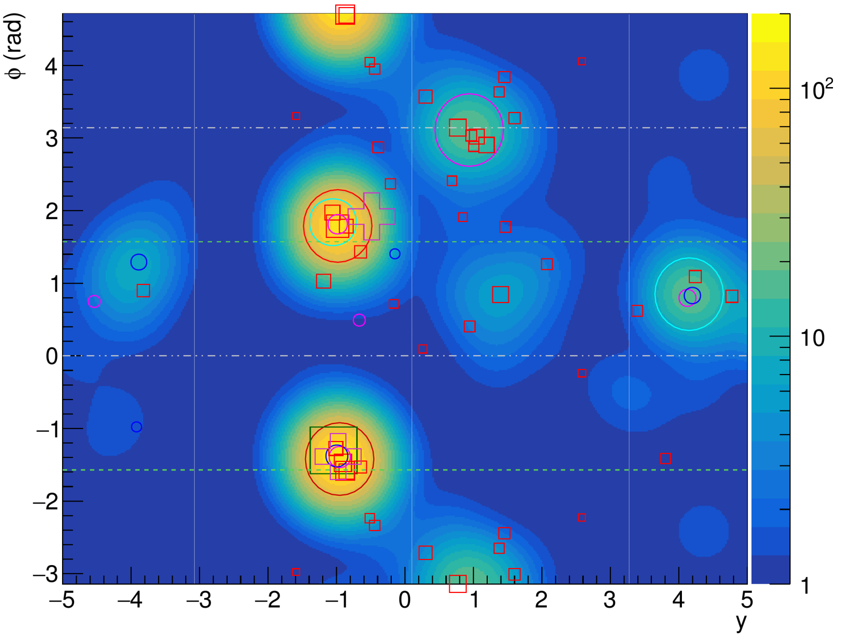
Aalto University
School of Science
Degree Programme in Engineering Physics and Mathematics

Hannu Siikonen

Jet flavors:

From the standard candles to the top quark mass



Master's Thesis
Espoo, March 20, 2016

Supervisor: Professor Mikko Alava
Advisor: Assistant Professor Mikko Voutilainen

Aalto University
School of Science
Degree Programme in Engineering Physics and Mathematics

ABSTRACT OF
MASTER'S THESIS

Author:	Hannu Siikonen		
Title: Jet flavors: From the standard candles to the top quark mass			
Date:	March 20, 2016	**Pages:**	viii + 89
Major:	Computational Physics	**Code:**	Tfy-105
Supervisor:	Professor Mikko Alava		
Advisor:	Assistant Professor Mikko Voutilainen		

The LHC began its second run in 2015 with upgraded hardware and software and an increased collision energy. With the higher energy scale the importance of *jet physics* has grown even larger than before. Jets are collimated sprays of hadrons that are produced in high-energy particle collisions. Understanding jets makes it possible to analyze the proton-proton collisions occurring at the LHC.

This work studies *jet flavors* and their definitions in the context of the CMS experiment. The motivation for this is that the jet energy corrections applied to the CMS data depend on the jet flavors. A jet flavor is typically understood as the flavor of the quark or gluon from which the jet originated. In other contexts, e.g. b-tagging, the meaning of a jet flavor can be slightly different.

Focus is given to the study of jet flavor definitions in simulations of proton-proton collisions. Due to the structure of simulations the flavor definitions have an algorithmic form. The flavor studies begin by inspecting the robustness of a previously favored jet flavor definition between three different simulation software packages. Here the robustness of a flavor means that the physical properties of each flavor are the same in three different collision event types (*standard candle* events).

Good robustness properties are observed between the software packages, but an excessive amount of jets is left without any flavor tag. A solution for this problem is sought for by developing enhanced flavor definitions. Two prominent new flavor definitions are found in the studies.

The knowledge gained in the flavor studies is then applied to the studies of top quark production. The abundance of jets is particularly high in collisions that produce top quarks, so jet-related knowledge is important. It turns out that here the jet flavor properties are somewhat similar to those observed in the standard candle collision events. However, there are some differences that require further study. To conclude, a simulated measurement of the top quark mass is made. This provides valuable understanding of the practical issues related to a top mass measurement.

Keywords:	LHC, CMS, Jet, Jet flavor, Top quark mass
Language:	English

Aalto-yliopisto
Perustieteiden korkeakoulu
Teknillisen fysiikan ja matematiikan koulutusohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Hannu Siikonen		
Työn nimi:			
Jettien maut: Referenssi-prosesseista top-kvarkin massaan			
Päiväys:	20. maaliskuuta 2016	**Sivumäärä:**	viii + 89
Pääaine:	Laskennallinen fysiikka	**Koodi:**	Tfy-105
Valvoja:	Professori Mikko Alava		
Ohjaaja:	Apulaisprofessori Mikko Voutilainen		

LHC aloitti toisen toimintajaksonsa vuonna 2015 kasvatetulla törmäysenergialla ja päivitetyllä laitteistolla. Energiaskaalan kasvun myötä *jettien fysiikan* merkitys on entistäkin suurempi. Jetit ovat kapeita hadroniryöppyjä, joita syntyy korkean energian hiukkastörmäyksissä. LHC:lla tapahtuvien protoni–protoni -törmäysten analysointi on mahdollista jettien ymmärtämisen ansiosta.

Tässä työssä tutkitaan jettien makