficiency by $\sim$10%. Some analyses in the Higgs sector that particularly benefit from DeepJet are highlighted.

One recent analysis using the DeepJet algorithm is a search of Higgs boson pair produced in the $HH \to b\bar{b}b\bar{b}$ decay channel at the CMS experiment [136]. Since the standard model Higgs boson pair production cross section is small and the main background of QCD multijets are strongly suppressed by requiring b-tagged jets, a high efficiency for heavy flavour identification is important. The analysis signal region is defined using 4 jets that pass the DeepJet b-tagging medium working point. The b-tagging efficiency is therefore $\sim$47% whereas with DeepCSV it would roughly have been $\sim$30%, yielding a rough estimate of an increased signal to background ratio of $\sim$1.6. This along with many other significant improvements and a larger data sample compared to older analyses, enabled this search to put the most stringent limits on the Higgs boson pair production cross section to date.

Another recent CMS analysis [137] is a direct search of $H \rightarrow c\bar{c}$ in the VH channel where V is either a Z or W Boson decaying leptonically. The result is a combination of a merged jet and a resolved jet analysis. The resolved jet analysis relies on using two charm tagged AK4 jets and it improves on the previous CMS analysis [138] by using the DeepJet charm identification capabilities instead of DeepCSV, which improves the signal rate. This analysis sets the most stringent limit on the charm Yukawa coupling modifier κ_c to date.

4.5. Conclusion

In this chapter a multiclass flavour tagging algorithm, DeepJet, was presented. It exploits the full jet information, and it relies on low-level variables with loose selection criteria. The architecture is designed such that it can process these inputs in a highly efficient manner. Compared to the previous CMS heavy flavour identification algorithm of DeepCSV, a significant gain in performance is seen for flavour tagging. For some high jet p_t topologies the efficiency is improved by a factor 2 for the same misidentification rate. Calibration of the tagger reveals that the performance gain seen in simulation, is applicable to data as well for heavy flavour identification. The model can also be used for quark-gluon discrimination, achieving state of the art performance.

Chapter 5.

Deep Neural Network for Jet Energy Regression

Machine learning has been employed in many avenues of jet physics, for instance in terms of jet flavour identification described in the previous chapter. One of the areas where machine learning is not the primary technique is the development of methods to estimate jet energy corrections (JEC). These corrections try to mitigate the difference in the reconstructed jet p_t as observed in the detector and the particle level jet p_t ($p_{t,\text{ptcl}}$) as given to the detector. The current CMS jet energy corrections [72], are derived using a pileup correction based on the jet area, the diffuse offset energy density, jet p_t and η, called the level 1 jet energy correction. A second correction on the jet energy response is also made based on the jet p_t and η called level 2 and level 3 jet energy corrections. This method has performed very well even compared to multivariate methods, which has made it favorable not to adopt more complicated approaches. However this method of correcting the jet energy does not take into account the jet flavour. As jets with different flavours have different kinematic jet properties, different jet multiplicities and different distributions of energy shared among the jet constituents, they also have different jet energy responses. This means that there is a clear connection between jet flavour identification and the task of estimating energy corrections. A neural network that is able to perform jet flavour identification could make a guess of the jet flavour and then propose a suitable energy correction. There has been significant improvements in jet resolution and response when applying b-jet energy corrections using neural networks in the past. In reference [139] a neural network was trained to predict an additional jet correction to b jets improving their resolution in $H \to b\bar{b}$. A similar approach to this analysis is developed in this chapter.

The main differences are to train a network to apply a correction to all jet flavours, and to apply an architecture similar to the DeepJet architecture, allowing the network to estimate the jet energy correction without prior knowledge or assumptions on the jet flavour.

One core reason for investigating this neural network based estimation of jet energy corrections is to understand, if it is possible to decrease the flavour modelling uncertainty on the jet energy corrections [72]. The simulated jet response for the different jet flavours depends on which matrix element generator, parton showering model and hadronization model is used. Herwig++ [140] and Pythia8 [39] are two different general purpose event generators that can be used for this, and they each result in different corrections. For instance, Herwig++ heavy flavour jets have a significantly larger jet energy response at high p_t compared to those simulated by Pythia8. This is due to the jets having different charged and neutral particle compositions and jet shapes. A jet correction only targeting p_t and η cannot take into account such effects on the jet response, and it would most likely make them sensitive to the difference between Herwig++ and Pythia8. However, a multivariate correction using the full jet shape, might be less sensitive, since the detector is the same, and the jet response should be almost fully describable by physical observables. This flavour dependent jet energy correction uncertainty is the largest for several top quark analyses, such as the top quark mass measurement.

5.1. Neural Network model

The DeepJet JEC model has a lot of similarities with DeepJet in terms of network architecture, jet flavour labeling, input features and preprocessing. The same training sample is used as for the DeepJet algorithm. The target variable is defined using the particle level jet transverse momentum $p_{t,\text{ptcl}}$, which is determined by matching generator jets clustered using the anti-k_t algorithm to the measured jets within $\Delta R < 0.4$. The training label is then $p_{t,\text{ptcl}}/p_t$. Two target variables are constructed using the transverse momentum of generator jets where neutrinos are included and excluded respectively. The standard CMS jet definition excludes neutrinos, but the neural network can potentially correct for neutrinos in leptonically decaying b or c jets, which would improve the jet energy resolution. The network architecture is identical to the enlarged DeepJet model from the previous chapter, i.e. the 1x1 convolutional branches

contain 100 filters each, as well as enlarging the final convolutional layer of each branch to 256 filters for charged particles, whereas the vertex and the neutral branch are set to 128 filters. The output layer is also modified to contain 2 nodes with linear activation functions. The training approach of reference [139] is emulated by using the Huber loss function, making the predicted corrections less sensitive to outliers compared to the use of the mean squared error. The input variables are identical to DeepJet with a few additions. The azimuthal angle ϕ of the particle candidates and jets are included as an input. The charges of particle flow candidates are included, and a flag is added to indicate if a charged particle flow candidate is a muon or electron. This was not included in flavour tagging to avoid a calibration bias, however since the network should correct for potential neutrinos, the lepton information is essential for this task. The jet energy and mass are included as well. The full set of input variables can be found in Appendix A.

5.2. Performance of the Neural Network

The performance of the DeepJet JEC network is compared to the standard jet energy corrections, which include a level 1, level 2 and level 3 correction. The DeepJet algorithm has the advantage of being able to perform a jet energy correction leveraging the information of the jet flavour. Beyond this, the network can also make a correction leveraging more information than just the jet p_t and η values. The standard JEC uses a jet definition excluding neutrinos. Here, the jet definition includes neutrinos. The network also includes a second output node, predicting the jet energy without neutrinos. The performance is similar for light quark jets and gluon jets, however including neutrinos yields better performance for b and c jets. Only the corrections with neutrinos are therefore shown here.

The ability to correct the jet energy is compared on a sample of jets gathered from simulated Pythia8 QCD multijet events. The main metric used to quantify the performance is the mean jet energy response, which is defined as the mean of the ratio of the measured jet p_t to the particle level jet p_t, $p_t / p_{t,\text{ptcl}}$. A comparison of the mean jet energy response is shown in Figure 5.1 for the four different jet flavours using the same flavour truth labeling as in Chapter 4. In the case of each jet flavour the mean jet energy response is closer to 1 when applying the DeepJet JEC method. The main improvement is below $p_t < 100$ GeV for light quark jets and gluon jets. The entire

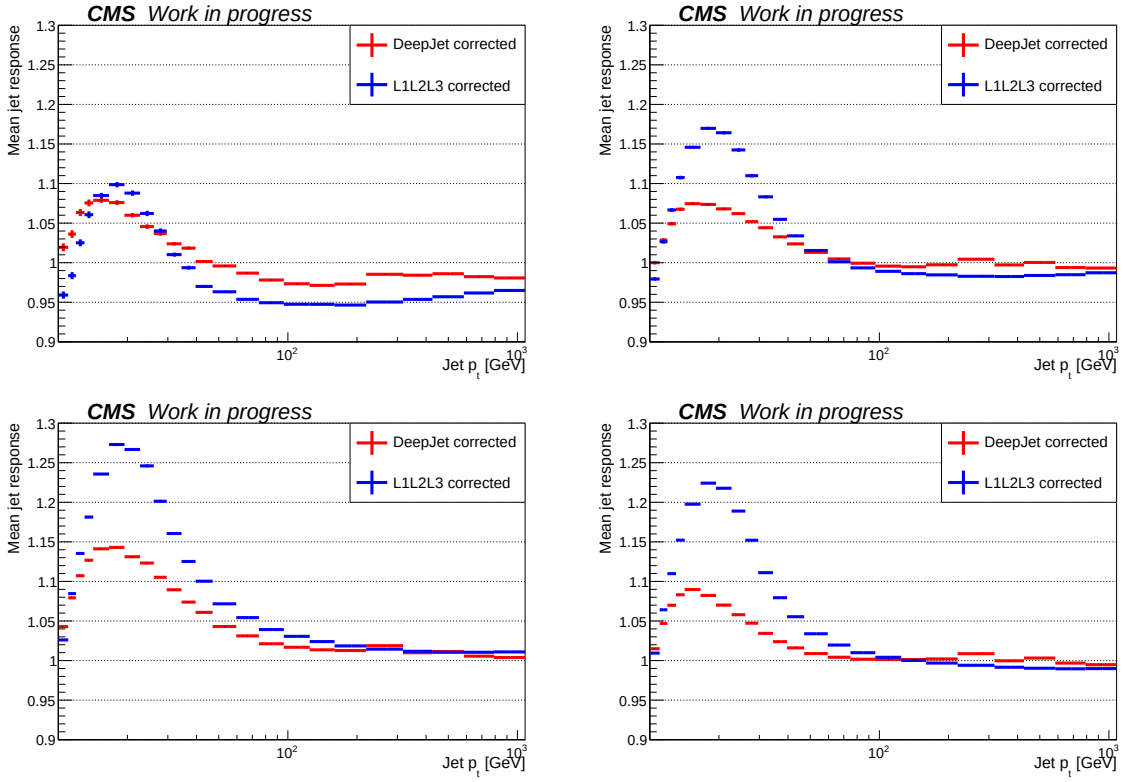


Figure 5.1.: The mean jet response of the DeepJet corrected jets is compared with the jets that are corrected with standard L1L2L3 jet corrections. Shown are b-jets (top left), c jets (top right), light quark jets (bottom left) and gluon jets (bottom right).

spectrum improves for heavy flavour jets since the neutrino contribution is added.

The improvements in jet resolutions were studied in the context of a top quark reconstruction analysis. Jets from $t\bar{t}$ events where one W boson decays leptonically and one W boson decays hadronically are used. The partons are matched to the jets by requiring $\Delta R < 0.3$. Only events where an unambiguous matching exists are used. The resolution is estimated as the σ of a Gaussian fitted to the $p_t - p_{t,\text{parton}}$ distribution in bins of η and p_t. Figure 5.2 shows the resolution of b-jets in different bins of η. The DeepJet corrected jets are compared to the L1L2L3 corrected jets. For all the different η regions, a $15-20\%$ improvement in resolution is observed below 100 GeV. Above this the resolution improvement becomes smaller, perhaps indicating that a larger training statistics is required in the high p_t region. Shown in Figure 5.3 is the resolution of the

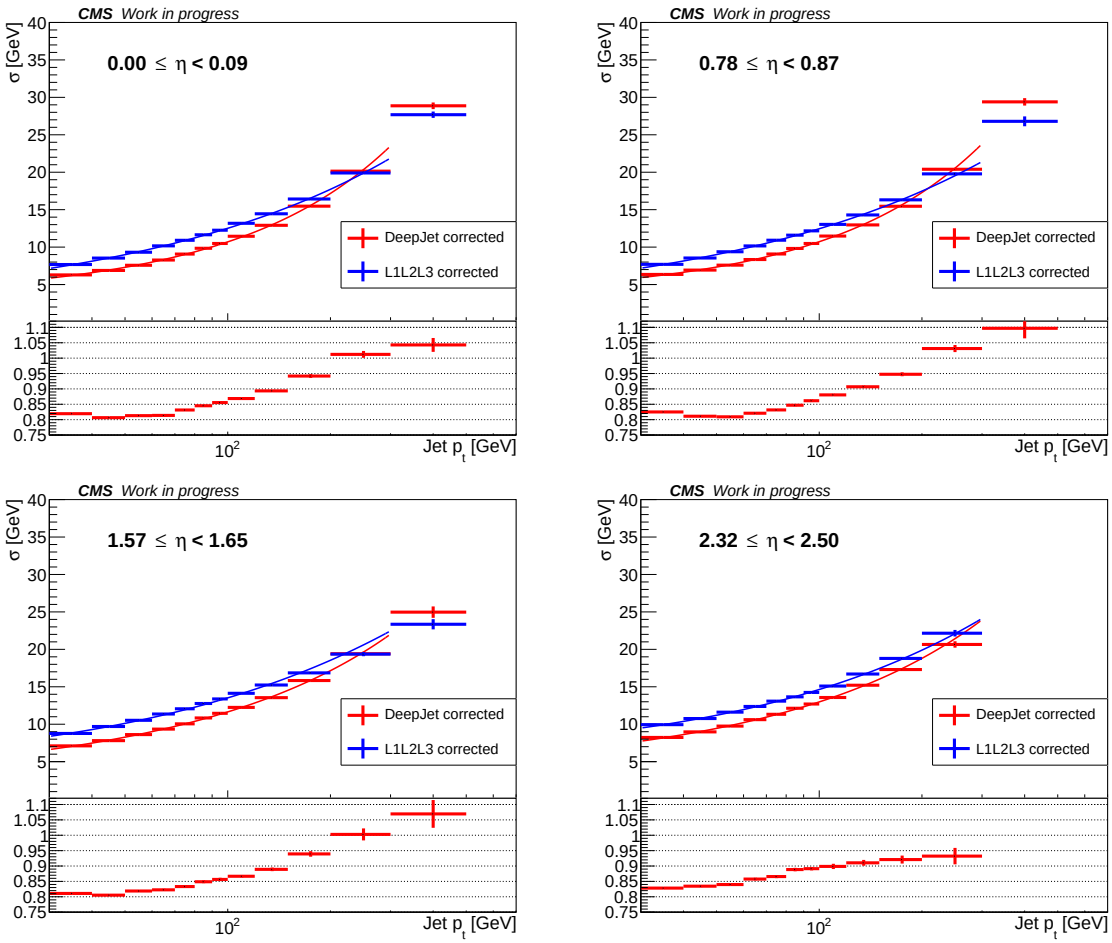


Figure 5.2.: The σ of a Gaussian fit to the $p_t - p_{t,\text{parton}}$ distribution is plotted as a function of the p_t of the reconstructed jet for b-jets from a top quark decay in $t\bar{t}$ events. Four different η bins are shown.

udsc jets from the W boson decay. A $\sim$15% improvement in the resolution is observed. This indicates that the better resolution does not solely come from including neutrinos in the jet energy estimation. The improvement in the jet energy reconstruction can

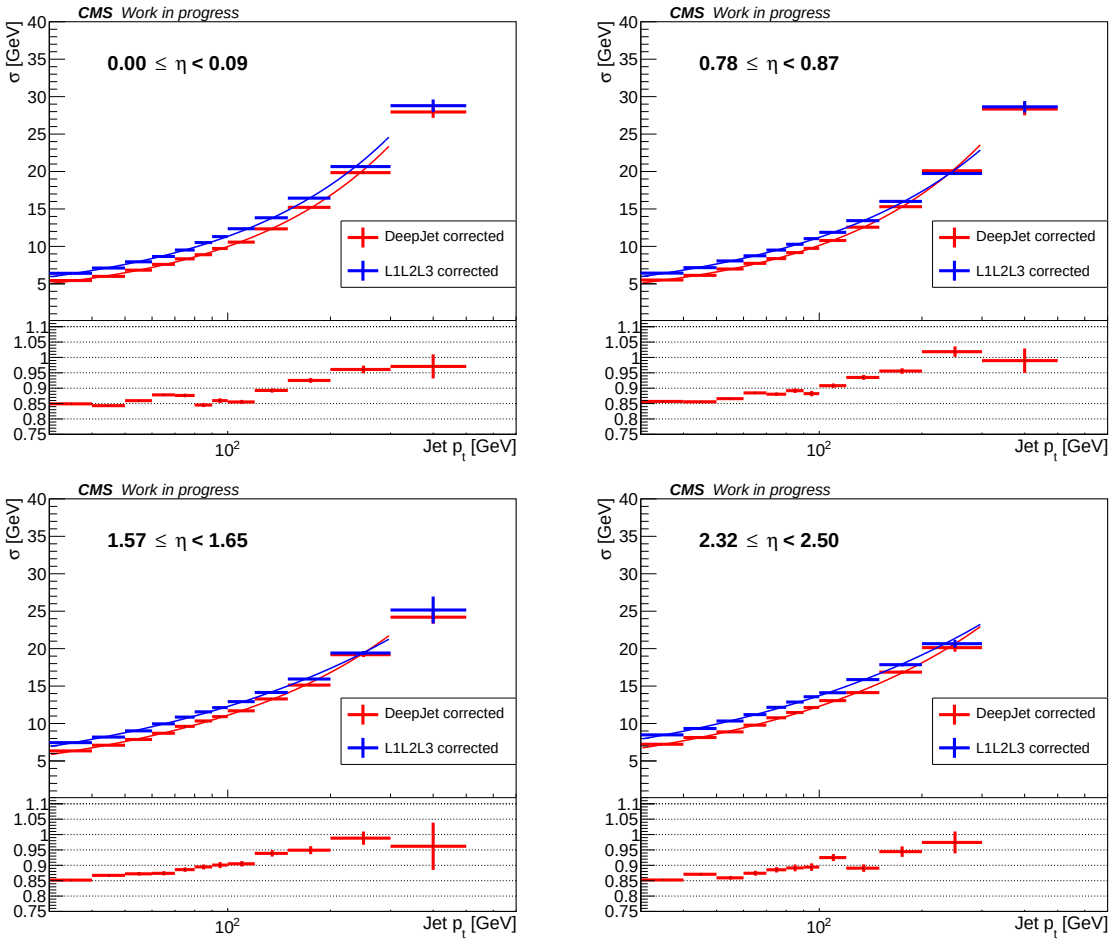


Figure 5.3.: The σ of a Gaussian fit to the $p_t - p_{t,\text{parton}}$ distribution is plotted as a function of the p_t of the reconstructed jet for udsc-jets from W boson decay in $t\bar{t}$ events. Four different η bins are shown.

also be quantified by examining the reconstructed physics objects. As an example, the $W \to q\bar{q}$ decay is studied in $t\bar{t}$ events. The jets associated with the W boson are identified via parton matching and the invariant mass of the jets is calculated. Figure 5.4 shows the obtained mass spectrums with and without DeepJet JEC applied. An improvement of around $\sim$15% is obtained in the width of the distribution. The mean of the distribution also moves closer to the generator value of W Boson mass of 80.4 GeV.

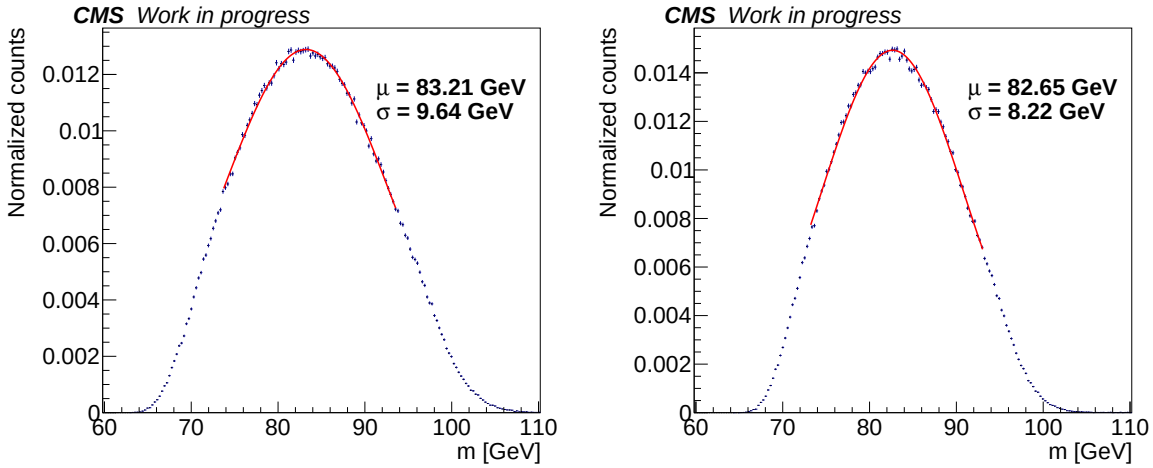


Figure 5.4.: The invariant mass of two jets produced from a $W \to q\bar{q}$ decay in top quark pair events. Left is the mass spectrum without DeepJet corrections applied to the jets, and right is the mass spectrum with the corrections applied. Gaussian fits are applied around the peaks to estimate the parameters of the distributions.

5.3. Jet Energy Correction Flavour uncertainties

The jet energy response depends on the jet flavour. The jet energy corrections derived for each flavour would vary depending on whether Herwig++ or Pythia8 is used as the general purpose event generator, due to differences in how the jet flavours are modelled. The difference between these two event generators can be used to assign an uncertainty on the jet energy flavour response. In Figure 5.5, the mean jet energy response as a function of jet p_t is shown for Herwig++ and Pythia8 for jets with the DeepJet JEC applied in QCD multijet events. For comparison, Figure 5.6 shows the mean jet energy response as function of jet p_t with simply the L1L2L3 corrections applied. Figure 5.7 shows the ratios between Herwig++ and Pythia8 for both methods. As can be seen from the ratios, there is no significant improvement in the agreement between the generators after applying the multivariate correction. It was also tried to train the network on a mix of QCD Herwig++ and Pythia8 jets, but this did not significantly improve the agreement. The lack of improvement seems to imply that the difference in the mean jet response is caused by complicated effects that are hard for the network to learn. Possibly a more advanced neural network is required to improve on this. In order to estimate the DeepJet JEC flavour uncertainty, a JEC set is derived for each flavour using both generators. This is done in 5 bins of η of the reconstructed jet. A ratio of Herwig++ versus Pythia8 is taken between these JECs and subsequently polynomials are fit to parameterize the tendency in p_t. The obtained function of these

ratios can then be used to apply an additional energy scale correction to the jets in p_t and η for uncertainty estimation. Figure 5.8 shows the uncertainties for each flavour in the central η bin. The plots for the remaining η bins can be found in Appendix C.

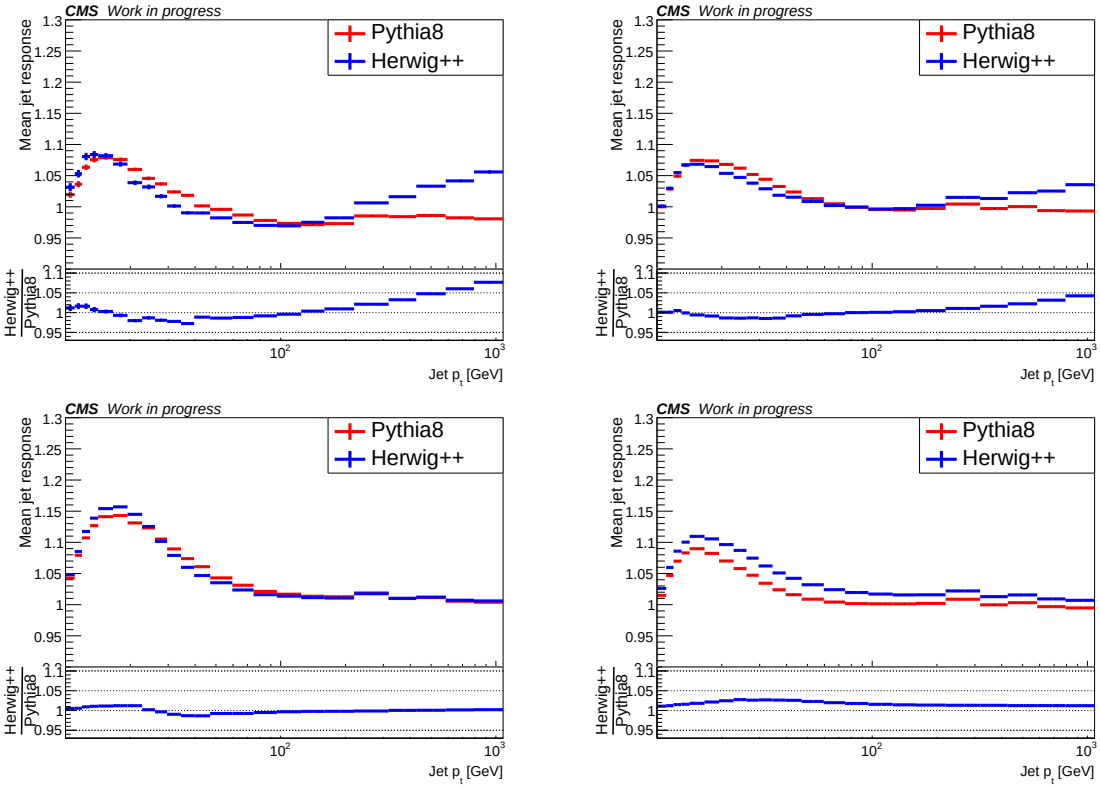


Figure 5.5.: The mean jet response of the DeepJet corrected jets are compared on a QCD multijet sample generated using Pythia8 and Herwig++. Shown are b-jets (top left), c jets (top right), light jets (bottom left) and gluon jets (bottom right).

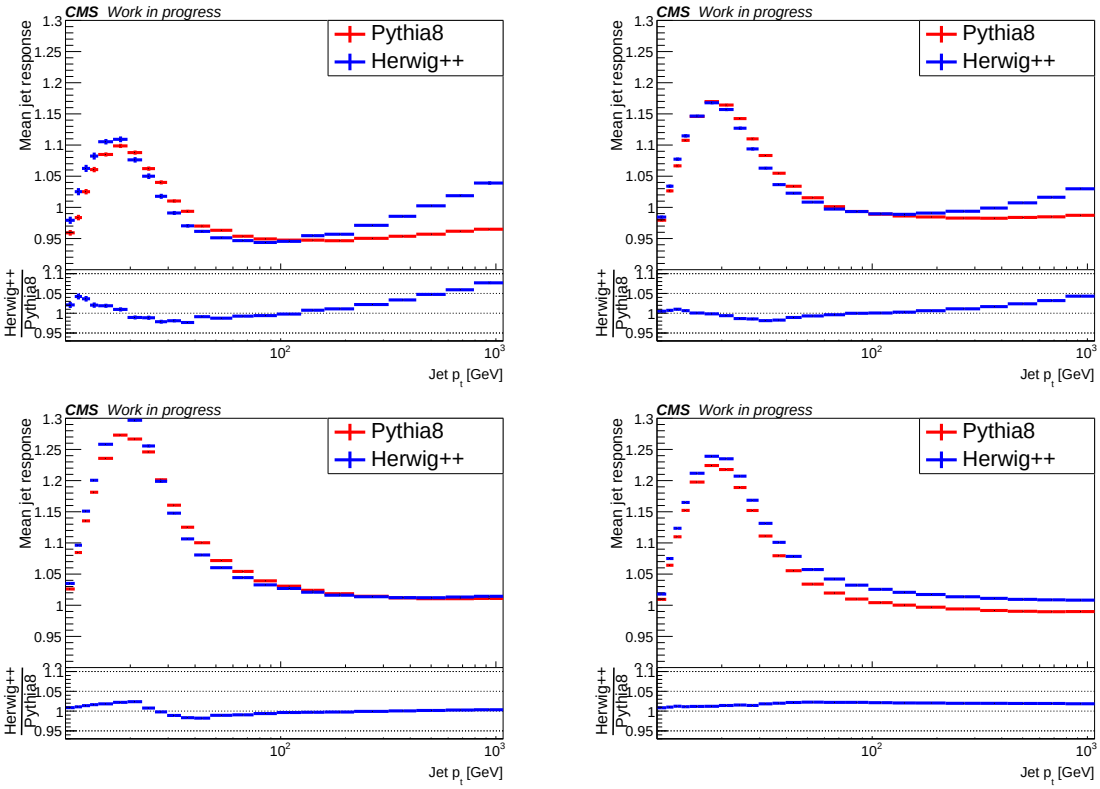


Figure 5.6.: The mean jet response of the L1L2L3 corrected jets are compared on a QCD multijet sample generated using Pythia8 and Herwig++. Shown are b jets (top left), c jets (top right), light quark jets (bottom left) and gluon jets (bottom right).

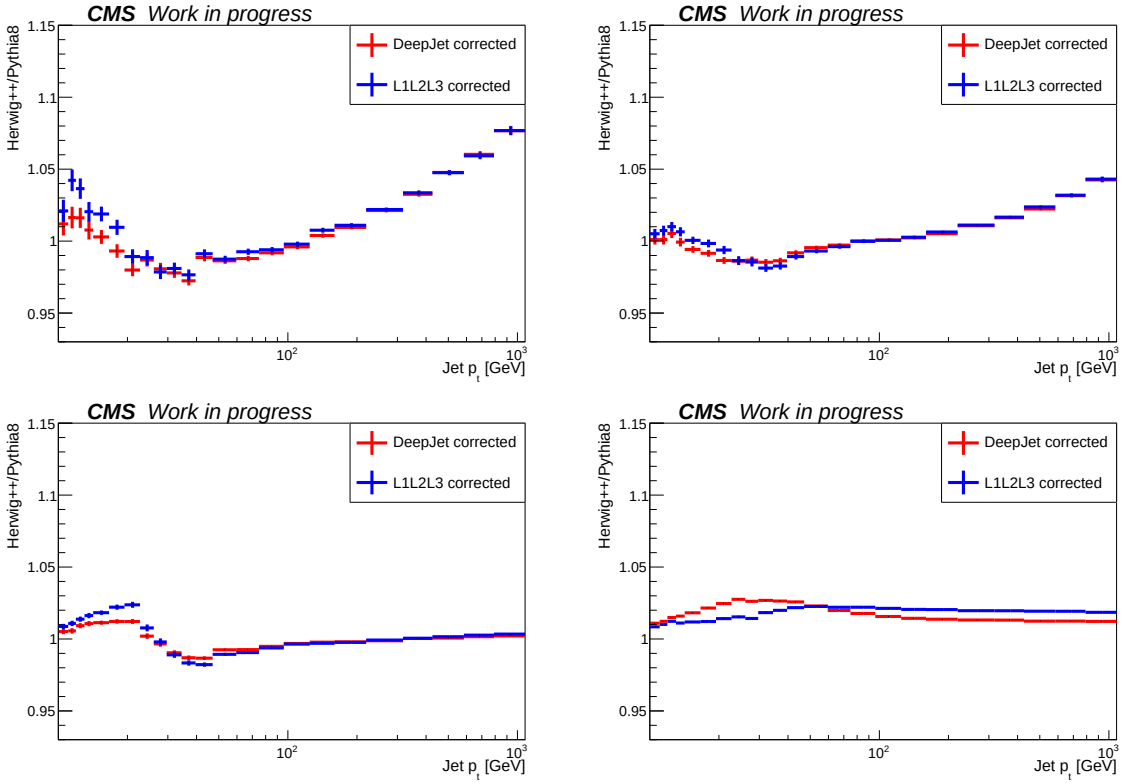


Figure 5.7.: The ratio of the Herwig++ and Pythia8 mean jet response for DeepJet corrected and L1L2L3 corrected jets are compared. Shown are b jets (top left), c jets (top right), light quark jets (bottom left) and gluon jets (bottom right).

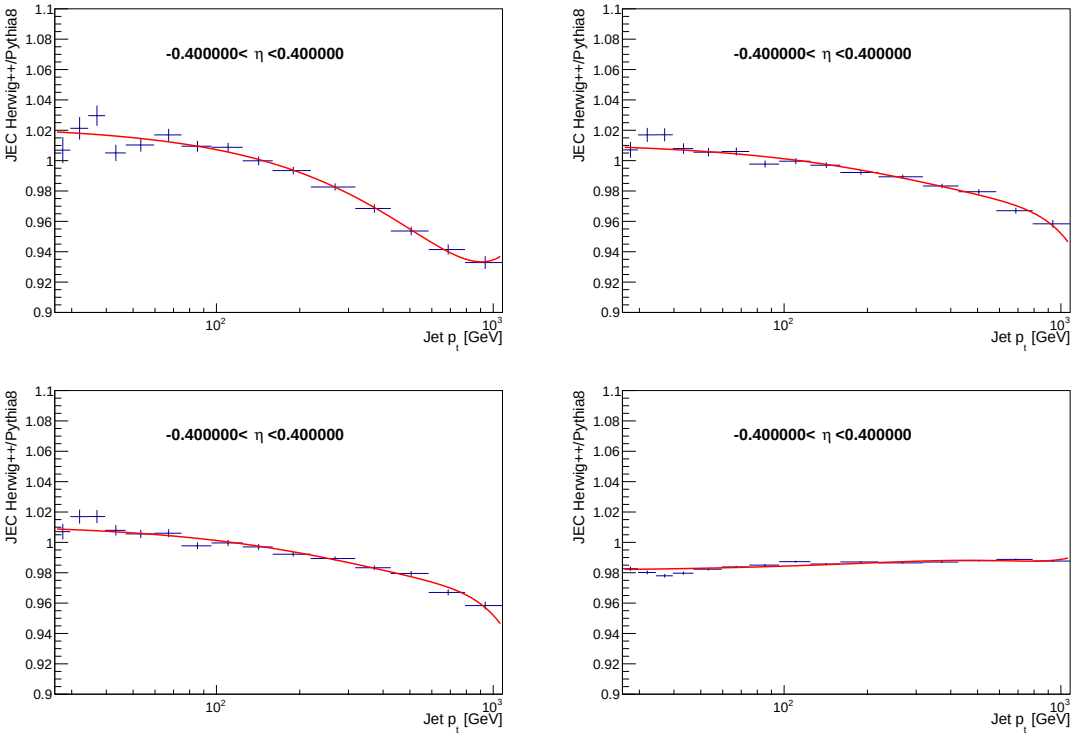


Figure 5.8.: The ratio of DeepJet JEC derived in the central η bin using Herwig++ and Pythia8. Shown are the ratios for b jets (top left), c jets (top right), uds jets (bottom left) and gluon jets (bottom right).

5.4. Conclusion

In this chapter the DeepJet model was extended to perform jet energy regression. This allows to estimate a jet energy correction that can use the full flavour information of the jet to propose a correction. Furthermore, the jet energy corrections can be parameterized in more variables than the standards of p_t and η of the reconstructed jets and pileup variables. An improvement in the mean jet response was observed, and a roughly 10-20% improvement was obtained in the jet energy resolution for the different jet flavours. The difference in mean jet energy response for different generators was however unchanged, indicating that the learned jet energy corrections still is not utilizing the full jet shape information. This might indicate that a larger network is required to learn the more subtle details of jet energy regression.

Chapter 6.

Reducing the Top Quark Mass Systematic Uncertainty with Machine Learning

As machine learning has been rapidly developed in the last years due to the steady rise in computer processing power, increasingly complex techniques have been adopted in the field of particle physics. For event selection, machine learning has mainly been employed with the aim of improving the statistical measurement precision by reducing background rates and increasing selection efficiencies. A common concern when applying machine learning is increasing systematic uncertainties, as the machine learning algorithms usually are trained on simulated collision data, and require a high number of inputs. For example a large scale neural network has not yet been deployed on collision data for quark/gluon tagging at the CMS experiment due to the difficulties in modelling the difference of gluon and quark jets in simulation [141]. Some techniques have already been developed in machine learning to tackle the issues of retaining the machine learning algorithm performance, while controlling systematic uncertainties [142–145]. However, the avenue of using machine learning to specifically reduce systematic uncertainties for a precision measurement, is not well explored. In general, physics analyses resulting in a statistical estimator of a physics parameter have a broad portfolio of systematic uncertainties. Accordingly finding a phase space selection that optimally minimizes all of them is highly challenging, making it a suitable task for a neural network. The main difficulty to address this problem was to define a suitable metric for the machine learning algorithm to minimize. The ReSYST method [16] defines a non observable metric, called the ReSYST quantifier R_i

for each collision event i consider by the estimator of a physics parameter. The metric can be used to identify and quantify the relationship between the individual events and the systematic uncertainties. Some events in the selected event sample could have a large impact on the systematic uncertainty, while some other events might have a much smaller impact. Using machine learning, this metric can be predicted for each event from observable variables. Accordingly, the machine learning estimated value of the metric becomes an event observable that can then be used to make a further event selection that reduces the total systematic uncertainty of the measurement. In the original ReSYST paper, a DELPHES top quark mass study is performed to test the concept by relating the non observable metric to a single event observable and then to propose a further event selection based on this single observable. In this chapter, the ReSYST method will be deployed with realistic simulations, machine learning and an advanced top quark mass estimator. The aim is to learn the viability of the ReSYST method in a complex and realistic analysis setting, as well as demonstrating how machine learning can improve the method.

6.1. Top quark mass measurement

The top quark mass measurement method that will be studied follows closely Ref. [146], which is a CMS top quark mass measurement in the lepton+jets channel published in 2018. The main extension and the focus of this chapter is trying to add an additional event selection cut using the method of ReSYST enhanced by machine learning techniques. The analysis strategy is also extended by applying the machine learning methods developed in Chapters 4 and 5 for jet flavour tagging and jet energy estimation. The mass is measured in top quark pair events, where one of the W bosons from the top quark decay is hadronically decaying, and the other decays leptonically to a muon and a neutrino. In the CMS paper the electron channel is also considered, however this is not required to demonstrate the relative improvement of using ReSYST with machine learning. The decay channel which is considered is referred to as the muon+jets channel. The channel is chosen because it results in a dominating systematic uncertainty relative to an almost negligible statistical uncertainty on the top quark mass estimator. However, in principle both the all-jets channel, where both W bosons decay hadronically, and the dilepton channel, where both W bosons

decay leptonically, should be possible to use with the ReSYST methodology. Since the top quark mass measurement in the muon+jets channel is extremely sensitive to uncertainties on the jet energy scale (JES), the Ideogram methodology is applied [147]. This method exploits that since the mass of the W boson is known to high precision, it can be used to constrain the JES in-situ on the observed $W \to jj$ decay. A 2D likelihood function is used to infer the top quark mass and a jet energy scale factor (JSF) from the reconstructed W boson mass (before it is constrained in a kinematic fit) and the fitted top quark mass (after it has been constrained in a kinematic fit) in the event. Again, a more simple top quark mass estimator, for example a 1D likelihood of the top quark mass, could have been used, and it should work fine with ReSYST as demonstrated in the original paper [16]. However, the extended method is used as it results in a lower uncertainty by default, and it allows us to explore the robustness of ReSYST in a more complex environment. Indeed, if the ReSYST method works with the Ideogram method, then it might be an indication that it can be applied with higher dimensional likelihood methods and profile likelihood methods, which have shown great promise for top quark mass measurements [148].

6.1.1. The $pp \to t\bar{t} \to bq\bar{q}b\mu\nu$ event selection

The process in which the measurement is done is $pp \to t\bar{t} \to bq\bar{q}b\mu\nu$, and an event selection is defined accordingly. A CMS 2017 simulated sample of $t\bar{t} \to$ jets + leptons events are used. The events are generated with the POWHEGv2 [121–124] matrix element generator. PYTHIA8 [39] with the CP5 tune [149] is used for parton showering and hadronization. Table 6.1 shows the different stages of selection, as well as the expected total number of $t\bar{t} \to$ jets + $\mu/e/\tau$ for the CMS 2017 dataset of 41.5 fb^{-1}. The event trigger used is a high level isolated muon trigger with a p_t requirement of at least 27GeV and $|\eta| < 2.5$. This removes roughly 80% of the $t\bar{t} \to$ jets + $\mu/e/\tau$ events, consisting of the e/τ + jets events, and events where the muon did not pass the trigger selection or wasn't in the detector acceptance. Some light event cleaning is performed putting requirements on the quality of the primary vertex reconstruction of the event as well as removing events with detector signals indicating fake MET. These quality requirements are inspired by comparisons between real collision data and simulated collision. An isolated signal muon with $p_t > 27$GeV, $|\eta| < 2.4$ is also required using offline reconstruction. Additionally tight muon ID requirements [150] are imposed on the signal muon. To reduce backgrounds from the $t\bar{t}$ dileptonic channel

as well as from Z+jets, a veto is applied on events containing additional muons that are reconstructed globally as well as potential electrons with $p_t > 15$GeV that meet the electron ID requirements. At this stage no $t\bar{t} \rightarrow \text{jets} + e$ events are remaining. In order to select the four jets coming from the top quarks and the W boson at least four AK4 jets passing tight "JetID" [151] with $p_t > 30$GeV and $|\eta| < 2.4$ are required. Since the top quarks almost always decay to one b quark, two of the four leading jets are required to be b-tagged. These jets are selected with the DeepJet algorithm, where the medium working point is used providing an efficiency of 83% to select b quark jets with a 1% efficiency for light quark jets. Besides identifying the b-jets for top quark reconstruction, this also removes events from the W+jets background process. A final selection is based on the probability of the kinematic fit used for the top quark reconstruction, which will be discussed in the following section. A number of background channels exists for this process, such as a Z or W boson produced with additional jets or single top quark production. However previous studies [146] with this baseline event selection found them to be only 4% of the total event yield, and only have a negligible impact on the systematic uncertainty on the measured top quark mass. Therefore for simplicity background events are not included in this study, because they do not affect our objective to demonstrate the relative improvement in systematic uncertainty on the top quark mass estimator when using machine learning with ReSYST.

Selection stage	Fraction of total events	Expected number of events for 41.53 fb^{-1} with σ = 347.4 pb^{-1} at 13TeV [40]
Before selection	100.0%	14420000 ± 420000
Muon Trigger (L1 & HLT)	20.0%	2890000 ± 85000
Event Cleaning	19.6%	2820000 ± 83000
Signal muon selection	18.2%	2620000 ± 77000
Muon veto	18.1%	2600000 ± 76000
Electron veto	17.7%	2540000 ± 75000
Four jet selection	8.2%	1180000 ± 35000
B-Tagging	4.6%	430000 ± 19000
HitFit probability cut	1.1%	150000 ± 5000

Table 6.1.: Sequential steps in the event selection starting from a $pp \rightarrow t\bar{t} \rightarrow \text{jets} + \text{lepton}$ sample with top quark mass 172.5 GeV.

6.1.2. Top Quark Reconstruction

A constrained kinematic fit [152, 153] is performed on the selected $pp \to t\bar{t} \to bq\bar{q}\bar{b}\mu\nu$ events in order to improve the resolution of the measured top quark mass and to select events that are in agreement with the $pp \to t\bar{t} \to bqqb\mu\nu$ hypothesis. The momentum of the four leading jets and the muon as well as the transverse missing momentum are used as constraints to the fitter giving 17 constraints. Additional constraints are set for the masses to be zero for the muon and the light quarks and 4.7 GeV for the b-quark. The χ^2 to be minimized is shown in Equation 6.1.

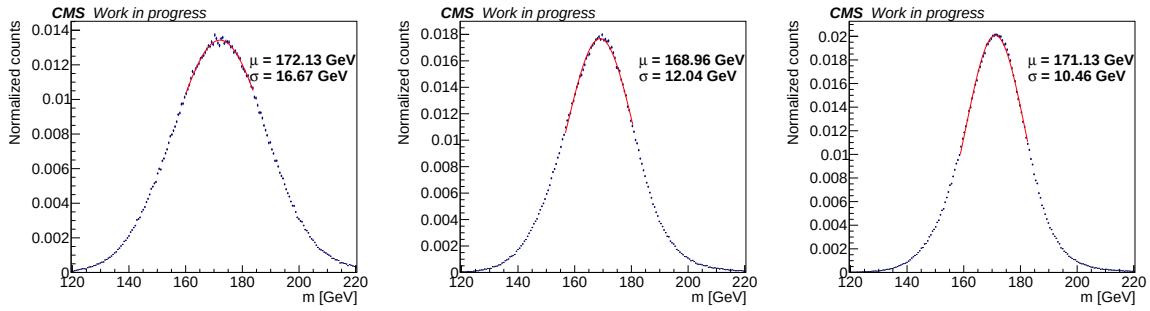
$$\chi^2 = (\mathbf{x} - \mathbf{x}_m)^T G (\mathbf{x} - \mathbf{x}_m) \tag{6.1}$$

Here $\mathbf{x}$ is the vector of the fitted variables, and $\mathbf{x}_m$ are their measured values provided as input. The matrix G is the inverse error matrix. To derive the inverse error matrix, the resolutions of the kinematic quantities are utilized. These resolutions were obtained using the methods of Reference [154] [1]. The χ^2 is minimized with three additional constraints. The invariant mass of the two light jets and of the neutrino and muon are fixed to the W boson mass assumed in the simulation, namely of 80.4 GeV. Furthermore the two top quarks in the event topology are constrained to have the same mass. Given that the constraints of the top quark and W boson masses are not linear in the fit variables, it is not possible to exactly solve the system of equations with Lagrange multipliers. Therefore an iterative technique called the HitFit [153] method is used. A power series expansion of the constraints is developed around the starting value given by the measured quantities, such that they can be made linear. With these linear constraints, the system is solved. The result of the fit is used as the starting point of the next iteration, and the process is continued until χ^2 value is no longer improved. This method requires that a starting value is specified for the momentum of the z-component of the neutrino. This can be obtained from the quadratic equation that the two top quarks should have the same mass resulting in two possible solutions. Additionally, there are two possible parton to jet assignments as each of the b-jets can either be associated with the hadronically or leptonically decaying top quark. Therefore every event has in total four possible permutations. These permutations can be classified in simulation as either being correct, wrong or unmatched. Correct permutations occur in events where the four leading jets can be associated to the four

[1] The code used to calculate the top quark fit resolutions was updated by Hannu Siikonen, Helsinki Institute of Physics, with additional improvements on Ref. [154]. Details can be found in his soon published PhD thesis.

quarks coming from the $t\bar{t}$ decay within $\Delta R < 0.3$, and the jets are then correctly assigned to the partons, such that the correct W boson is associated with the correct b jet to form a top quark. Wrong permutations occur in events when the correct jets are present and selected in the event, but they are assigned to the wrong partons, either by incorrect flavour tagging or by associating the W boson with the wrong b-jet. Finally, unmatched permutations occur in the events where the four leading jets cannot be associated with the partons. This can happen due to many reasons, such as two partons being within the matching distance of the same jet, initial state radiation producing one of the four leading jets in the event or simply because a parton develops a jet that fails the selection criteria.

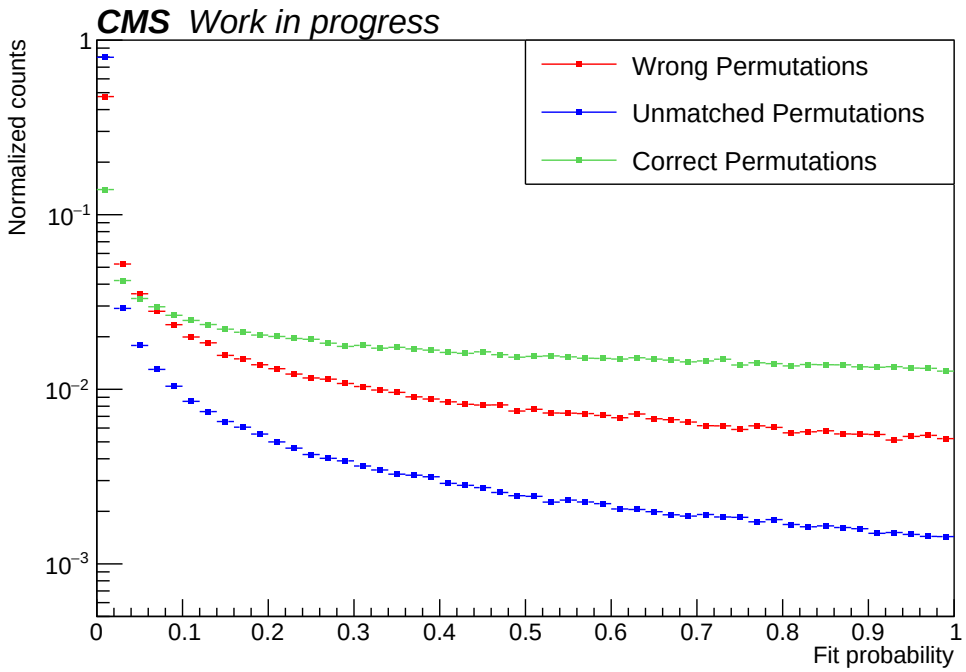
Figure 6.1.: The mass distribution of a correctly reconstructed top quark ($t \to bjj$) is shown before being constrained by the kinematic fit (left), after being fitted (middle) and after being fitted with DeepJet JEC applied (right). The mean and standard deviation of a Gaussian fit of the peak of the distribution (shown in red) is indicated on the plots.



The distribution of the reconstructed top quark mass for correct permutations before and after being improved by the kinematic fit can be seen in Figure 6.1. Also shown is the top quark mass distribution after the kinematic fit is performed using jets where the DeepJet JEC has been applied. The fit improves the resolution of the top quark mass by about 30%. A further $\sim$10% improvement is obtained when using DeepJet JEC.

Each of the four permutations in the event are initially considered, and each will have a fit probability associated with it. A plot of the fit probability distributions for each type of permutation can be found in Figure 6.2. As can be seen, the correct permutation becomes the dominant contribution for higher fit probabilities, as expected. In order to increase the fraction of correct permutations, only permutations with a fit probability of at least 0.2 are considered.

Figure 6.2.: The fit probability distribution is shown for each permutation category.



Permutation type	No cut	$p_{\text{fit}} > 0.2$	$p_{\text{fit}} > 0.2$ with DeepJet JEC
Correct	15%	44%	47%
Wrong	16%	23%	23%
Unmatched	69%	33%	30%

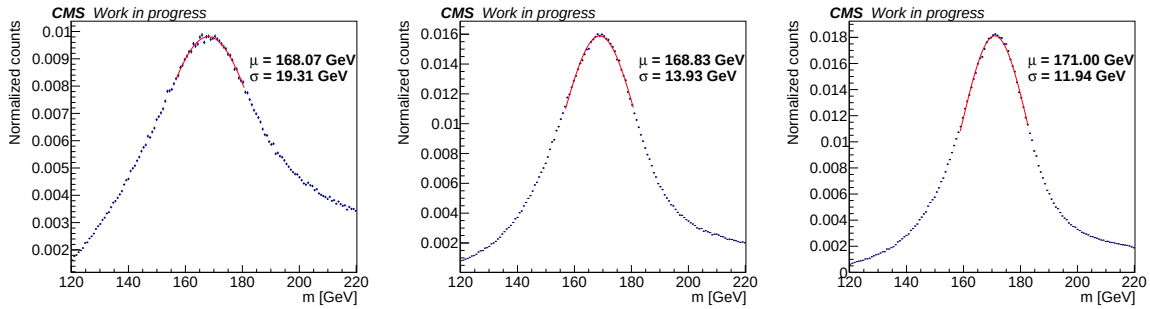
Table 6.2.: Fractions of the permutation types are shown with and without the fit probability cut, and with the DeepJet jet corrections applied to the jets.

In the Table 6.2 the fraction of all permutations considered for different selections are shown. When applying the additional $p_{\text{fit}} > 0.2$ criteria there is a significant increase in the fraction of correct permutations. Additionally it is seen that when applying the additional jet energy corrections with the DeepJet JEC method developed in Chapter 5, a further small improvement is obtained.

In Figure 6.3 the distribution of the fitted top quark mass is shown before and after applying the fit probability cut. A Gaussian is fitted to the peak of the distribution, and it can be seen that the resolution is significantly improved mainly due to the increase of correct permutations. Additionally the distribution is shown when the DeepJet JEC is applied as well giving a further improvement in resolution. This improvement is a

combination of both the larger fraction of correct permutations and the better mass resolution for correct permutations.

Figure 6.3.: The mass distribution of the fitted top quark is shown with no fit probability cut (left), with a fit probability cut (middle) and with both a fit probability cut and DeepJet JEC (right). The mean and standard deviation of a Gaussian fit of the peak of the distribution (shown in red) is indicated on the plots.



6.1.3. Ideogram Method

The Ideogram method [147] estimates the top quark mass and a jet energy scale factor by utilizing a 2D likelihood function parameterized in two observable variables. The first observable is the top quark mass as fitted by HitFit. The second observable is the invariant mass of the two leading light quark jets. For correct permutations this invariant mass corresponds to the hadronically decaying W boson in the $pp \rightarrow t\bar{t} \rightarrow bqqb\mu\nu$ process. As the W boson mass is known with high precision from previous measurements, it can be used to make an in-situ measurement of an additional jet energy scale correction such that the invariant mass of the reconstructed jets indeed corresponds to the best known value of the W boson mass. With this approach the experimental uncertainty on the top quark mass from jet energy corrections is decreased significantly. Accordingly, the Ideogram method allows for a simultaneous measurement of the top quark mass (m_t) and an additional jet energy scale factor (JSF).

In order to construct this 2D likelihood, templates are obtained of the reconstructed W boson mass ($m_{W,\text{reco}}$) and fitted top quark mass ($m_{t,\text{fit}}$) distributions for various points of an underlying true value (a generated value) of the top quark mass ($m_{t,\text{gen}}$) and of the jet scale factor (JSF_{gen}). The set of values used for the $m_{t,\text{gen}}$ variable corresponds to $\{166.5, 169.5, 171.5, 172.5, 173.5, 175.5, 178.5\}$ GeV. For the JSF_{gen} variable, the used set of values is $\{0.96, 0.98, 1.00, 1.02, 1.04\}$. The combination of these sets of values, yield

a total of 35 samples. These template distributions are further separated for correct, wrong and unmatched permutations, since the template shape of the fitted top quark mass is significantly different for these categories. An example of these templates can be seen on Figure 6.4, where the distributions of $m_{W,\text{reco}}$ and $m_{t,\text{fit}}$ are shown for each permutation category for the generated values $(m_{t,\text{gen}}, \text{JSF}_{\text{gen}}) = (172.5\,\text{GeV}, 1.00)$.

Figure 6.4.: Templates after the described selection for a generated top quark mass 172.5 GeV and a generated JSF value of 1.00. The top row shows the fitted top quark mass distribution for correct (left), wrong (center) and unmatched (right) permutations. The bottom row shows this for the reconstructed W boson mass.

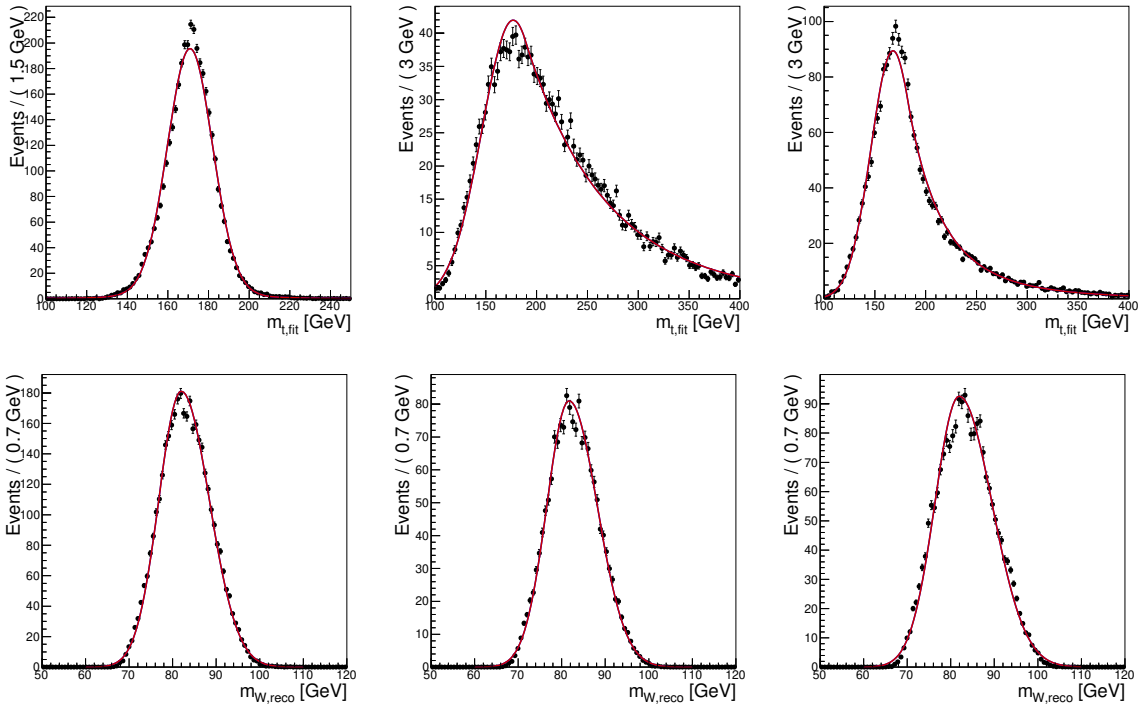
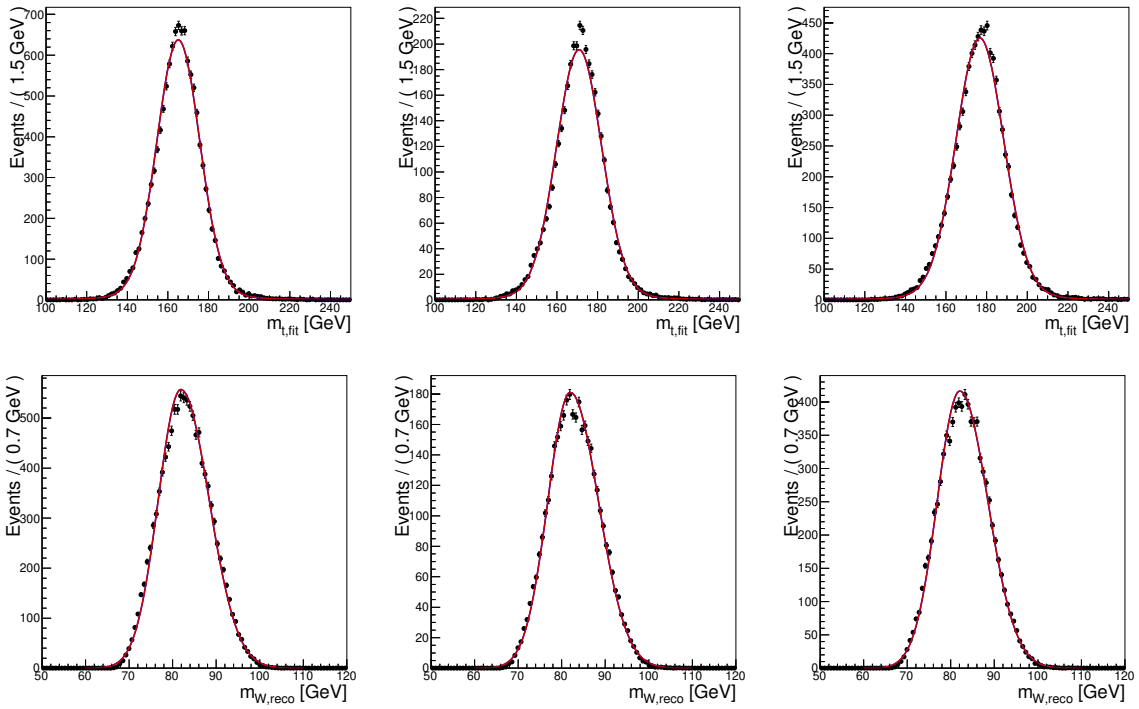


Figure 6.5 shows the fitted top quark mass distribution for correct permutations with $\text{JSF}_{\text{gen}} = 1.00$, but varying values of $m_{t,\text{gen}}$. As expected the $m_{t,\text{fit}}$ variable is seen to be very sensitive to the generated top quark mass, and the mean of the distribution is observed to be strongly correlated with the used value of $m_{t,\text{gen}}$. On the same figure the reconstructed W boson mass distribution is also shown for different values of $m_{t,\text{gen}}$ with $\text{JSF}_{\text{gen}} = 1.00$. As expected the reconstructed W boson mass distribution is not sensitive to the variations of $m_{t,\text{gen}}$.

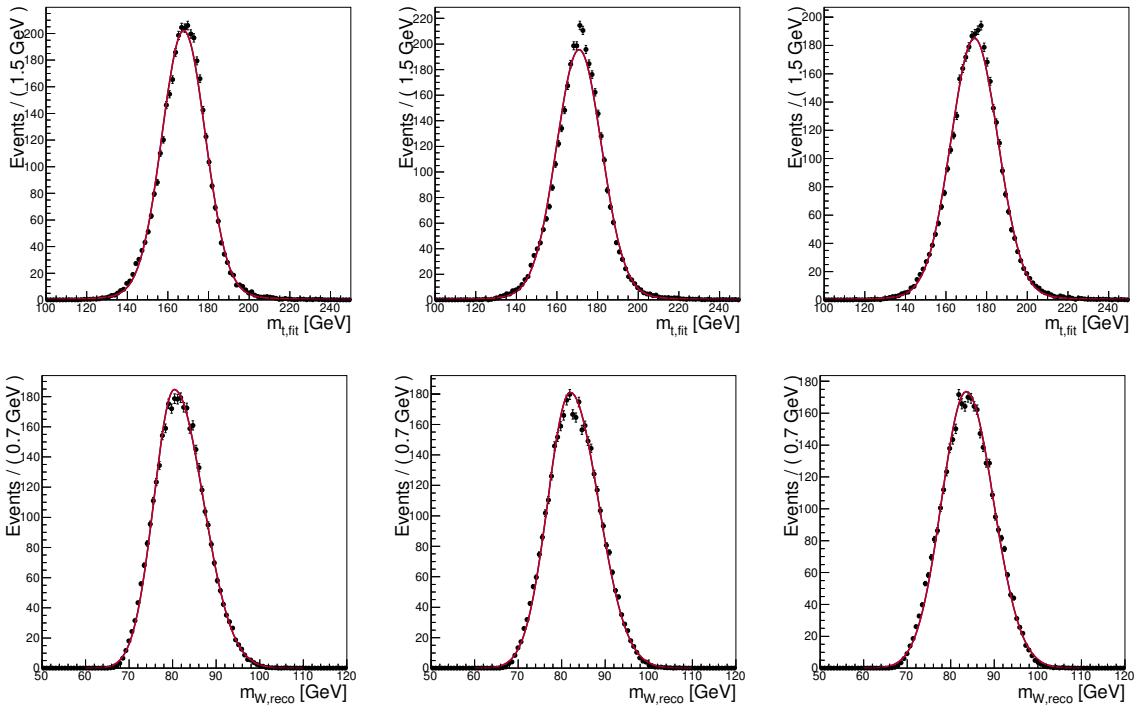
Figure 6.6 shows both the $m_{W,\text{reco}}$ and $m_{t,\text{fit}}$ distributions for $m_{t,\text{gen}} = 172.5$ GeV but with varying values of JSF_{gen}. As both of these variables are reconstructed from jets, they are both seen to be very sensitive to the generated jet energy scale factor.

Figure 6.5.: Templates for the standard top mass measurement for different top quark generator values with JSF_{gen} fixed to 1.00. The top row shows correct permutations for the fitted top quark mass for generator top quark mass of 166.5 GeV (left), 172.5 GeV (center) and 178.5 GeV (right). The bottom row shows the reconstructed W boson mass for the same samples.



The fitted top quark mass distributions for correct permutations are fitted with a Voigtian, whereas for wrong and unmatched permutations they are fitted with a Crystal Ball function. The reconstructed W boson mass distributions are fitted with a Bifurcated Gaussian for all permutations. In order to smoothly interpolate between the template functions of the $m_{W,\text{reco}}$ and $m_{t,\text{fit}}$ distributions at discrete sets of generated top quark mass and true JSF points the evolution of the parameters of the functions with $m_{t,\text{gen}}$ and JSF_{gen} are fitted. The dependence of the parameters α_i is modelled as $\alpha_i = a_i + b_i \cdot (m_{t,\text{gen}} - 172.5) + (c_i + d_i \cdot (m_{t,\text{gen}} - 172.5)) \cdot (JSF_{gen} - 1.)$ with a_i, b_i, c_i, d_i being free parameters. The value of a_i, b_i, c_i, d_i are obtained from an unbinned likelihood fit of the template functions on the template distributions performed with RooFit [155]. Using the template functions a permutation based 2D

Figure 6.6.: Templates for the standard top quark mass measurement for different $\mathrm{JSF}_{\mathrm{gen}}$ values with generator top quark mass fixed to 172.5 GeV. The top row shows the fitted top quark mass distribution for correct permutations for $\mathrm{JSF}_{\mathrm{gen}}$ of 0.96 (left), 1.00 (center) and 1.04 (right). The bottom row shows the reconstructed W boson mass for the same samples.



likelihood is constructed. This function takes the form of Equation 6.2.

$$\begin{aligned}
&\mathcal{L}_{\mathrm{perm}}(m_{t,\mathrm{fit}}, m_{W,\mathrm{reco}} | m_{t,\mathrm{gen}}, \mathrm{JSF}_{\mathrm{gen}}) = \\
&f_{\mathrm{CP}} \cdot L^{\mathrm{CP}}_{\mathrm{m_t}}(m_{t,\mathrm{fit}}, m_{W,\mathrm{reco}} | m_{t,\mathrm{gen}}, \mathrm{JSF}_{\mathrm{gen}}) \cdot L^{\mathrm{CP}}_{\mathrm{m_W}}(m_{t,\mathrm{fit}}, m_{W,\mathrm{reco}} | m_{t,\mathrm{gen}}, \mathrm{JSF}_{\mathrm{gen}}) \\
+&f_{\mathrm{WP}} \cdot L^{\mathrm{WP}}_{\mathrm{m_t}}(m_{t,\mathrm{fit}}, m_{W,\mathrm{reco}} | m_{t,\mathrm{gen}}, \mathrm{JSF}_{\mathrm{gen}}) \cdot L^{\mathrm{WP}}_{\mathrm{m_W}}(m_{t,\mathrm{fit}}, m_{W,\mathrm{reco}} | m_{t,\mathrm{gen}}, \mathrm{JSF}_{\mathrm{gen}}) \\
+&f_{\mathrm{UN}} \cdot L^{\mathrm{UN}}_{\mathrm{m_t}}(m_{t,\mathrm{fit}}, m_{W,\mathrm{reco}} | m_{t,\mathrm{gen}}, \mathrm{JSF}_{\mathrm{gen}}) \cdot L^{\mathrm{UN}}_{\mathrm{m_W}}(m_{t,\mathrm{fit}}, m_{W,\mathrm{reco}} | m_{t,\mathrm{gen}}, \mathrm{JSF}_{\mathrm{gen}})
\end{aligned} \tag{6.2}$$

Here the L functions indicate the template functions parameterized in $m_{t,\mathrm{gen}}$ and $\mathrm{JSF}_{\mathrm{gen}}$. The factors f indicate the fraction of permutations that are correct, wrong or unmatched in the event sample. The event based likelihood is constructed with a weighted sum of the likelihoods for each permutation. Of course only the permutations in the event that meet the fit probability requirement of $p_{\mathrm{fit}} > 0.2$ are included. The permutations are weighted with the HitFit fit probability of the given permutation as

in Equation 6.3.

$$\mathcal{L}_{\text{event}}(m_{t,\text{fit}}, m_{W,\text{reco}}|m_{t,\text{gen}}, \text{JSF}_{\text{gen}}) = \sum_{\text{perms}} p_{\text{fit}} \mathcal{L}_{\text{perm}}(m_{t,\text{fit}}, m_{W,\text{reco}}|m_{t,\text{gen}}, \text{JSF}_{\text{gen}}) \tag{6.3}$$

Finally the full sample likelihood is constructed by taking the product of event likelihoods. An event weight is introduced as $w_{\text{evt}} = c \sum_{\text{perms}} p_{\text{fit}}$ in order to reduce the impact of events without correct permutations. Here c is a normalization constant such that $\sum_{\text{events}} w_{\text{evt}} = N_{\text{events}}$.

$$\mathcal{L}_{\text{sample}}(m_{t,\text{fit}}, m_{W,\text{reco}}|m_{t,\text{gen}}, \text{JSF}_{\text{gen}}) = \prod_{\text{events}} (\mathcal{L}_{\text{event}}(m_{t,\text{fit}}, m_{W,\text{reco}}|m_{t,\text{gen}}, \text{JSF}_{\text{gen}}))^{w_{\text{evt}}} \tag{6.4}$$

The estimator of the top quark mass and JSF is found by maximizing the likelihood. The properties of the estimators are studied with a bootstrap procedure. The expectation value of the top quark mass and JSF estimators are slightly biased from the generated values. This is due to several reasons, for example the assumption that the fraction of permutations are the same for every mass and JSF point. For each pseudo-experiment 10000 events are sampled. The probability of including a given event in the pseudo-experiment is determined based on its simulated event weight. Roughly 1000 pseudo-experiments are conducted for each of the 35 $(m_{t,\text{gen}}, \text{JSF}_{\text{gen}})$ samples. If biases are present, a calibration procedure is performed to correct the estimators to account for these. This is done by fitting the following 2D calibration functions to the biases.

$$\begin{aligned} m_t^{\text{calib}}(m_t, \text{JSF}) &= m_t + a_1 + b_1 \cdot (m_t - 172.5) \\ &+ c_1 \cdot (\text{JSF} - 1.0) + d_1 \cdot (m_t - 172.5) \cdot (\text{JSF} - 1.0) \end{aligned} \tag{6.5}$$

$$\begin{aligned} \text{JSF}^{\text{calib}}(m_t, \text{JSF}) &= \text{JSF} + a_2 + b_2 \cdot (m_t - 172.5) \\ &+ c_2 \cdot (\text{JSF} - 1.0) + d_2 \cdot (m_t - 172.5) \cdot (\text{JSF} - 1.0) \end{aligned}$$

The parameters $a_1, b_1, c_1, d_1, a_2, b_2, c_2$ and d_2 are free in the fit. These calibration functions are then applied to the m_t and JSF estimators respectively. Figure 6.7 shows the bias on the mass estimator before and after calibration as a function of $m_{t,\text{gen}}$ for

three of the five $\mathrm{JSF}_{\mathrm{gen}}$ values. Similar trends are observed for the remaining $\mathrm{JSF}_{\mathrm{gen}}$ values. As can be seen a mass shift of roughly -0.25 GeV is observed before calibration for all $m_{t,\mathrm{gen}}$ and $\mathrm{JSF}_{\mathrm{gen}}$ samples. After the calibration procedure, this is significantly reduced everywhere. Potential smaller biases remaining will be accounted for with a systematic uncertainty. Figure 6.8 shows the bias on the JSF estimator before and after calibration. Before calibration a JSF shift of roughly -0.5% is observed for all $m_{t,\mathrm{gen}}$ and $\mathrm{JSF}_{\mathrm{gen}}$ samples. Similarly to the mass estimator, biases are significantly reduced with the calibration procedure.

Figure 6.7.: Biases are shown on the top quark mass estimator as a function of the generated mass before and after the calibration procedure. The biases are shown for $\mathrm{JSF}_{\mathrm{gen}}$ values of 0.96, 1.00 and 1.04.

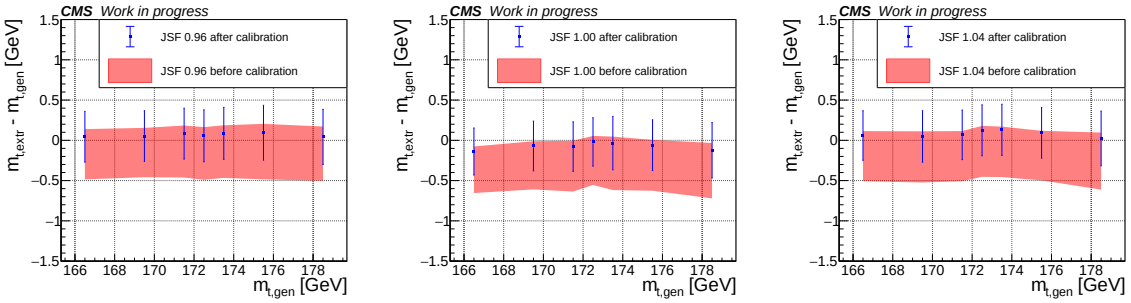
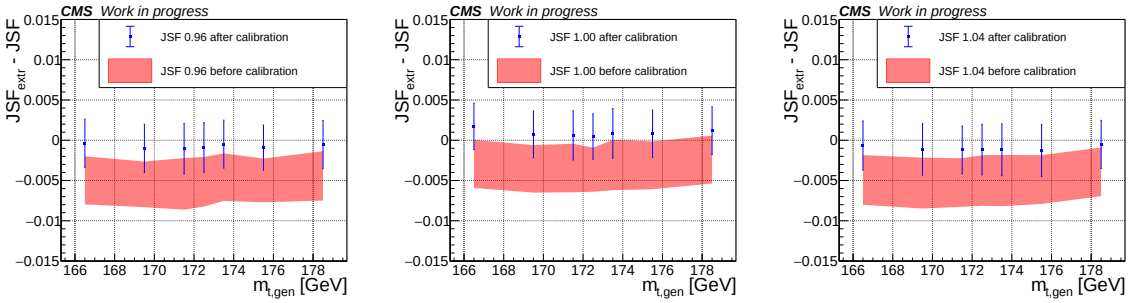


Figure 6.8.: Biases are shown on the JSF estimator as a function of the generated mass before and after the calibration procedure. The biases are shown for $\mathrm{JSF}_{\mathrm{gen}}$ values of 0.96, 1.00 and 1.04.



6.1.4. Systematic Uncertainties

In order to study the properties of the top quark mass and JSF estimators, it is required to use simulations of proton collision events at 13 TeV. When simulation does not agree with the real collisions observed by the CMS experiment, systematic uncertainties are induced. These discrepancies between data and simulation can originate from an incomplete understanding of the modelling of the physics phenomena in proton collisions, but they can also originate from an imperfect simulation of the CMS detector. An overview of the relevant systematic uncertainties on the top quark mass measurement is listed below. Additionally, the chosen method for estimating them are described.

6.1.5. Experimental uncertainties

Jet Energy Corrections The jet energy is scaled up and down in correspondence to the uncertainty on the jet energy response induced by differences between data and simulation [72]. This is the largest uncertainty in the 1D top quark mass measurement, but due to the in-situ calibration in the Ideogram method, it becomes quite small in the 2D top quark mass and JSF measurement.

Mass Calibration A systematic uncertainty is assigned to this procedure by fitting $f(x) = a$ to the calibrated biases on the m_t and JSF estimators, and then taking the fit uncertainty of the parameter a as the systematic uncertainty.

Jet Energy Resolution The jet energy resolution is better in simulation compared to data. To account for this a smearing of all jet energies is performed in simulated samples to match the resolution found in data. The systematic uncertainty is evaluated by varying the nominal smearing up and down based on the derived uncertainties from the measured resolution in data [72].

b-tagging The b-tagging scale factors are varied up and down within their uncertainties [117].

Pileup The total inelastic pp cross section is varied up and down within its mea-

surement uncertainty. This alters the estimated number of pileup collisions in the event, therefore affecting the m_t and JSF measurement.

6.1.6. Modeling of hadronization

JEC Flavour uncertainty The largest systematic uncertainty in the top quark mass measurement is the JEC flavour uncertainty. It is evaluated by applying an additional flavour based jet energy correction, based on the difference in jet energy response for the separate jet flavours using simulated event samples generated with Herwig++ and Pythia8. Since this applies an energy correction to each flavour separately the Ideogram Method cannot reduce this uncertainty with the in-situ jet calibration to the W boson mass. Each flavour variations is done separately. The obtained shifts on the m_t and JSF estimators from each flavour are combined by a linear sum, since the variations built up together the total difference between the two generators.

b fragmentation The Bowler-Lund fragmentation function for B hadrons has been tuned to match measurements in e^+e^- data by the ALEPH and DELPHI collaborations [156, 157]. The uncertainties on this tuning are used to evaluate the systematic uncertainties on m_t and JSF from b-fragmentation.

Peterson fragmentation The alternative Peterson fragmentation function [158] is used instead of Bowler-Lund and the difference observed on the m_t and JSF estimators is taken as a systematic uncertainty due to an incomplete understanding of the phenomenon from first principles.

Semileptonic B hadron decays The branching fraction of the semileptonic B hadron decay is varied within its experimental measurement uncertainties [17]. This increases the rate of neutrinos in the b-jets and therefore affects the energy response and the top quark mass measurement.

6.1.7. Modeling of pertubative QCD

Matrix element generator The alternative MadGraph5 aMC@NLO generator with the FxFx matching [159, 160] is used instead of POWHEGv2. The difference observed on the m_t and JSF estimators is taken as the systematic uncertainty.

Matrix element/parton shower (ME/PS) matching The matrix element/parton shower matching is set by the h_{damp} parameter in POWHEGv2. This variable is varied up and down within its uncertainties to define a systematic uncertainty on the estimators.

Initial state radiation (ISR) parton shower scale The scale value of the parton shower for the initial state radiation in PYTHIA is varied by 4 and 0.25 using event weights [161]. The differences it induces on the estimators relative to the nominal value are used as systematic uncertainties.

Final state radiation (FSR) parton shower scale The scale value of the parton shower for the final state radiation in PYTHIA is varied by 4 and 0.25 using event weights [161]. The differences it induces on the estimators relative to the nominal value are used as systematic uncertainties.

Top quark transverse momentum modeling The top quark transverse momentum distribution is observed to be different in data and simulation, possibly due to the NNLO effects. The simulated distribution is reweighted to match the distribution observed in CMS data [162, 163]. The differences this induces on the estimators when applying these weights are taken as systematic uncertainties.

6.1.8. Modeling of soft QCD

Underlying event To account for soft QCD effects PYTHIA is tuned to measurements of the underlying event with the CP5 tune [149]. The tune parameters are varied within their uncertainties in order to estimate the systematic uncertainties on the estimators.

Early resonance decay Early resonance decays (ERD) can be turned on in PYTHIA to allow color reconnection between the top quark decay particles and particles from the underlying event. By default this is turned off. The difference between on and off induces systematic uncertainties on the estimators.

Color reconnection models Two different color reconnection models [164, 165] are compared with the default PYTHIA model. The first model allows string formation beyond leading color. The second model allows gluons to be moved to another string.

Both models have been tuned to the underlying event based on CMS data [166]. The larger differences induced on the estimators between the nominal model and either of these two models is assigned as a systematic uncertainty.

6.1.9. Total systematic uncertainties

In Table 6.3 the total systematic and statistical uncertainties for a top quark mass in a simulated sample with a generator top quark mass of 172.5 GeV can be found. The uncertainty table contains both symmetric, asymmetric and one sided uncertainties. One sided uncertainties are represented by a single value in the "m_t 1st variation" column, and two sided asymmetric uncertainties are represented by two values in both columns. The symmetric uncertainties are indicated in the table by a single value in the "δm_t 1st variation" column with a $\pm$ sign. The total systematic uncertainty reported in the "δm_t 1st variation" column is calculated by combining all systematic variations yielding $\delta m_t > 0$ in a square sum and taking the square root. Similarly the total systematic uncertainty in the "δm_t 1st variation" column is calculated by combining all systematic variations yielding $\delta m_t < 0$ in a square sum and taking the square root. The JEC flavour uncertainty for single flavours are shown in italic, and the total JEC flavour uncertainty is found by taking the linear sum of these contributions. The b-jet modelling uncertainties are shown combined and independently. The table also shows what is referred to as reducible and irreducible systematic uncertainties. As will be discussed in the following section these refer to uncertainties that can and cannot be targeted with the ReSYST method. The irreducible uncertainties consists of the matrix element generator, ME/PS matching, underlying event, early resonance decay and the color reconnection uncertainties. The reducible uncertainties contain the rest. As can be seen from the table, the dominating systematic uncertainties for the top quark mass are the JEC flavour uncertainty, the jet energy resolution uncertainty as well as the FSR and ISR parton shower scale uncertainties. The expected statistical uncertainty is estimated from the 2017 CMS dataset luminosity. As can be seen the total systematic uncertainty is significantly larger than the expected statistical uncertainty. In the case of the jet scale factor measurement the largest uncertainty is from the jet energy correction uncertainties as expected.

Uncertainty source	δm_t 1st variation (MeV)	δm_t 2nd variation (MeV)	δJSF 1st variation	δJSF 2nd variation
Experimental uncertainties				
Jet Energy Corrections	-37	69	0.8%	-0.8%
Mass calibration	$\pm$ 54		$\pm$ 0.1%	
Jet Energy Resolution	-311	353	0.5%	-0.5%
b-tagging	14	-22	<0.1%	-<0.1%
Pileup	-116	103	0.1%	-0.1%
Modeling of hadronization				
JEC Flavour (linear sum)	$\pm$ 355		$\pm$ 0.4%	
-light quarks	*-343*		*0.5%*	
-bottom quarks	*683*		*-<0.1%*	
-gluon	*69*		*-0.1%*	
-charm quarks	*-54*		*-<0.1%*	
b-jet modeling (squared sum)	129	-183	0.1%	-0.1%
-Bowler-Lund fragmentation	*-170*	*95*	*0.1%*	*-0.1%*
-Peterson fragmentation	*-30*		*-<0.1%*	
-Semileptonic B hadron decays	*88*	*-61*	*-0.1%*	*-<0.1%*
Modeling of pertubative QCD				
Matrix element generator	193		0.1%	
ME/PS matching	-56	26	0.1%	-0.1%
ISR parton shower scale	371	-19	-0.4%	-<0.1%
FSR parton shower scale	-241	-17	0.3%	-<0.1%
Top quark p_t uncertainty	72		-0.2%	
Modeling of soft QCD				
Underlying event	-95	-<1	0.1%	-<0.1%
Early resonance decay	-183		0.6%	
Color reconnection models	-128		0.1%	
Total reducible systematic unc.	652	-574	1.1%	-1.2%
Total irreducible systematic unc.	202	-255	0.6%	-0.1%
Total systematic unc.	683	-628	1.2%	-1.2%
Expected statistical unc.	$\pm$ 106		$\pm$ 0.1%	
Expected total unc.	691	-637	1.2%	-1.2%

Table 6.3.: Uncertainty table

6.2. Reducing the systematic uncertainty: ReSYST

Systematic uncertainties are estimated using the effect induced on an estimator evaluated on an ensemble of events. For example by changing the modeling parameter in the fragmentation of simulated $t\bar{t}$ events and calculating the effect this has on the estimator. In the case of the likelihood based estimate for the top quark mass considered in this study, the systematic effect on the ensemble of events is the sum over the systematic effect on the likelihood of each event. These conditions allows us to employ the ReSYST method [16] to quantify the systematic impact of individual events. In this section the ReSYST method will be briefly reviewed, and it will then be deployed to make an additional neural network based event selection criteria with the aim of reducing the systematic uncertainties on the top quark mass estimator.

6.2.1. The ReSYST Method

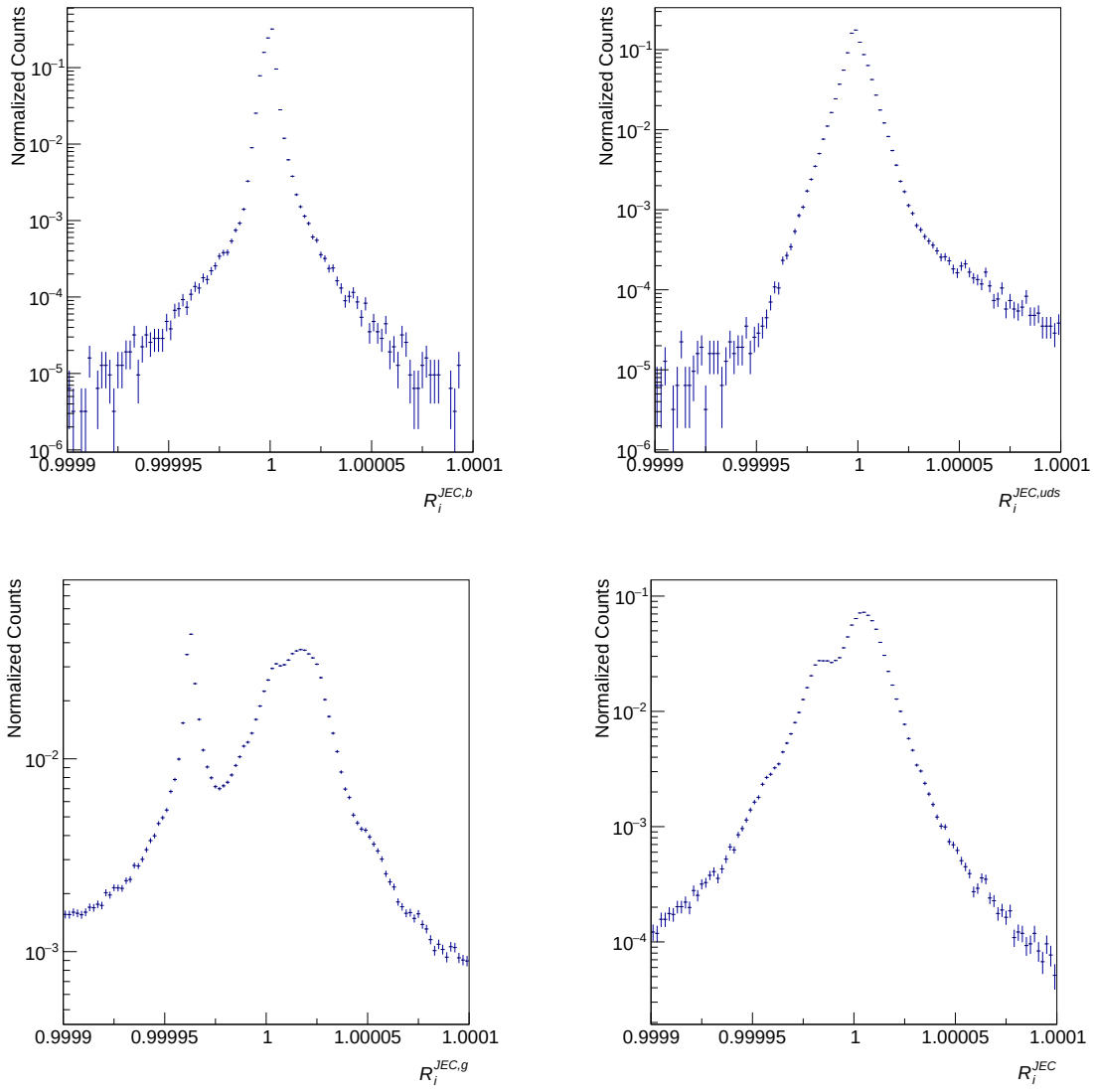
In the ReSYST method a generator level event variable is constructed which is correlated with the systematic impact of each individual event in the ensemble. This is in order to quantify the contribution each individual event has in the evaluation of a specific systematic uncertainty of the ensemble. This is done by removing a given event from the ensemble, and then calculating the impact on the total systematic uncertainty. The ReSYST quantifier is shown in Equation 6.6.

$$R_i = \frac{\sqrt{\sum_j (m_{t,(i)}^{\sigma_j} - m_{t,(i)})^2}}{\sqrt{\sum_j (m_t^{\sigma_j} - m_t)^2}} \tag{6.6}$$

Here m_t is the measured top quark mass of the ensemble, $m_t^{\sigma_j}$ is the measured top quark mass with systematic variation j applied, $m_{t,(i)}$ is the measured top quark mass after removing event number i and $m_{t,(i)}^{\sigma_j}$ is the measured top quark mass with systematic variation j applied after removing event i. Therefore the denominator in Equation 6.6 is the magnitude of the total systematic uncertainty for the full event ensemble, whereas the numerator is the magnitude of the total systematic uncertainty after event i is removed from the event ensemble. The R_i quantity will be 1 for events that have no impact on the systematic uncertainty when they would be removed from the ensemble, it will be greater than 1 for events where removing the event increased the total systematic uncertainty, and it will be smaller than 1 for events where removing

the event decreased the total systematic uncertainty. Therefore, if an event selection criterion can be identified that removes events with R_i value smaller than 1, the total systematic uncertainty should decrease. Since the R_i quantity is non-observable, it is required to find observable features of the event that are correlated with it, such that an event selection criterion can be constructed.

Figure 6.9.: The R_i distributions are shown for the JEC flavour uncertainties for bottom quarks (top left), light quarks (top right), gluons (bottom left) and the combined JEC flavour uncertainty (bottom right).



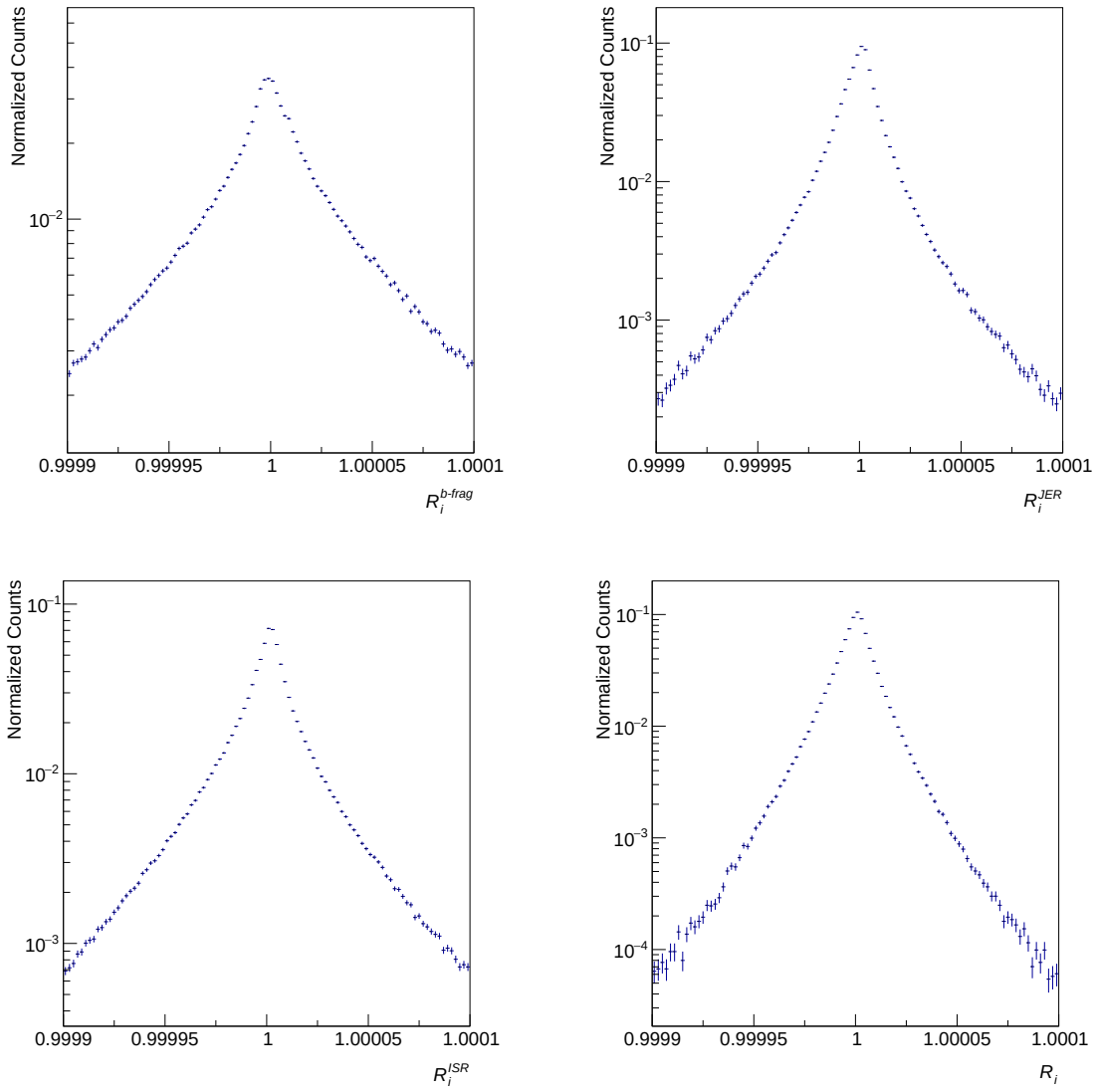
The distribution of the R_i quantifier can be seen for the JEC flavour uncertainties in Figure 6.9. The quantifier distributions for the light quark (uds) and bottom quark JEC

uncertainty is fairly symmetric around 1. The R_i quantifier distribution for the gluon JEC uncertainty has a multi-peak shape, since, unlike for light quarks and bottom quark, events can have a varying number of gluon induced jets present among the four leading jets in the event. Therefore a peak exists for events that have no gluons among the leading jets, a peak exists for events where the gluon jet is associated with the top quark, and a peak exists for events where the gluon is associated with the W boson jets. The R_i quantifier distribution for the combined JEC uncertainty is constructed by taking a linear sum for each JEC flavour uncertainty, $\sigma_{\text{JEC}} = \sum_j m_t^{\sigma_j} - m_t$, and then calculate the quantifier with this systematic uncertainty. The distribution for the combined JES quantifier has elements of each flavour component distribution.

The R_i quantifier can be seen for the Bowler-Lund b fragmentation uncertainty, the jet energy resolution uncertainty and the ISR uncertainty in Figure 6.10. As these uncertainties have an impact on every event in the sample a single peak structure is seen. The R_i quantifier for the full set of uncertainties is also shown in Figure 6.10, where it is seen to be fairly symmetric around 1.

The R_i quantity can only be constructed for systematic uncertainties where it is possible to link an event in the nominal event sample with the same event in the systematic uncertainty induced event sample. For example, it is possible to construct R_i for the jet energy correction uncertainties because in the nominal event sample, an event can be matched to the same event in the systematic uncertainty induced sample, where the jet energies have been scaled by an additional factor. Similarly, it is possible to construct R_i for the systematic uncertainties that are quantifiable using an event weight. It is not possible for the systematic uncertainties that require an independent event sample, where individual events cannot be matched to the nominal sample, such as the matrix element generator uncertainty, where MadGraph is compared to POWHEG. Accordingly, only those systematic uncertainties that can relate individual events between the nominal and systematic ensemble are considered in the calculation of R_i. The systematic uncertainties that are included are referred to as reducible systematic uncertainties. These are the jet energy resolution, pileup, b-tagging, JEC, JEC Flavour, b-jet modeling, ISR, FSR and top quark p_t modeling uncertainties. The systematic uncertainties where it is not possible to calculate R_i are referred to as irreducible systematic uncertainties. These are the matrix element generator, ME/PS matching, underlying event, early resonance decay and the color reconnection uncertainties.

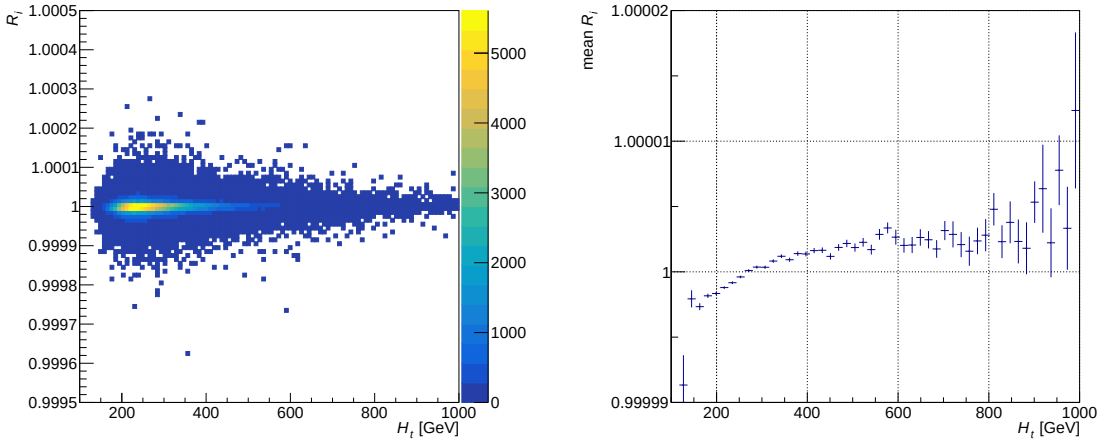
Figure 6.10.: The R_i distributions are shown for the Bowler-Lund b fragmentation uncertainty (top left), the jet energy resolution uncertainty (top right), the ISR uncertainty (bottom left) and the full combination of all uncertainties (bottom right).



6.2.2. Mono-variable approach of ReSYST

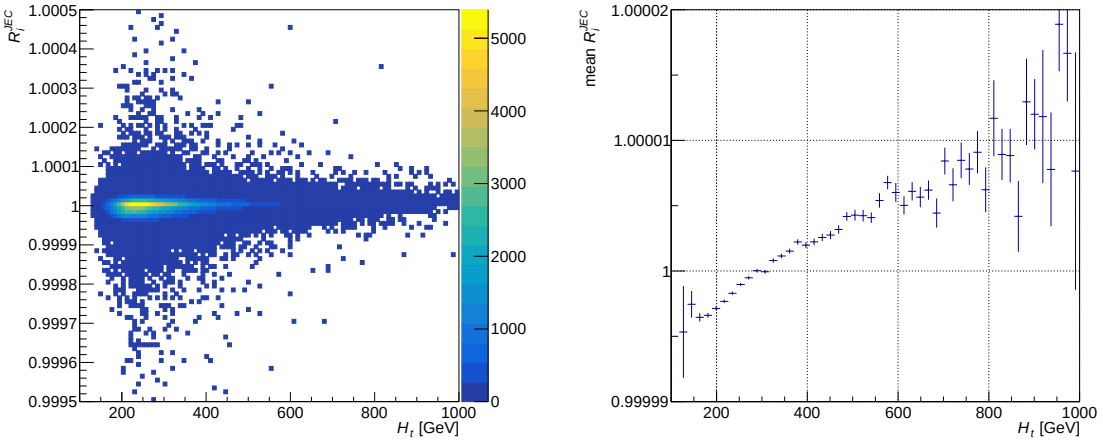
The ReSYST quantifier, R_i, is a generator level variable. Therefore it is required to correlate it to observable variables of the event. Such a variable can be found by inspecting the profile histogram of the ReSYST quantifier against one observable variable of the event. If the mean of the ReSYST quantifier in a specific observable variable bin is lower than 1, removing the events of that bin should reduce the systematic uncertainty on the top quark mass estimator.

Figure 6.11.: The ReSYST quantifier, R_i, targeting the full set of systematic uncertainties is shown against the variable H_t in the left plot. In the right plot the profile is shown.



Shown in Figure 6.11 is the ReSYST quantifier plotted against the scalar sum of the transverse momenta of jets in the event with $p_t > 30$ GeV, H_t, which has a relatively strong correlation with the R_i quantifier, compared to other event variables. From the profile plot a trend can be seen showing an increasing value of R_i for increasing H_t, and the mean R_i crosses 1 at roughly $H_t = 250$ GeV. This indicates that an additional event selection criteria on H_t, i.e requiring events with $H_t > 250$ GeV, should lower the total systematic uncertainty on the top quark mass estimator.

Figure 6.12.: The ReSYST quantifier targeting only the JEC flavour uncertainty, R_i^{JEC}, is shown against the variable H_t in the left plot. In the right plot the profile is shown.



In order to study the methodology, a ReSYST quantifier targeting a limited number of systematic uncertainties can be constructed. Shown on Figure 6.12 is the ReSYST quantifier targeting only the JEC flavour uncertainty, R_i^{JEC}, plotted against the H_t variable. From the profile plot it can be seen that the R_i^{JEC} variable is particularly connected with the H_t variable, as the slope of the profile plot is steeper than for the R_i quantifier with the full set of systematic uncertainties, indicating a stronger correlation for the JEC flavour uncertainty than for the combined set of systematic uncertainties. Therefore, a cut on H_t should particularly affect the JEC flavour uncertainty.

Two additional event selection cuts are studied in order to demonstrate the general ReSYST quantifier behaviour. To check the H_t dependence on the systematic uncertainties an $H_t > 250$ GeV event selection requirement was applied as well as an $H_t < 250$ GeV requirement. Based on the findings from Figure 6.11, these selections should result in a respectively smaller and larger systematic uncertainty compared to the nominal selection. When including these additional criteria in the event selection it is necessary to remake the templates distributions from which the event likelihoods are calculated as well as to obtain new top quark mass calibration curves, resulting in a top quark estimator with new statistical and systematic properties. A comparison of the template distributions with the additional selection requirements can be seen in Figure 6.13, and the new calibration curves can be seen in Figure 6.14 and 6.15. The new estimators looks to be stable and robust.

In Table 6.4 the different components of the systematic uncertainty after requiring the additional event selection criterion $H_t > 250$ GeV are listed. A reduction of the total systematic uncertainty of roughly 15% is observed compared to the nominal event selection. This arises due to some large systematic components being reduced. The JEC flavour uncertainty is reduced from $\sim$350 MeV to $\sim$200 MeV, the ISR uncertainty is reduced from $\sim$370 MeV to $\sim$290 MeV and the jet energy resolution is decreased from $\sim$330 MeV to $\sim$260 MeV. It is worth noting that the reduction of the total systematic uncertainty would be slightly larger if not for the irreducible uncertainties, which by coincidence increase with the additional selection criterion from $\sim$200 MeV to $\sim$300 MeV. Similarly, in Table 6.5 the different components of the systematic uncertainty after requiring the additional event selection criterion $H_t < 250$ GeV are listed. The total systematic uncertainty is much higher for this event selection confirming the expectation observed in Figure 6.11. The JEC Flavour uncertainty is particularly affected increasing to 570 MeV from its previous 350 MeV value, which is in agreement with the expectation observed in Figure 6.12.

Figure 6.13.: Distributions with fitted template functions for different generator top quark mass values with JSF fixed to 1.00. The top row shows the fitted top quark mass distribution for correct permutations for generator top quark masses of 166.5 GeV(left), 172.5 GeV (center) and 178.5 GeV (right) with an event selection requirement of $H_t < 250$ GeV . The bottom row shows the fitted top quark mass distributions for the same samples with an event selection requirement of $H_t > 250$ GeV.

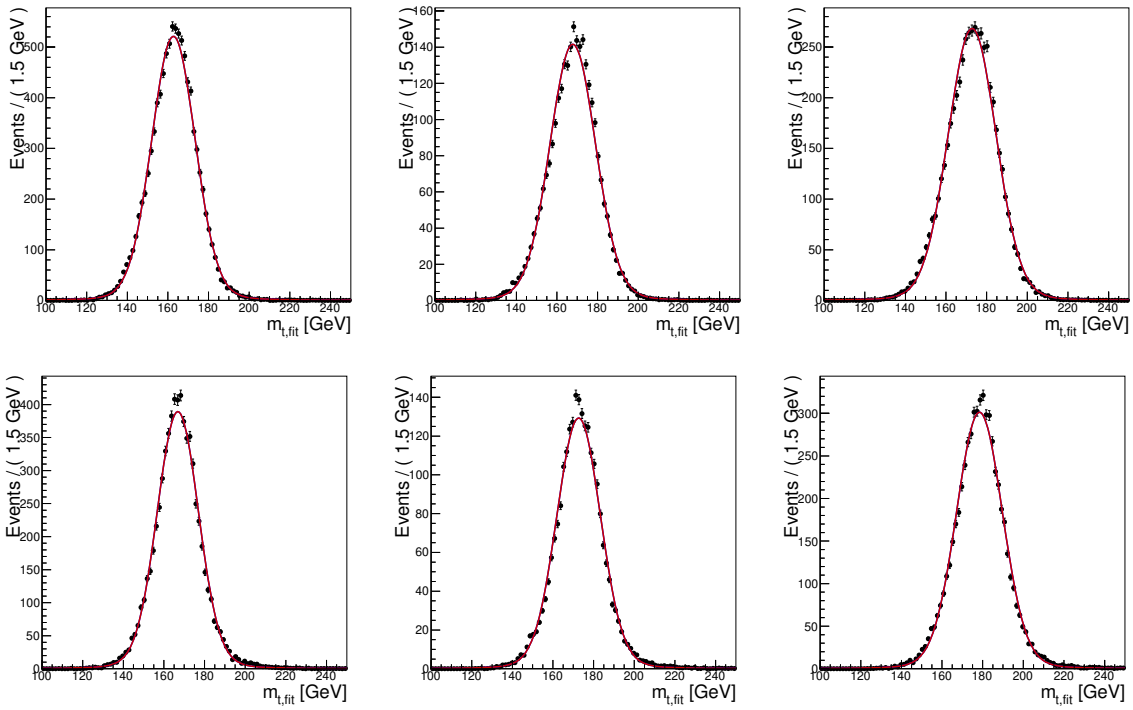


Figure 6.14.: Biases are shown on the top quark mass estimator with the additional $H_t > 250$ GeV selection criteria as a function of the generated mass after the calibration procedure. The biases are shown for $\mathrm{JSF}_{\mathrm{gen}}$ values of 0.96, 1.00 and 1.04. For comparison shaded bands indicating the mass biases with the nominal event selection are shown.

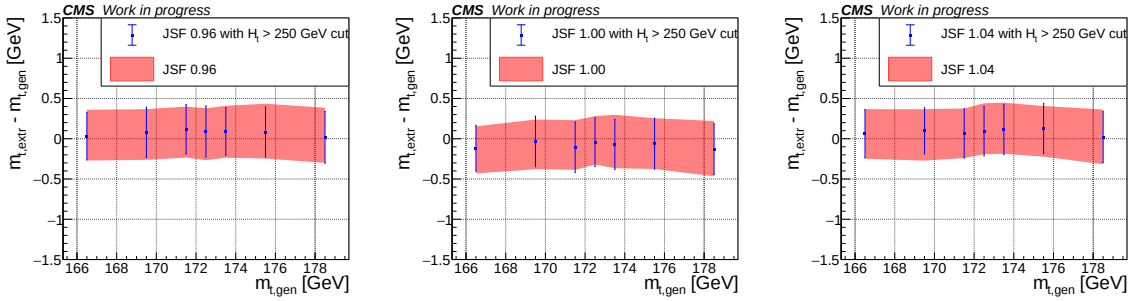
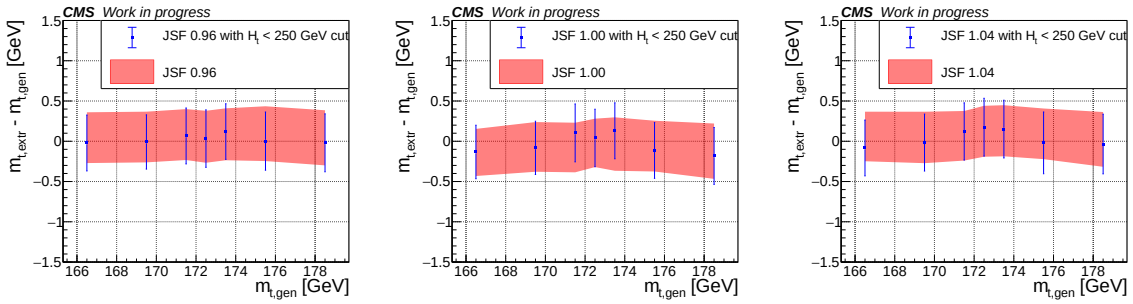


Figure 6.15.: Biases are shown on the top quark mass estimator with the additional $H_t < 250$ GeV selection criteria as a function of the generated mass after the calibration procedure. The biases are shown for $\mathrm{JSF}_{\mathrm{gen}}$ values of 0.96, 1.00 and 1.04. For comparison shaded bands indicating the mass biases with the nominal event selection are shown.



Uncertainty source	δm_t 1st variation (MeV)	δm_t 2nd variation (MeV)	δJSF 1st variation	δJSF 2nd variation
Experimental uncertainties				
Jet Energy Corrections	-3	66	0.7%	-0.8%
Mass calibration	$\pm$ 52		$\pm$ -<0.1%	
Jet Energy Resolution	-253	274	0.5%	-0.5%
b-tagging	12	-7	<0.1%	-<0.1%
Pileup	-92	93	0.1%	-0.1%
Modeling of hadronization				
JEC Flavour (linear sum)	$\pm$ 201		$\pm$ 0.4%	
-light quarks	*-350*		*0.4%*	
-bottom quarks	*544*		*-<0.1%*	
-gluon	*46*		*-<0.1%*	
-charm quarks	*-39*		*0.1%*	
b-jet modeling (squared sum)	114	-171	-<0.1%	-<0.1%
-Bowler-Lund fragmentation	*-151*	*84*	*-<0.1%*	*-<0.1%*
-Peterson fragmentation	*-49*		*-<0.1%*	
-Semileptonic B hadron decays	*77*	*-63*	*-<0.1%*	*-<0.1%*
Modeling of pertubative QCD				
Matrix element generator	175		-<0.1%	
ME/PS matching	-70	-5	-<0.1%	-0.1%
ISR parton shower scale	290	1	-0.3%	-<0.1%
FSR parton shower scale	-185	-<1	0.2%	-<0.1%
Top quark p_t uncertainty	114		-0.1%	
Modeling of soft QCD				
Underlying event	-116	-31	0.1%	-<0.1%
Early resonance decay	-225		0.5%	
Color reconnection models	-190		0.1%	
Total reducible systematic unc.	488	-420	1.0%	-1.1%
Total irreducible systematic unc.	183	-330	0.5%	-0.1%
Total systematic unc.	521	-534	1.1%	-1.1%
Expected statistical unc.	$\pm$ 124		$\pm$ 0.1%	
Expected total unc.	536	-548	1.1%	-1.1%

Table 6.4.: Overview of the systematic uncertainties after imposing the ReSYST guided selection criterion of requiring $H_t > 250$ GeV on the events.

Uncertainty source	δm_t 1st variation (MeV)	δm_t 2nd variation (MeV)	δJSF 1st variation	δJSF 2nd variation
Experimental uncertainties				
Jet Energy Corrections	-15	-3	1.0%	-0.9%
Mass calibration	$\pm$ 59		$\pm$ 0.1%	
Jet Energy Resolution	-471	505	0.5%	-0.6%
b-tagging	32	-30	<0.1%	-<0.1%
Pileup	-115	95	0.1%	-0.1%
Modeling of hadronization				
JEC Flavour (linear sum)	$\pm$ 569		$\pm$ 0.7%	
-light quarks	*-428*		*0.6%*	
-bottom quarks	*1029*		*-<0.1%*	
-gluon	*28*		*-<0.1%*	
-charm quarks	*-60*		*0.1%*	
b-jet modeling (squared sum)	159	-229	-<0.1%	-<0.1%
-Bowler-Lund fragmentation	*-207*	*119*	*-<0.1%*	*-<0.1%*
-Peterson fragmentation	*-39*		*-<0.1%*	
-Semileptonic B hadron decays	*105*	*-88*	*-<0.1%*	*-<0.1%*
Modeling of pertubative QCD				
Matrix element generator	491		0.1%	
ME/PS matching	-48	39	0.1%	-<0.1%
ISR parton shower scale	533	-5	-0.7%	-<0.1%
FSR parton shower scale	-306	31	0.4%	-<0.1%
Top quark p_t uncertainty	-28		-<0.1%	
Modeling of soft QCD				
Underlying event	-39	9	-<0.1%	-0.1%
Early resonance decay	-145		0.7%	
Color reconnection models	115		0.1%	
Total reducible systematic unc.	947	-840	1.4%	-1.5%
Total irreducible systematic unc.	509	-168	0.8%	-0.1%
Total systematic unc.	1075	-857	1.6%	-1.5%
Expected statistical unc.	$\pm$ 181		$\pm$ 0.2%	
Expected total unc.	1091	-876	1.6%	-1.5%

Table 6.5.: Overview of the systematic uncertainties after imposing the selection criterion of requiring $H_t < 250$ GeV on the events.

6.2.3. Caveats

Several aspects need to be monitored and controlled to successfully apply the ReSYST method. The first important point is that an event selection can reduce the overall magnitude of a systematic variation by making the estimator itself statistically less sensitive to the generated top quark mass, i.e. increasing the statistical variance of the estimator. There is no term in the ReSYST quantifier that ensures the estimator's variance remains unchanged. Therefore, it is needed to carefully monitor that a given ReSYST inspired event selection does not simply decrease the estimator's statistical sensitivity. When using a neural network to predict R_i from a larger set of observable event variables, it is useful to exclude the fitted top quark mass from this set, as a neural network can use this variable to break the estimators sensitivity to the generator top quark mass.

Another caveat is caused by the fact that the ReSYST quantifier R_i is calculated on an event by event basis. For a given systematic variation removing a single event from the ensemble might reduce the systematic uncertainty making it slightly smaller, and in order to proceed, the quantifier R_i would have to be reevaluated for all remaining events with new values in the denominator. Optimally, one would proceed by identifying the single event in the ensemble with the smallest R_i value. After removing this event from the ensemble, the systematic uncertainty of the estimator is reduced and one would proceed by recalculating R_i for all remaining events. However, this is computationally not feasible. When events are removed from the sample, the magnitude of the systematic uncertainty can change, and therefore it can also change which systematic uncertainties are the dominant contribution in the total systematic uncertainty. For systematic uncertainties where it is possible to define a selection that puts the systematic uncertainty to zero, you can reach a cross over point, where removing additional events from the ensemble will increase the systematic uncertainty, this time with the opposite sign. A solution that can address this, is to remove events in smaller batches, before any of the uncertainties reach a "swing-over" point, and then recalculate the ReSYST quantifier to target the new uncertainties that become dominant. A different and perhaps more advisable approach, is that by employing a more simple neural network, it appears that no uncertainties ever reach a swing-over point. The downside is that significantly more statistics have to be removed in order to achieve the same reduction of systematic uncertainties.

A final caveat is that not all uncertainties can be targeted by the ReSYST method.

Systematic uncertainties that are determined using separate event samples where an event to event mapping cannot be made between the nominal sample and the alternate sample cannot be used in the R_i quantifier, for instance matrix element uncertainties.

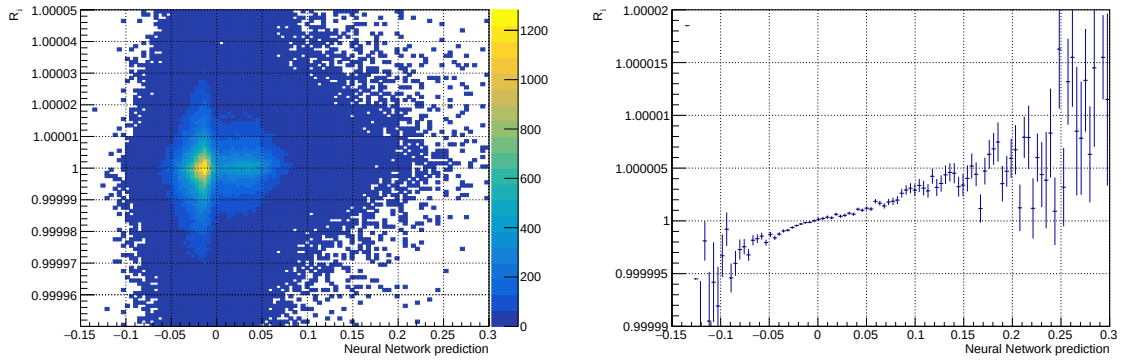
6.2.4. Multi-variable approach of ReSYST: Neural Network

In order to exploit the ReSYST quantifier optimally, it is needed to find an observable that is strongly correlated with the generator level R_i quantifier. This can be done by performing regression with a neural network. A too large and deep neural network architecture is to be avoided. As mentioned among the caveats it can easily result in a multi-variable observable that is strongly correlated with the fitted top quark mass itself, despite using input variables that make it difficult to calculate the fitted top quark mass. A powerful network would learn that by removing events that provide a lot of discriminating power for the top quark mass itself, the systematic uncertainty variations would become small, given that the estimator had lost its sensitivity to the generator top quark mass. A more simple network will however just combine the input variables that are correlated with the ReSYST quantifier in an optimal way, which already provides opportunity for a good reduction of the systematic uncertainties. To ensure no potential statistical bias, the training sample is kept separate from the sample that is used to calculate the final systematic uncertainty magnitudes.

The architecture used consists of three fully connected layers with 75 nodes each, however other machine learning methods such as BDTs or SVMs would be viable as well. Dropout of 0.5 is used in between the layers to prevent overtraining. Additionally, ReSYST scores for single events that are significant outliers are removed from the training sample, accounting for less than 1% of the training sample. The activation function used throughout is ReLU [128] and the loss function is the mean squared error. In order to easier predict the ReSYST quantifier with a neural network, the R_i quantity is re-scaled by subtracting it with its mean and dividing the result by the standard deviation of the R_i distribution. This is done since the neural network can struggle with predicting scores in the training process that only have very small numerical difference. Around 35 input event variables are used, provided to the network in a flat structure. These mainly consist of kinematic variables of the four leading jets as well as a few event level variables. The full list can be found in Appendix B. The performance of the network is evaluated by comparing the neural network output score with the

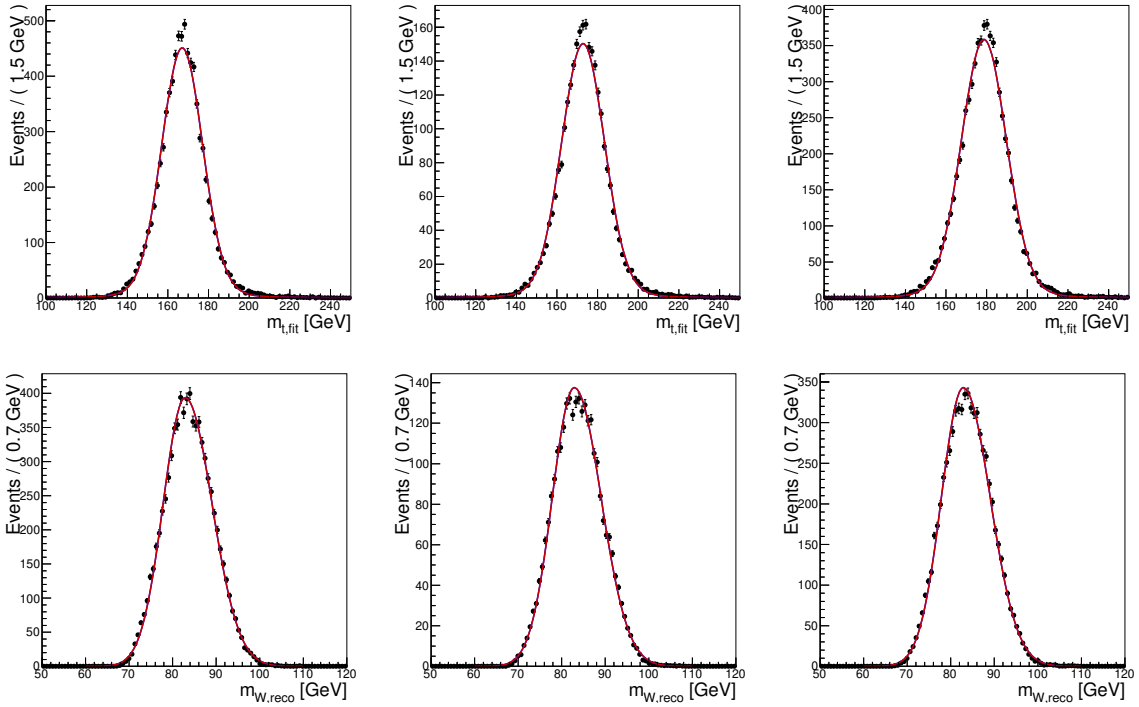
ReSYST quantifier. From Figure 6.16 it can be seen that there is a correlation between the neural network output and the ReSYST quantifier, which is much stronger than observed in any mono-variable approach. From the profile histogram (right plot in Figure 6.16) it can be seen that the neural network prediction is smoothly increasing towards larger values of R_i, indicating that the event selection threshold can be made for any given value. Larger networks with more inputs can easily get a stronger correlation with R_i however due to the caveats described in subsection 6.2.3, it was difficult to define a proper event selection criterion.

Figure 6.16.: A 2D histogram (left) itself of the neural network output score is showed against the ReSYST quantifier for all reducible systematic uncertainties, and a profile histogram (right) showing the mean of the ReSYST quantifier as a function of the reduced neural network prediction score.



Based on the neural network prediction, an additional event selection is to be determined while avoiding the swing-over effect described in section 6.2.3. When using larger neural networks it was observed to be an essential procedure to construct the event selection by making a series of progressively more tight event selection cuts, and then stop when the total systematic uncertainty starts to increase. However, when limited to a smaller set of inputs and a smaller network size, this is not as important, as the swing-over effect is not prevalent. In this case removing events with the neural network prediction score $R_i^{\text{pred}} < 0.01$ was chosen, removing roughly 65% of the events in the ensemble. After the event selection has been updated the template distributions are remade to built up the likelihood estimator. As can be seen in Figure 6.17, the templates are modified compared to before applying the additional R_i cut. For example the mean of the distributions are shifted to higher values. The overall effect of the neural network based selection on the templates is comparable to that observed with the mono-variable $H_t > 250\,\text{GeV}$ event selection criterion. The ReSYST

Figure 6.17.: Distributions of $m_{t,\text{fit}}$ and $m_{W,\text{reco}}$ with the fitted template functions for different generated top quark mass values with JSF_{gen} fixed to 1.00. The top row shows the fitted top quark mass distributions for correct permutations for generated top quark masses of 166.5 GeV (left), 172.5 GeV (center) and 178.5 GeV (right) with the additional neural network event selection cut. The bottom row shows the reconstructed W mass distribution for the same samples.



cut also changes the fractions of permutation types as can be seen in Table 6.6. The fraction of correct permutations notably goes up by roughly 6%. As the H_t variable

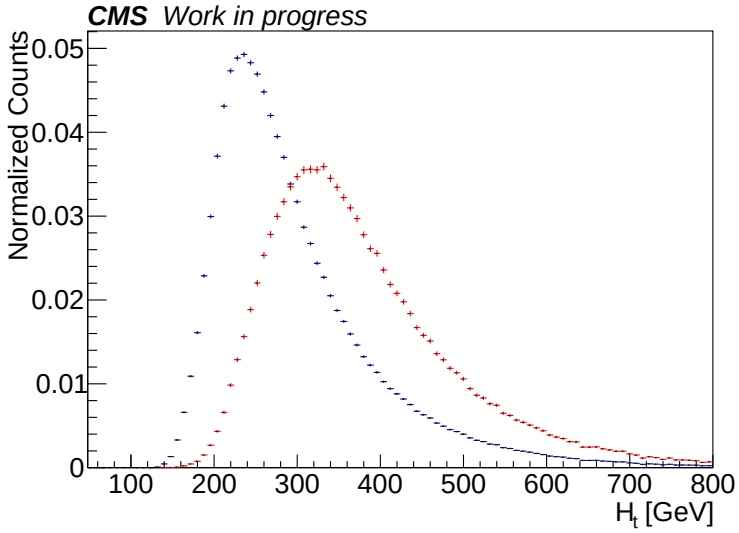
Permutation type	Nominal	$H_t > 250$ GeV	ReSYST
Correct	47%	49%	53%
Wrong	23%	22%	22%
Unmatched	30%	29%	25%

Table 6.6.: Fractions of the permutation types are shown with and without ReSYST neural network cut. The $H_t > 250$ GeV is also shown for comparison.

was the most correlated single variable in the neural network inputs, it is interesting to compare the distribution before and after the neural network ReSYST inspired event selection criterion is imposed. Figure 6.18 shows these H_t distributions. As can be

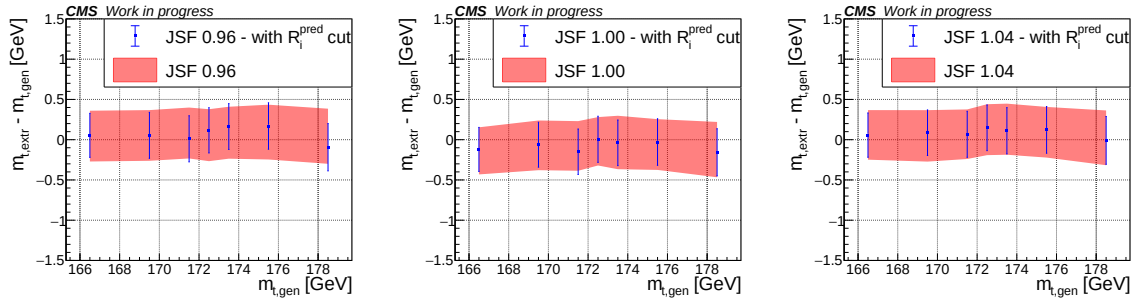
seen the H_t variable clearly is heavily correlated with the neural network score, which is to be expected since it was the input variable with the strongest correlation to R_i.

Figure 6.18.: The H_t distribution of the selected events before (blue) and after (red) the ReSYST inspired event selection criterion is applied.



After imposing the new event selection criterion, it is also required to perform a new calibration of the top quark mass estimator. Figure 6.19 shows the remaining biases after calibration, and compares them to the biases obtained with the nominal event selection. It can be clearly seen that the obtained biases on the top quark mass estimator are similar to those after the nominal selection.

Figure 6.19.: Biases are shown on the top quark mass estimator as a function of the generated top quark mass after the calibration procedure. Each plot shows the biases for a different $\mathrm{JSF}_{\mathrm{gen}}$ value. The biases are shown with the additional neural network based cut, and for comparison shaded bands indicate the mass biases with the nominal event selection.



After performing these secondary steps, the magnitude of the systematic uncertainties can be calculated. They are highlighted in Table 6.7. A significant improvement with respect to the $H_t > 250$ GeV event selection is observed. While 65% of the event are removed, the expected statistical uncertainty is still low compared to the systematic uncertainty, despite here only considering the muon channel and the size of the 2017 dataset. The ISR uncertainty is particularly reduced from 371 MeV to 95 MeV compared to the nominal event sample. Also the FSR uncertainty is reduced from -241 MeV to -67 MeV. The JEC flavour uncertainty is reduced from 355 MeV to 108 MeV, as an event selection that cancels out the contributions from different flavour has been found. The jet energy resolution uncertainty is lowered from 353 MeV to 184 MeV. The pileup uncertainty, the top quark p_t and the components of the b-fragmentation uncertainty are not reduced by much. It was verified that these uncertainties are possible to reduce using ReSYST, however since the magnitude of these uncertainties were small in comparison to the other uncertainties after the nominal event selection, they did not have a large impact on the ReSYST scores. If the ReSYST scores were recalculated with the new event selection criterion applied, it should be possible to target these as well. Compared to the $H_t > 250$ GeV event selection the irreducible systematic uncertainties happen to be slightly larger.

Uncertainty source	δm_t 1st variation (MeV)	δm_t 2nd variation (MeV)	δJSF 1st variation	δJSF 2nd variation
Experimental uncertainties				
Jet Energy Corrections	1	44	0.7%	-0.7%
Mass calibration	$\pm$ 53		$\pm$ <0.1%	
Jet Energy Resolution	-172	184	0.4%	-0.4%
b-tagging	13	-7	-<0.1%	-<0.1%
Pileup	-71	90	0.1%	-0.1%
Modeling of hadronization				
JEC Flavour (linear sum)	$\pm$ 108		$\pm$ 0.4%	
-light quarks	*-319*		*0.4%*	
-bottom quarks	*425*		*-0.1%*	
-gluon	*32*		*-<0.1%*	
-charm quarks	*-30*		*-<0.1%*	
b-jet modeling (squared sum)	82	-154	-<0.1%	-<0.1%
-Bowler-Lund fragmentation	*-131*	*70*	*-<0.1%*	*-<0.1%*
-Peterson fragmentation	*-66*		*-<0.1%*	
-Semileptonic B hadron decays	*43*	*-46*	*-<0.1%*	*-<0.1%*
Modeling of pertubative QCD				
Matrix element generator	72		0.1%	
ME/PS matching	-135	-70	0.1%	-0.1%
ISR parton shower scale	95	6	-0.1%	-<0.1%
FSR parton shower scale	-67	6	0.1%	-<0.1%
Top quark p_t uncertainty	164		-0.1%	
Modeling of soft QCD				
Underlying event	-173	-72	0.1%	-0.1%
Early resonance decay	-243		0.4%	
Color reconnection models	-202		0.2%	
Total reducible systematic unc.	314	-273	0.9%	-0.9%
Total irreducible systematic unc.	89	-389	0.5%	-0.0%
Total systematic unc.	326	-475	1.0%	-0.9%
Expected statistical unc.	$\pm$ 157		$\pm$ 0.1%	
Expected total unc.	362	-500	1.0%	-1.0%

Table 6.7.: Uncertainties after the ReSYST inspired neural network event selection.

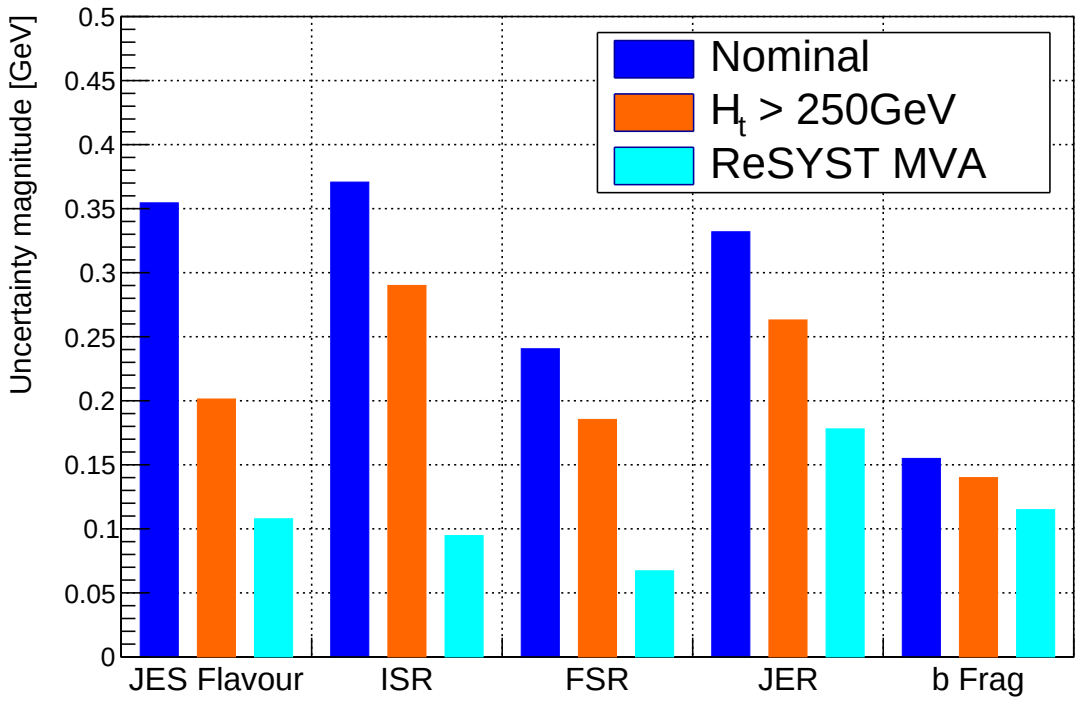
In the original selection the reducible systematic uncertainty is ± 691MeV and after applying the ReSYST inspired event selection criteria this value becomes ± 314MeV, yielding more than a 50% reduction, while removing 65% of the total number of events, which increases the statistical uncertainty to an acceptable level. Considering the slope of the profile histogram shown in Figure 6.16, it should be possible to reduce the systematic uncertainty further, but in the case of the top quark mass measurement, the ReSYST irreducible uncertainties seems to become the dominant uncertainty component, making further cuts pointless. Of the reducible systematic uncertainties some are more easily reduced with ReSYST. The ISR, FSR, jet resolution uncertainty, JEC flavour uncertainty are all significantly reduced. Presumably this is possible since there exist a phase space for the jets in these events, where these uncertainties have a small impact.

6.3. Conclusion and Outlook

Machine learning has not yet been used in particle physics as a tool to reduce systematic uncertainties of precise measurements of physical parameters. In this chapter the ReSYST method has been studied [16] in the context of a top quark mass measurement in the CMS experiment. It has been demonstrated that the ReSYST method is applicable in the context of a complex and realistic analysis setting. A mono-variable (H_t) event selection developed using the ReSYST quantifier, R_i, was shown to reduce the total systematic uncertainty on the top quark mass with 15% compared to the nominal event selection. This result was extended by predicting the ReSYST quantifier using a neural network approach, and then developing a new event selection reducing the contributions of systematic uncertainties targeted by ReSYST by about a factor of two, while removing 65% of the total number of events in the ensemble. An overview of the reduction of the largest systematic uncertainties can be seen on Figure 6.20. All these five uncertainties are reducible with the ReSYST method, and ReSYST inspired event selections can reduce their magnitudes. It is seen that by using a multi-variable machine learning approach it is possible to reduce the uncertainty magnitudes even further compared to the mono-variable approach.

One feature of the method is that certain uncertainties can swing-over with a more stringent event selection and thereafter start to increase as you remove more events, since the ReSYST quantifier is calculated from the full event sample based on the

Figure 6.20.: The magnitude of five of the largest systematic uncertainty sources on the top quark mass estimator are shown for the nominal event selection, an event selection with an additional requirement of $H_t > 250$ GeV, as well as an event selection based on a neural network prediction of R_i.



nominal event selection. This easily happens when employing larger neural networks. Alternative methods of combining the uncertainties were tried to mitigate this effect, such as using in the definition of the R_i variable the absolute value in a linear sum, instead of the square root of the squared sum. In the end, the squared sum approach seemed to be the most efficient to reduce the total systematic uncertainty, as it made the quantifier focus on the systematic uncertainties that were most important. For some systematic uncertainties that are estimated using an alternate event sample, referred to as irreducible systematic uncertainties, it is not possible to calculate an event by event ReSYST quantifier, because no event mapping can be made between the events in the nominal sample and the events in the alternate sample. An example of such an systematic uncertainty is the matrix element generator uncertainty. This was seen to put a limit on the ability to reduce further the total systematic uncertainty.

When applying the ReSYST method on real data collisions, it would be needed to

perform some additional checks to validate that the behaviour seen in simulation extends to data. The neural network prediction of R_i can be compared in simulation and data to ensure agreement, and potential corrections can be made accordingly to verify the impact of the mismatch on the estimation of the systematic uncertainties. A further useful study would be to examine the relative evolution of the measured top quark mass in bins of R_i^{pred} and examine if the measured top quark mass evolution is the same in data and simulation.

To improve the ReSYST method further, it might be possible to include the ReSYST irreducible uncertainties in the ReSYST quantifier by using a proxy weight for calculating the ReSYST quantifier R_i. Such a proxy weight could be constructed by training a neural network to discriminate on an event by event basis between the nominal event sample and the alternate sample. That would remove the limiting factor of the irreducible systematic uncertainties becoming the dominant contribution allowing to further reduce the systematic uncertainty.

A demonstration of the capabilities of the ReSYST method has been done for a top quark mass measurement focusing on the channel $pp \to t\bar{t} \to bqqb\mu\nu$. However, it is not expected that the method is limited to this measurement. It is possible to employ the ReSYST method in the top quark mass measurement in both the all-jets channel and the dilepton channel. Beyond the top quark mass measurements, any analyses that utilizes an event based likelihood with a systematic uncertainty component that is dominant with respect to the statistical component, could profit from the ReSYST method.

Chapter 7.

Conclusions and Outlook

In Chapter 4, a multiclass flavour tagging deep learning algorithm, DeepJet, was presented. It improves upon the traditional CMS heavy flavour tagging algorithm by utilizing more basic variables and employing a very loose input selection criteria. This contrasts previous heavy flavour identification algorithms by moving away from relying on high-level engineered features along with stringent selection criteria. This is made possible by using a new machine learning architecture that is capable of exploiting and processing the larger set of inputs. The jet flavour identification performance is improved in every part of the kinematic phase space of the jet and for all flavours. For b jet identification, the efficiency is relatively improved with $20 - 30\%$ for the same misidentification rate in simulation. The largest improvement is observed in the very high p_t jet region, which is the most challenging b jet identification phase space. Significant improvements are also obtained for c-tagging identification as well as quark-gluon discrimination, resulting in an all flavour jet identification algorithm. The performance of the tagger has been calibrated by the CMS b-tagging group using collision data for c-tagging and b-tagging. The results show that the performance gain observed in simulation translates to real collision data.

In Chapter 5, a neural network was designed to perform jet energy regression. Since different jet flavours have different jet energy responses, a similar network architecture and set of inputs to DeepJet was used. The network is able to estimate the jet flavour and can therefore propose a suitable energy correction. While traditional jet energy corrections are only parameterized in the p_t and η values of the reconstructed jets as well as some pileup variables, the neural network can additionally use the full jet shape and kinematics. The mean jet energy response improves when using the deep neural network. The jet energy resolution was also seen to improve relatively by roughly

10-20%. One hope of this approach would be to reduce the systematic uncertainty on the jet flavour energy response, as ideally this algorithm would be independent of the generator flavour labels, and instead utilize the full jet information. However, comparing the difference in jet energy response for different event generators yields very comparable numbers.

Many future studies can be performed to improve the DeepJet and DeepJet JEC algorithms. Interesting studies have been conducted for machine learning architectures for jet physics, highlighting promising approaches such as graphs and transformer networks. However, while the performance might improve with still $\sim 5\%$, a larger improvement will require more than an architecture switch. The IVF vertexing algorithm improved on the AVR algorithm by utilizing tracks beyond the jet cone for vertexing. Initial studies [167] have shown great promise for improving the performance of the b-tagger by incorporating a similar approach, allowing tracks beyond the jet cone among the inputs. Great care must however be taken not to leak event level information to the neural network, as this can bias calibration scale factors. For b jets at very high energy, interesting studies [168] have shown that instead of using tracks, better performance can be achieved by using pixel hit information directly. This could be incorporated as an additional input to DeepJet, and potentially improve the performance further at very high p_t. Both DeepJet and DeepJet JEC improved the performance by using more basic inputs, and allowing deep neural networks to select the important information. The approach builds on the philosophy that unnecessary processing done by physicist removes information from the inputs, and therefore will not perform as well compared to what can be done by a machine learning algorithm fully optimizing the process. However, both algorithms are still fully reliant on physicist-processed objects. Particle flow candidates are processed objects consisting of tracks and calorimeter clusters. Both of these are in turn processed objects from silicon tracker hits and energy deposits in calorimeter cells. It is clear that it would be possible to train algorithms on an even more basic set of inputs closer to the detector itself, which might improve performance significantly. The main challenges are simply that the lowest level data format consisting of the raw detector readouts is very impractical to work with directly. From a machine learning standpoint, due to the many different geometries of the CMS detector, additional care is needed to find a proper architecture to combine all this input information.

Beyond reconstruction, machine learning has also in recent years become a standard

in physics analysis for designing optimal event selection for improving the statistical measurement precision. However, the next step will be to enable it to reduce the systematic uncertainties that are increasingly becoming the dominant uncertainty component in several particle physics precision measurements. In Chapter 6, a neural network that performs an event selection designed to reduce systematic uncertainties in the context of a top quark mass measurement was presented, building on top of the ReSYST method [16]. The study provides a demonstration of the utility of ReSYST in a realistic and complex analysis as well as the capability of integrating machine learning methods. By using the ReSYST quantifier, R_i, as a guide, it was possible to design a mono-variable (H_t) event selection that was able to reduce the top quark mass total systematic uncertainty with 15% with respect to the nominal event selection. By utilizing a neural network to predict the ReSYST quantifier, and then using the neural network prediction to define an additional event selection criterion, it was possible to reduce the systematic uncertainties that were targeted by ReSYST by factor of ~ 2. The additional event selection criterion removes $\sim 65\%$ of the total number of events in the ensemble, increasing the statistical uncertainty by an acceptable level.

The ReSYST method shows great promise, however it is still in its infancy, and there is a need to further study its many potential benefits. The bottleneck for reducing the total systematic uncertainty even further with the ReSYST method, in the case of the top quark mass study, are systematic uncertainties that are not calculated utilizing the same simulated event sample, i.e. when no event to event map can be made between events in the nominal sample and the alternate uncertainty sample. This could be mitigated by using proxy weights that could be designed using neural networks. Additional improvements could be found by examining new potential input variables that are identified to be connected with a specific systematic uncertainty.

It is important to validate that the ReSYST quantifier behaves in a similar manner on real collision data as in simulation. However, presumably this will be more dependent on the chosen input variables to the neural network than the actual method itself.

The ReSYST method relies on scrutinizing the defined systematic uncertainties in order to identify the event selection that minimize them. However, it is not a given that all analyses systematic uncertainties are defined in a rigorous enough manner to allow for this procedure to be consistent. Some systematic uncertainties might in reality be correlated with specific observable variables, but at the analysis level it is applied as a flat correction across the full variable space. Clearly, the ReSYST method will not be

able to reduce such an uncertainty. However, more importantly is that the systematic uncertainty definitions reduced by the ReSYST method, are reverified to be an accurate estimate of the uncertainty. For instance in the case of the discussed JEC flavour uncertainty definition, it was shown in the study that it is possible to find a phase space where the jet energy response is similar for jets generated with Pythia and Herwig, making the currently defined systematic uncertainty small. However, of course the real quantity of interest, is the difference between the jet energy response of jets in real collision data and Pythia, and it is unclear if the ReSYST selection really coincides with that.

In conclusion, several new machine learning based methods were developed in this thesis that improve upon current methods that are used today, in order to conduct top quark physics at the CMS experiment at the LHC. These improvements were made possible by utilizing novel machine learning methods for defining analyses event selections as well as jet reconstruction. Significant advances were made in the ability to identify the flavour of the quark that initiated a reconstructed jet flavours. The ability to properly estimate the initial quark energy was also improved. Finally, using a novel machine learning method, it was shown to be possible to construct an optimized event selection for reducing the total systematic uncertainty in a top quark mass measurement.

Appendix A.

Input Variables: DeepJet and DeepJet JEC

The input variables are shown for the DeepJet algorithm described in Chapter 4. The DeepJet JEC algorithm described in Chapter 5 uses the same variables, except for a few additional variables that are marked in red.

A.1. List of global variables

- Jet p_t
- Jet η
- The number of charged particle flow candidates in the jet
- The number of neutral particle flow candidates in the jet
- The number of secondary vertices in the jet
- The number of primary vertices in the event
- Jet invariant mass
- Jet energy
- Jet ϕ

A.2. List of charged candidate variables

- Charged track η relative to the jet axis
- Charged track p_t relative to the jet axis
- Dot product of the jet and track momentum
- Dot product of the jet and track momentum divided by the magnitude of the jet momentum
- ΔR between the jet axis and the track
- The track 2D impact parameter value
- The track 2D impact parameter significance
- The track 3D impact parameter value
- The track 3D impact parameter significance
- The track distance to the jet axis
- Fraction of the jet momentum carried by the track.
- ΔR between the track and the closest secondary vertex
- An integer flag that indicate whether the track was used in the primary vertex fit.
- The charged candidates PUPPI weight
- χ^2 of the charged track fit.
- A integer flag which indicate the quality of the fitted track, based on number of detector hits used for the reconstruction as well as the overall χ^2 of the charged track fit.
- The sign of charge of the particle flow candidate
- Charged track ϕ relative to the jet axis
- Integer flag indicating if the track was flagged as an electron track
- Integer flag indicating if the track was flagged as a muon track

A.3. List of neutral candidate variables

- Fraction of the jet momentum carried by the neutral candidate
- ΔR between the jet axis and the neutral candidate
- A integer flag indicating whether the neutral candidate is a photon.
- Fraction of the neutral candidate energy deposited in the hadronic calorimeter.
- ΔR between the neutral candidate and the closest secondary vertex
- The neutral candidates PUPPI weight
- Neutral candidate η
- Neutral candidate ϕ

A.4. List of secondary vertex variables

- Secondary vertex p_t
- ΔR between the jet axis and the secondary vertex
- Secondary vertex mass
- Number of tracks in the secondary vertex
- χ^2 of the secondary vertex fit
- Reduced χ^2 of the secondary vertex fit
- The secondary vertex 2D impact parameter value
- The secondary vertex 2D impact parameter significance
- The secondary vertex 3D impact parameter value
- The secondary vertex 3D impact parameter significance
- Cosine of the angle between the secondary vertex flight direction and the direction of the secondary vertex momentum.
- Ratio of the secondary vertex energy to the jet energy

Appendix B.

Input Variables: ReSYST

B.1. Event Variables

- Missing transverse energy
- H_t
- Sum of E_t

B.2. Jet Variables

- Jet p_t
- Jet η
- Number of soft electrons in the jet
- Number of soft muons in the jet
- DeepJet b jet probability
- DeepJet c jet probability
- DeepJet quark jet probability
- DeepJet gluon jet probability
- DeepJet JEC correction

- DeepJet JEC resolution estimate

Appendix C.

DeepJEC Herwig Pythia

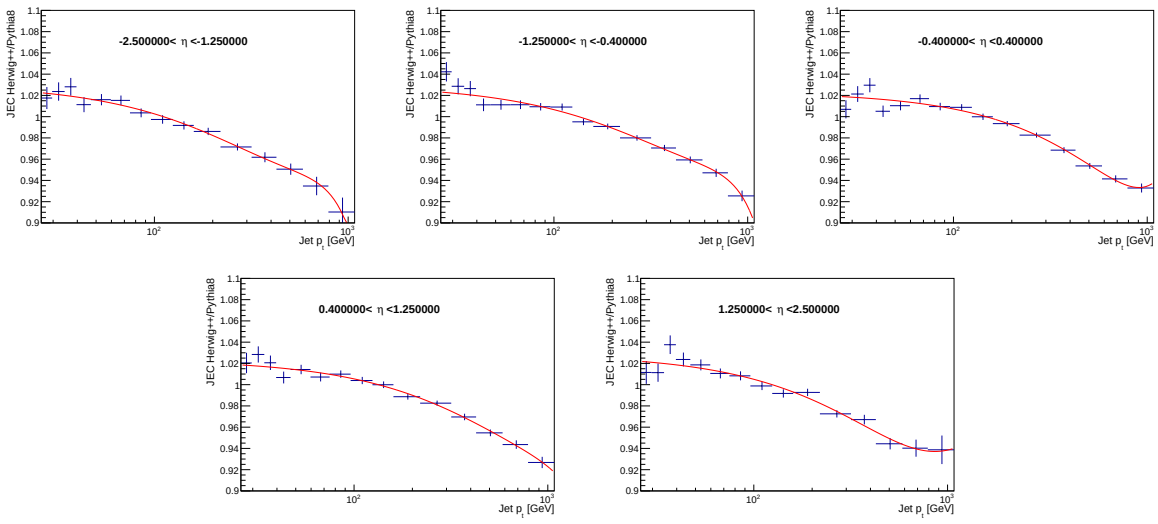


Figure C.1.: The ratio of JECs as calculated with b jets using QCD multijets samples with Herwig++ samples and QCD multijet samples.

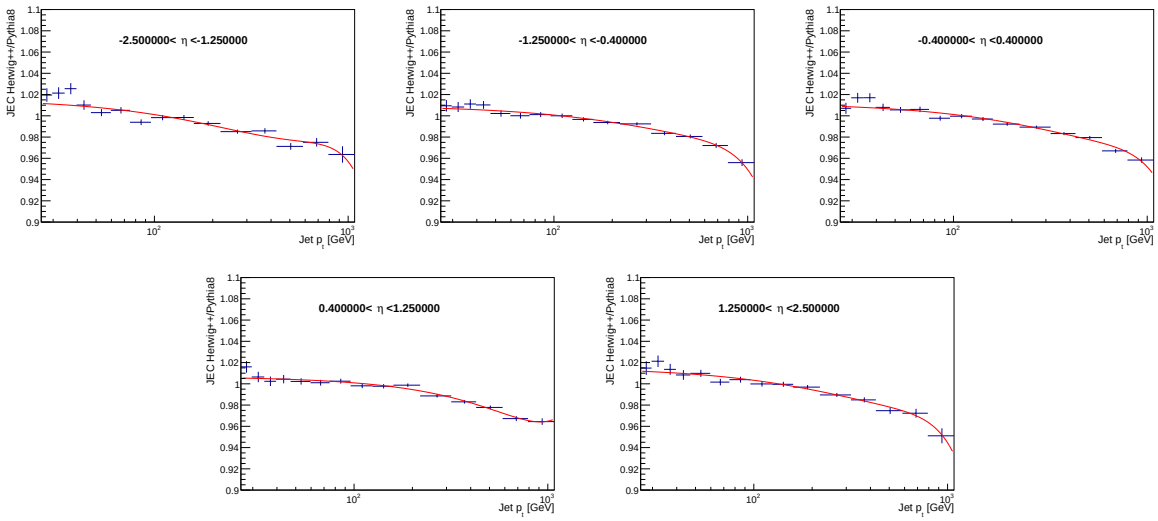


Figure C.2.: The ratio of JECs as calculated with c jets using QCD multijets samples with Herwig++ samples and QCD multijet samples.

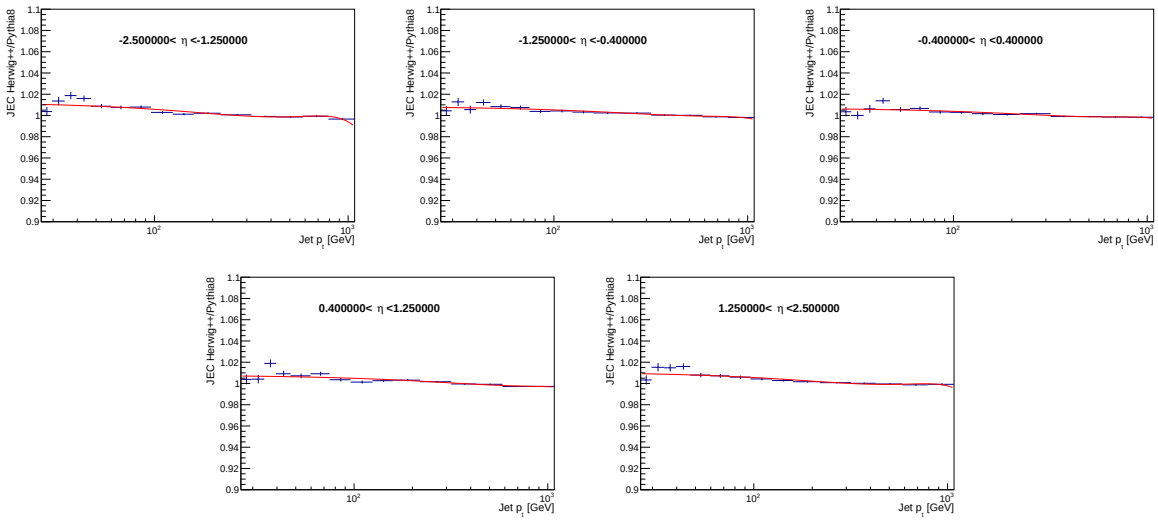


Figure C.3.: The ratio of JECs as calculated with uds jets using QCD multijets samples with Herwig++ samples and QCD multijet samples.

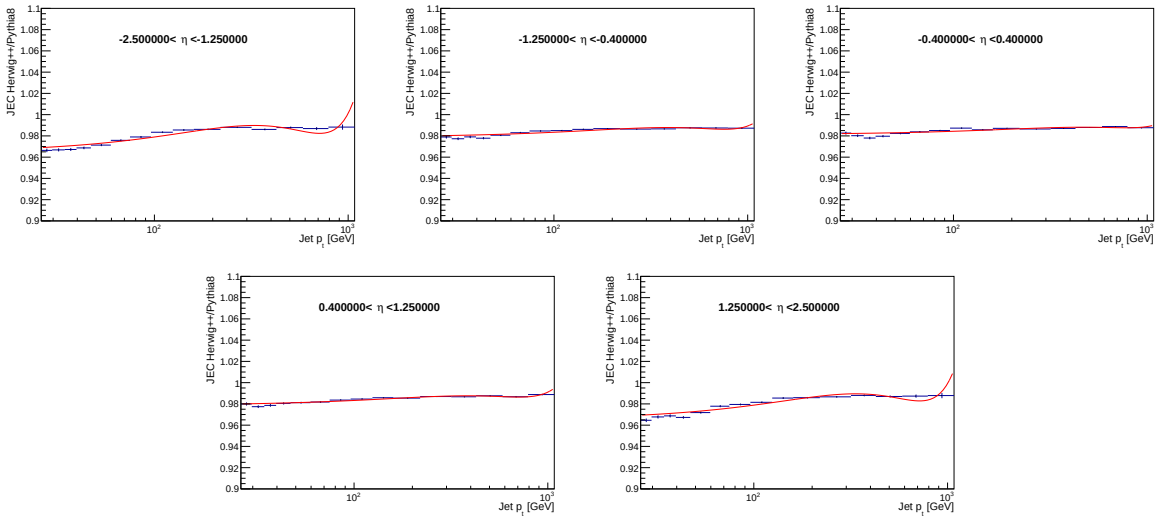


Figure C.4.: The ratio of JECs as calculated with g jets using QCD multijets samples with Herwig++ samples and QCD multijet samples.

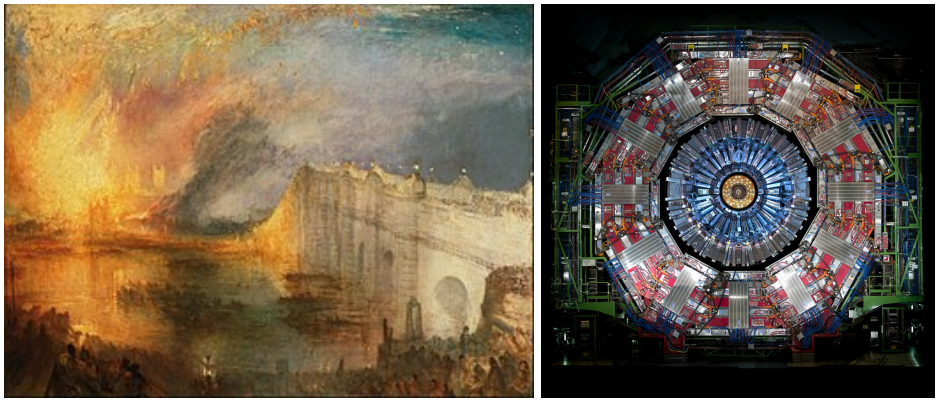
Appendix D.

Thesis cover

The thesis cover was generated using a neural network via the technique of style transfer. The method originates from Ref. [169]. It exploits the observation that an image recognition algorithm usually encodes the higher level concepts in early layers, and then examines more and more specific elements of the image. An input image is fed to a convolutional neural network, in this case the VGG-19 [170] image recognition algorithm is used. The network activations of one of the late convolutional layers are saved as an output that is meant to encode the content of the network. A style image is picked, which is then input to the same CNN and the network activations of an early convolutional layer is picked. The style representation is then made by building the Gram matrix from the vectorized feature map. Finally a new image with both the style and content can be synthesized by generating a new image that when input to the VGG-19 network, yields the same style and content representation. This image can be found via gradient descent.

In the case of the thesis, the style image used for the cover is J. M. W. Turners 'The Burning of the Houses of Lords and Commons', and the content image is the CMS detector at CERN. This can be synthesized into the image seen on the left in Figure D.2. The original image is in $\sim 400 \times 400$ resolution, and it is then scaled using the approach of Ref. [171]. This method is called Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN), and it is a convolutional neural network trained with low resolution images to generate the same image in high resolution. After applying the ESRGAN the picture for the cover is created in $\sim 2000 \times 2000$ resolution, as seen on the right. In the case of the back cover a illustration of the LHC was synthesized as shown in Figure D.3. Additional images were tried with the styles of different paintings. These can be seen in Figure D.4.

Figure D.1.: The style image for the cover is shown on the left, and the content image is shown on the right.



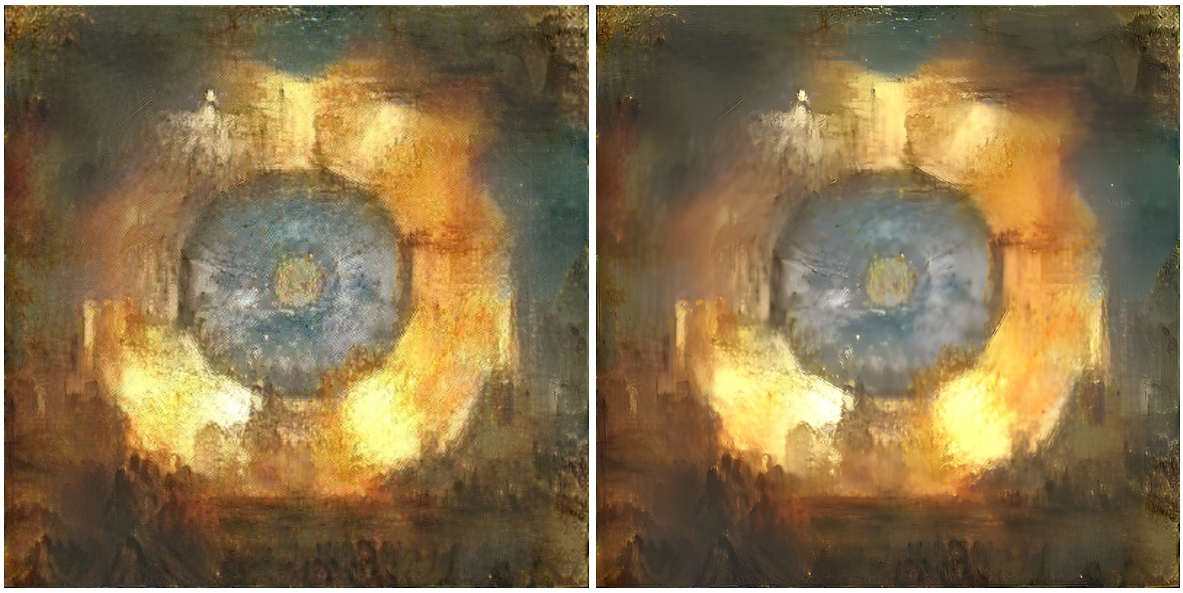


Figure D.2.: The synthesized image before and after being upscaled using the Real-ESRGAN neural network.

Figure D.3.: The original content image of the LHC is shown on the left. On the right a synthesized image with the style of the Turner painting "The Fighting Temeraire Tugged to Her Last Berth to Be Broken up" is shown.

Figure D.4.: The CMS detector with styles of Picasso (top left), Magritte (top right), Van Gogh (bottom left) and Delaunay (bottom right).

Author contributions

This text aims to emphasize the research in the thesis that the author directly contributed to, as well as to highlight additional research that was performed, but not included in the thesis.

The DeepJet algorithm described in Chapter 4 was published in Ref. [113] of which I am the main contact author. In connection with my involvement in research of flavour tagging algorithms, I served as CMS b-tagging and vertexing (BTV) software and algorithms convener from 2019-2021. In addition I have been functioning since 2020 as reconstruction contact for BTV as well as machine learning contact for both BTV and the CMS jet and missing ET group (JETMET).

I have performed additional work on the DeepJet algorithm not described in the thesis. Notably a retraining was made for PUPPI jets, and the algorithm was extended to tag pileup jets as well. Additionally, studies of feature importance ware performed using the technique of layerwise relevance propagation, and using this information feature pruning was performed.

I performed the work described in Chapter 5 on extending DeepJet to perform jet energy corrections as well as the work on improving and applying the ReSYST method described in Chapter 6.

Beyond the work shown in the thesis I participated in the assembly of new outer tracker modules of the 2S type for the CMS upgrade towards HL-LHC. Particularly I helped to develop and setup a metrology system for measuring the assembly precision.

Bibliography

[1] ATLAS, *Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc*, *Physics Letters B* **716** (2012) 1.

[2] CMS, *Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc*, *Physics Letters B* **716** (2012) 30.

[3] S. Weinberg, *A model of leptons*, *Phys. Rev. Lett.* **19** (1967) 1264.

[4] S. Willenbrock, *Symmetries of the standard model*, hep-ph/0410370.

[5] S. L. Glashow, J. Iliopoulos and L. Maiani, *Weak interactions with lepton - hadron symmetry*, *Phys. Rev.* **D2** (1970) 1285.

[6] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*, *Phys. Lett.* **12** (1964) 132.

[7] A. D. Sakharov, *Violation ofiCP/iin variance,ic/iasymmetry, and baryon asymmetry of the universe*, *Soviet Physics Uspekhi* **34** (1991) 392.

[8] and N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini et al., *Planck 2018 results*, *Astronomy & Astrophysics* **641** (2020) A6.

[9] Y. Fukuda, T. Hayakawa, E. Ichihara, K. Inoue, K. Ishihara, H. Ishino et al., *Evidence for oscillation of atmospheric neutrinos*, *Physical Review Letters* **81** (1998) 1562.

[10] Q. R. Ahmad, R. C. Allen, T. C. Andersen, J. D. Anglin, G. Bühler, J. C. Barton et al., *Measurement of the rate of* $nu_e + d \rightarrow p + p + e^-$ *interactions produced by 8b solar neutrinos at the sudbury neutrino observatory*, *Physical Review Letters* **87** (2001) .

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

[12] CMS collaboration, *The CMS Experiment at the CERN LHC*, *JINST* **3** (2008)

S08004.

[13] M. Kobayashi and T. Maskawa, *CP-Violation in the Renormalizable Theory of Weak Interaction*, *Progress of Theoretical Physics* **49** (1973) 652.

[14] S. Abachi, B. Abbott, M. Abolins, B. S. Acharya, I. Adam, D. L. Adams et al., *Observation of the top quark*, *Physical Review Letters* **74** (1995) 2632.

[15] F. Abe, H. Akimoto, A. Akopian, M. G. Albrow, S. R. Amendolia, D. Amidei et al., *Observation of top quark production in pbarp collisions with the collider detector at fermilab*, *Physical Review Letters* **74** (1995) 2626.

[16] P. van Mulders, *Resyst: a novel technique to reduce the systematic uncertainty for precision measurements*, *J. High Energ. Phys.* **132** (2019) [1809.07700].

[17] P. D. Group, P. A. Zyla, R. M. Barnett, J. Beringer, O. Dahl, D. A. Dwyer et al., *Review of Particle Physics*, *Progress of Theoretical and Experimental Physics* **2020** (2020) [https://academic.oup.com/ptep/article-pdf/2020/8/083C01/34673722/ptaa104.pdf].

[18] M. Thomson, *Modern particle physics*. Cambridge University Press, New York, 2013.

[19] M. J. Herrero, *The standard model*, 1998. 10.48550/ARXIV.HEP-PH/9812242.

[20] M. Srednicki, *Quantum field theory*. Cambridge University Press, 1, 2007.

[21] P. A. M. Dirac, *On the Theory of quantum mechanics*, *Proc. Roy. Soc. Lond. A* **112** (1926) 661.

[22] E. Noether, *Invariante variationsprobleme*, *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* **1918** (1918) 235.

[23] Wikimedia, *Coupling strength image*, https://commons.wikimedia.org/wiki/File:Running$_c$oupling$_c$onstants.svg.

[24] D. J. Gross and F. Wilczek, *Ultraviolet behavior of non-abelian gauge theories*, *Phys. Rev. Lett.* **30** (1973) 1343.

[25] H. D. Politzer, *Reliable perturbative results for strong interactions?*, *Phys. Rev. Lett.* **30** (1973) 1346.

[26] S. Borsanyi, S. Durr, Z. Fodor, C. Hoelbling, S. D. Katz, S. Krieg et al., *Ab initio*

calculation of the neutron-proton mass difference, *Science* **347** (2015) 1452.

[27] F. Englert and R. Brout, *Broken symmetry and the mass of gauge vector mesons*, *Phys. Rev. Lett.* **13** (1964) 321.

[28] P. W. Higgs, *Broken symmetries and the masses of gauge bosons*, *Phys. Rev. Lett.* **13** (1964) 508.

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

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

[31] J. Ellis, *Higgs Physics*, 1312.5672.

[32] J. Haller, , A. Hoecker, R. Kogler, K. Mönig, T. Peiffer et al., *Update of the global electroweak fit and constraints on two-higgs-doublet models*, *The European Physical Journal C* **78** (2018) .

[33] N. Cabibbo, L. Maiani, G. Parisi and R. Petronzio, *Bounds on the fermions and Higgs boson masses in grand unified theories*, *Nucl. Phys. B* **158** (1979) 295.

[34] G. Degrassi, S. D. Vita, J. Elias-Miró, J. R. Espinosa, G. F. Giudice, G. Isidori et al., *Higgs mass and vacuum stability in the standard model at NNLO*, *Journal of High Energy Physics* **2012** (2012) .

[35] A. Andreassen, W. Frost and M. D. Schwartz, *Scale-invariant instantons and the complete lifetime of the standard model*, *Physical Review D* **97** (2018) .

[36] A. H. Hoang, *What is the top quark mass?*, *Annual Review of Nuclear and Particle Science* **70** (2020) 225.

[37] A. H. Hoang, A. Jain, C. Lepenik, V. Mateu, M. Preisser, I. Scimemi et al., *The msr mass and the $\mathcal{O}\left(\Lambda_{\rm qcd}\right)$ renormalon sum rule*, *Journal of High Energy Physics* **2018** (2018) .

[38] M. Butenschoen, B. Dehnadi, A. H. Hoang, V. Mateu, M. Preisser and I. W. Stewart, *Top quark mass calibration for monte carlo event generators*, *Phys. Rev. Lett.* **117** (2016) 232001.

[39] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [1410.3012].

[40] CMS collaboration, *Measurement of differential tt ¯ production cross sections in the full kinematic range using lepton+jets events from proton-proton collisions at* $s = \sqrt{13}tev$, *Physical Review D* **104** (2021) .

[41] E. A. Mobs, *The CERN accelerator complex. Complexe des accélérateurs du CERN*, (Oct, 2016), http://cds.cern.ch/record/2225847.

[42] L. Arnaudon, P. Baudrenghien, M. Baylac, G. Bellodi, Y. Body, J. Borburgh et al., *Linac4 Technical Design Report*, tech. rep., CERN, Geneva, Dec, 2006.

[43] J.-P. Burnet, C. Carli, M. Chanel, R. Garoby, S. Gilardoni, M. Giovannozzi et al., *Fifty years of the CERN Proton Synchrotron: Volume 1*, CERN Yellow Reports: Monographs. CERN, Geneva, 2011, 10.5170/CERN-2011-004.

[44] ATLAS collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003. 437 p.

[45] CMS collaboration, *Cutaway diagrams of CMS detector*, https://cds.cern.ch/record/2665537.

[46] I. Neutelings, *Example spherical coordinates.*, https://wiki.physik.uzh.ch/cms/latex:example$_s$*pherical*$_c$*oordinates*.

[47] CMS collaboration, V. Karimäki, M. Mannelli, P. Siegrist, H. Breuker, A. Caner, R. Castaldi et al., *The CMS tracker system project: Technical Design Report*, Technical design report. CMS. CERN, Geneva, 1997.

[48] CMS collaboration, *The CMS tracker: addendum to the Technical Design Report*, Technical design report. CMS. CERN, Geneva, 2000.

[49] CMS collaboration, *CMS Technical Design Report for the Pixel Detector Upgrade*, Tech. Rep. CERN-LHCC-2012-016. CMS-TDR-11, CERN, Geneva, Sep, 2012. doi:10.2172/1151650.

[50] CMS Tracker Group of the CMS collaboration, *The CMS Phase-1 Pixel Detector Upgrade*, *JINST* **16** (2020) P02027. 84 p [2012.14304].

[51] CMS collaboration, *Track impact parameter resolution in the 2017 dataset with the CMS Phase-1 Pixel detector*, .

[52] CMS collaboration, *Description and performance of track and primary-vertex reconstruction with the CMS tracker*, *Journal of Instrumentation* **9** (2014) P10009.

[53] CMS collaboration, *The CMS electromagnetic calorimeter project: Technical Design Report*, Technical design report. CMS. CERN, Geneva, 1997.

[54] CMS collaboration, P. Bloch, R. Brown, P. Lecoq and H. Rykaczewski, *Changes to CMS ECAL electronics: addendum to the Technical Design Report*, Technical design report. CMS. CERN, Geneva, 2002.

[55] CMS collaboration, *CMS ECAL performance with 2017 data*, .

[56] CMS collaboration, *The CMS hadron calorimeter project: Technical Design Report*, Technical design report. CMS. CERN, Geneva, 1997.

[57] J. Mans, J. Anderson, B. Dahmes, P. de Barbaro, J. Freeman, T. Grassi et al., *CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter*, tech. rep., Sep, 2012.

[58] CMS collaboration, *Calibration of the CMS hadron calorimeters using proton-proton collision data at* $\sqrt{s} = 13$ *TeV*, *JINST* **15** (2019) P05002. 45 p [1910.00079].

[59] V. D. Elvira, *Measurement of the Pion Energy Response and Resolution in the CMS HCAL Test Beam 2002 Experiment*, .

[60] CMS collaboration, *Performance of CMS muon reconstruction in pp collision events at* $\sqrt{s} = 7$ *TeV*, *JINST* **7** (2012) P10002. 81 p [1206.4071].

[61] CMS collaboration, J. G. Layter, *The CMS muon project: Technical Design Report*, Technical design report. CMS. CERN, Geneva, 1997.

[62] CMS collaboration, G. L. Bayatyan, N. Grigorian, V. G. Khachatrian, A. T. Margarian and e. a. Sirunyan, A M, *CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems*, Technical design report. CMS.

[63] CMS collaboration, S. Cittolin, A. Rácz and P. Sphicas, *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. CERN, Geneva, 2002.

[64] CMS collaboration, *Particle-flow reconstruction and global event description with the CMS detector. Particle-flow reconstruction and global event description with the CMS detector*, *JINST* **12** (2017) P10003. 82 p [1706.04965].

[65] W. Adam, B. Mangano, T. Speer and T. Todorov, *Track Reconstruction in the CMS*

tracker, tech. rep., CERN, Geneva, Dec, 2006.

[66] W. Adam, R. Frühwirth, A. Strandlie and T. Todorov, *Reconstruction of electrons with the gaussian-sum filter in the CMS tracker at the LHC*, *Journal of Physics G: Nuclear and Particle Physics* **31** (2005) N9.

[67] CMS collaboration, *The performance of the cms muon detector in proton-proton collisions at $\sqrt{s} = 7$ tev at the lhc*, *Journal of Instrumentation* **8** (2013) P11002.

[68] T. Speer, K. Prokofiev, R. Frühwirth, W. Waltenberger and P. Vanlaer, *Vertex Fitting in the CMS Tracker*, tech. rep., CERN, Geneva, Feb, 2006.

[69] CMS collaboration, *Measurement of b anti-b angular correlations based on secondary vertex reconstruction at $\sqrt{s}$ = 7tev*, *Journal of High Energy Physics* **2011** (2011) .

[70] CMS collaboration, *Jet image*, https://cms.cern/news/jets-cms-and-determination-their-energy-scale.

[71] M. Cacciari, G. P. Salam and G. Soyez, *The Anti-k(t) jet clustering algorithm*, *JHEP* **0804** (2008) 063 [arXiv:0802.1189].

[72] CMS collaboration, *Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV*, *Journal of Instrumentation* **12** (2017) P02014.

[73] CMS collaboration, *Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s}$ = 13 tev using the cms detector*, *Journal of Instrumentation* **14** (2019) P07004.

[74] J. Collins, K. Howe and B. Nachman, *Anomaly detection for resonant new physics with machine learning*, *Phys. Rev. Lett.* **121** (2018) 241803.

[75] A. Andreassen, I. Feige, C. Frye and M. D. Schwartz, *Junipr: a framework for unsupervised machine learning in particle physics*, *The European Physical Journal C* **79** (2019) .

[76] J. Brehmer, S. Macaluso, D. Pappadopulo and K. Cranmer, *Hierarchical clustering in particle physics through reinforcement learning*, in *34th Conference on Neural Information Processing Systems*, 11, 2020, 2011.08191.

[77] W. Mcculloch and W. Pitts, *A logical calculus of ideas immanent in nervous activity*, *Bulletin of Mathematical Biophysics* **5** (1943) 127.

[78] F. Rosenblatt, *The perceptron: A probabilistic model for information storage and*

organization in the brain., *Psychological Review* **65** (1958) 386.

[79] A. B. Novikoff, *On convergence proofs on perceptrons*, in *Proceedings of the Symposium on the Mathematical Theory of Automata*, vol. 12, (New York, NY, USA), pp. 615–622, Polytechnic Institute of Brooklyn, 1962.

[80] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, USA, 1969.

[81] B. E. Boser, I. M. Guyon and V. N. Vapnik, *A training algorithm for optimal margin classifiers*, in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, ACM Press, 1992.

[82] K. Hornik, M. Stinchcombe and H. White, *Multilayer feedforward networks are universal approximators*, *Neural Networks* **2** (1989) 359.

[83] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning Representations by Back-propagating Errors*, *Nature* **323** (1986) 533.

[84] B. H. Denby, *Neural Networks and Cellular Automata in Experimental High-energy Physics*, *Comput. Phys. Commun.* **49** (1988) 429.

[85] C. Peterson, *Track Finding With Neural Networks*, *Nucl. Instrum. Meth. A* **279** (1989) 537.

[86] G. Stimpfl-Abele and L. Garrido, *Fast track finding with neural nets*, *Comput. Phys. Commun.* **64** (1991) 46.

[87] L. Lönnblad, C. Peterson and T. Rögnvaldsson, *Finding gluon jets with a neural trigger*, *Phys. Rev. Lett.* **65** (1990) 1321.

[88] L. Lönnblad, C. Peterson and T. Rögnvaldsson, *Using neural networks to identify jets*, *Nuclear Physics B* **349** (1991) 675.

[89] G. Aad, B. Abbott, J. Abdallah, S. Abdel Khalek, O. Abdinov, R. Aben et al., *Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s} = 7$ tev with the atlas detector*, *The European Physical Journal C* **74** (2014) .

[90] DELPHI, *Classification of the hadronic decays of the z0 into b and c quark pairs using a neural network*, *Physics Letters B* **295** (1992) 383.

[91] D0 collaboration, *Direct measurement of the top quark mass*, *Phys. Rev. Lett.* **79** (1997) 1197.

[92] J. K. Kohne et al., *Realization of a second level neural network trigger for the H1 experiment at HERA*, *Nucl. Instrum. Meth. A* **389** (1997) 128.

[93] P. Busson, R. Nóbrega and J. Varela, *Modular neural networks for on-line event classification in high energy physics*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **410** (1998) 273.

[94] B. Denby, *Neural networks in high energy physics: A ten year perspective*, *Computer Physics Communications* **119** (1999) 219.

[95] B. P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu and G. McGregor, *Boosted decision trees, an alternative to artificial neural networks*, *Nucl. Instrum. Meth. A* **543** (2005) 577 [physics/0408124].

[96] A. Krizhevsky, I. Sutskever and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds., vol. 25, Curran Associates, Inc., 2012, https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[97] V. Nair and G. E. Hinton, *Rectified linear units improve restricted boltzmann machines*, in *ICML*, 2010.

[98] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, *Improving neural networks by preventing co-adaptation of feature detectors*, *ArXiv* **abs/1207.0580** (2012) .

[99] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard et al., *Handwritten digit recognition with a back-propagation network*, in *Advances in Neural Information Processing Systems*, D. Touretzky, ed., vol. 2, Morgan-Kaufmann, 1989, https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf.

[100] IBM, *Convolutional neural networks image*, (2022), https://www.ibm.com/cloud/learn/convolutional-neural-networks.

[101] CMS, *Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques*, *Journal of Instrumentation* **15** (2020) P06005–P06005.

[102] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, *Neural Computation* **9** (1997) 1735.

[103] G. Chevalier, *Lstm image*, (2022), https://upload.wikimedia.org/wikipedia/commons/9/93/LSTM_Cell.svg.

[104] CMS, *Performance of deep tagging algorithms for boosted double quark jet topology in proton-proton collisions at 13 tev with the phase-0 cms detector*, (CERN), CERN DPS, 2018.

[105] M. Wielgosz, A. Skoczeń and M. Mertik, *Using lstm recurrent neural networks for monitoring the lhc superconducting magnets*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **867** (2017) 40.

[106] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov and A. J. Smola, *Deep sets*, in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

[107] P. T. Komiske, E. M. Metodiev and J. Thaler, *Energy flow networks: deep sets for particle jets*, *Journal of High Energy Physics* **2019** (2019) .

[108] H. Qu and L. Gouskos, *Jet tagging via particle clouds*, *Physical Review D* **101** (2020) .

[109] S. R. Qasim, J. Kieseler, Y. Iiyama and M. Pierini, *Learning representations of irregular particle-detector geometry with distance-weighted graph networks*, *The European Physical Journal C* **79** (2019) .

[110] Qasim, Shah Rukh, Long, Kenneth, Kieseler, Jan, Pierini, Maurizio and Nawaz, Raheel, *Multi-particle reconstruction in the high granularity calorimeter using object condensation and graph neural networks*, *EPJ Web Conf.* **251** (2021) 03072.

[111] A. Shmakov, M. J. Fenton, T.-W. Ho, S.-C. Hsu, D. Whiteson and P. Baldi, *Spanet: Generalized permutationless set assignment for particle physics using symmetry preserving attention*, *arXiv preprint arXiv:2106.03898* (2021) .

[112] V. Mikuni and F. Canelli, *Point cloud transformers applied to collider physics*, *Machine Learning: Science and Technology* **2** (2021) 035027.

[113] E. Bols, J. Kieseler, M. Verzetti, M. Stoye and A. Stakia, *Jet flavour classification using DeepJet*, *Journal of Instrumentation* **15** (2020) P12012.

[114] CMS collaboration, *Algorithms for b Jet identification in CMS*, Tech. Rep. CMS-PAS-BTV-09-001, CERN, Geneva, 2009.

[115] CMS collaboration, *Identification of b-Quark Jets with the CMS Experiment*, *JINST* **8** (2013) P04013 [1211.4462].

[116] CMS collaboration, *Identification of b quark jets at the CMS Experiment in the LHC Run 2*, Tech. Rep. CMS-PAS-BTV-15-001, CERN, Geneva, 2016.

[117] CMS collaboration, *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*, *JINST* **13** (2018) P05011 [arXiv:1712.07158].

[118] ATLAS collaboration, *Performance of b-jet identification in the ATLAS experiment*, *JINST* **11** (2016) P04008.

[119] ATLAS collaboration, *Atlas b-jet identification performance and efficiency measurement with tt ¯ events in pp collisions at s=13 tev*, *The European Physical Journal C* **79** (2019) .

[120] ATLAS, *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment*, Tech. Rep. ATL-PHYS-PUB-2017-003, CERN, Geneva, 2017.

[121] P. Nason, *A new method for combining NLO QCD with shower Monte Carlo algorithms*, *JHEP* **11** (2004) 040 [hep-ph/0409146].

[122] S. Alioli, P. Nason, C. Oleari and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, *JHEP* **06** (2010) 043 [1002.2581].

[123] S. Frixione, P. Nason and C. Oleari, *Matching NLO QCD computations with parton shower simulations: the POWHEG method*, *JHEP* **11** (2007) 070 [0709.2092].

[124] J. M. Campbell, R. K. Ellis, P. Nason and E. Re, *Top-pair production and decay at NLO matched with parton showers*, *JHEP* **04** (2015) 114 [1412.1828].

[125] GEANT4 collaboration, *GEANT4—A simulation toolkit*, *Nucl. Instrum. Meth. A* **506** (2003) 250.

[126] CMS collaboration, *The CMS experiment at the CERN LHC*, *JINST* **3** (2008)

S08004.

[127] CMS, *Pileup mitigation at CMS in 13 TeV data*, *Journal of Instrumentation* **15** (2020) P09018.

[128] V. Nair and G. Hinton, *Rectified linear units improve restricted boltzmann machines*, *Proceedings of ICML* **27** (2010) 807.

[129] S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, 2015. 10.48550/ARXIV.1502.03167.

[130] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. 10.48550/ARXIV.1412.6980.

[131] J. Kieseler, E. Bols, M. Stoye, M. Verzetti, P. Silva, S. S. MEHTA et al., *Deepjetcore*, Feb., 2020. 10.5281/zenodo.3670882.

[132] J. Kieseler, E. Bols, M. Verzetti, C. Vernieri, D. Majumder, L. Gouskos et al., *Deepntuples*, Feb., 2020. 10.5281/zenodo.3639231.

[133] J. Kieseler, E. Bols, M. Stoye, M. Verzetti, A. Stakia, H. KIRSCHENMANN et al., *Deepjet*, Feb., 2020. 10.5281/zenodo.3670523.

[134] CMS collaboration, *Jet algorithms performance in 13 TeV data*, Tech. Rep. CMS-PAS-JME-16-003, CERN, Geneva, 2017.

[135] CMS collaboration, *A new calibration method for charm jet identification validated with proton-proton collision events at* $\sqrt{s}$ *=13 TeV*, *JINST* **17** (2022) P03014 [2111.03027].

[136] CMS collaboration, *Search for Higgs boson pair production in the four b quark final state in proton-proton collisions at* $\sqrt{s}$ *= 13 TeV*, 2202.09617.

[137] CMS collaboration, *Direct search for the standard model Higgs boson decaying to a charm quark-antiquark pair*, tech. rep., CERN, Geneva, 2022.

[138] CMS, *A search for the standard model higgs boson decaying to charm quarks*, *Journal of High Energy Physics* **2020** (2020) .

[139] CMS collaboration, *A Deep Neural Network for Simultaneous Estimation of b Jet Energy and Resolution*, *Comput. Softw. Big Sci.* **4** (2020) 10 [1912.06046].

[140] J. Bellm, S. Gieseke, D. Grellscheid, S. Plätzer, M. Rauch, C. Reuschle et al.,

Herwig 7.0/herwig++ 3.0 release note, *The European Physical Journal C* **76** (2015) 1.

[141] P. Gras, S. Höche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer et al., *Systematics of quark/gluon tagging*, *Journal of High Energy Physics* **2017** (2017) .

[142] G. Louppe, M. Kagan and K. Cranmer, *Learning to pivot with adversarial networks*, 1611.01046.

[143] L. Lyons, *A method of reducing systematic errors in classification problems*, *NIMA A 324* (1993) .

[144] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski and D. Whiteson, *Parameterized machine learning for high-energy physics*, *Eur. Phys. J. C 76* (2016) .

[145] CMS collaboration, *A deep neural network to search for new long-lived particles decaying to jets*, *Machine Learning: Science and Technology* **1** (2020) 035012.

[146] CMS collaboration, *Measurement of the top quark mass with lepton+jets final states using pp collisions at* $\sqrt{s} = 13tev$, *The European Physical Journal C 78* (2018) [1805.01428].

[147] D0 collaboration, *Measurement of the top quark mass in the lepton+jets channel using the ideogram method*, *Phys.Rev.D75:092001,2007* (2007) [hep-ex/0702018].

[148] CMS collaboration, *A profile likelihood approach to measure the top quark mass in the lepton+jets channel at* $\sqrt{s} = 13$ TeV, tech. rep., CERN, Geneva, 2022.

[149] CMS collaboration, *Extraction and validation of a new set of CMS pythia8 tunes from underlying-event measurements*, *The European Physical Journal C* **80** (2020) .

[150] CMS collaboration, *Performance of CMS muon reconstruction in pp collision events at sqrt(s) = 7TeV*, *Journal of Instrumentation* **7** (2012) P10002.

[151] CMS COLLABORATION collaboration, *Jet Performance in pp Collisions at 7 TeV*, tech. rep., CERN, Geneva, 2010.

[152] DØ collaboration, *Direct measurement of the top quark mass by the dø collaboration*, *Phys. Rev. D* **58** (1998) 052001.

[153] S. S. Snyder, *Measurement of the top quark mass at D0*, Ph.D. thesis, 1995.

[154] S. Dildick, *Application of kinematic fitting to top quark mass reconstruction in the mu+jets channel at* $\sqrt{s} = 7tevwiththecmsdetector, Master'sthesis$.

[155] W. Verkerke and D. Kirkby, *The roofit toolkit for data modeling*, 2003. 10.48550/ARXIV.PHYSICS/0306116.

[156] A. Heister, S. Schael, R. Barate, I. De Bonis, D. Decamp, C. Goy et al., *Study of the fragmentation of b quarks into b mesons at the z peak*, *Physics Letters B* **512** (2001) 30.

[157] J. Abdallah, , P. Abreu, W. Adam, P. Adzic, T. Albrecht et al., *A study of the b-quark fragmentation function with the DELPHI detector at LEP i and an averaged distribution obtained at the z pole*, *The European Physical Journal C* **71** (2011) .

[158] C. Peterson, D. Schlatter, I. Schmitt and P. M. Zerwas, *Scaling violations in inclusive e^+e^- annihilation spectra*, *Phys. Rev. D* **27** (1983) 105.

[159] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *Journal of High Energy Physics* **2014** (2014) .

[160] R. Frederix and S. Frixione, *Merging meets matching in MC@NLO*, *Journal of High Energy Physics* **2012** (2012) .

[161] S. Mrenna and P. Skands, *Automated parton-shower variations in pythia 8*, *Physical Review D* **94** (2016) .

[162] CMS collaboration, *Measurement of differential cross sections for top quark pair production using the lepton+jets final state in proton-proton collisions at 13 TeV*, *Physical Review D* **95** (2017) .

[163] CMS collaboration, *Measurement of normalized differential $t\bar{t}$ cross sections in the dilepton channel from pp collisions at $\sqrt{s} = 13$ TeV*, *Journal of High Energy Physics* **2018** (2018) .

[164] J. R. Christiansen and P. Z. Skands, *String formation beyond leading colour*, *Journal of High Energy Physics* **2015** (2015) .

[165] S. Argyropoulos and T. Sjöstrand, *Effects of color reconnection on $t\bar{t}$ final states at the LHC*, *Journal of High Energy Physics* **2014** (2014) .

[166] CMS collaboration, *Study of the underlying event in top quark pair production in pp collisions at 13 tev*, *The European Physical Journal C* **79** (2019) .

[167] L. Giannini, *Deep Learning techniques for the observation of the Higgs boson decay to*

bottom quarks with the CMS experiment, Ph.D. thesis, 2020.

[168] B. T. Huffman, C. Jackson and J. Tseng, *Tagging b quarks at extreme energies without tracks*, *Journal of Physics G: Nuclear and Particle Physics* **43** (2016) 085001.

[169] L. A. Gatys, A. S. Ecker and M. Bethge, *A neural algorithm of artistic style*, 2015. 10.48550/ARXIV.1508.06576.

[170] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2014. 10.48550/ARXIV.1409.1556.

[171] X. Wang, L. Xie, C. Dong and Y. Shan, *Real-esrgan: Training real-world blind super-resolution with pure synthetic data*, 2021. 10.48550/ARXIV.2107.10833.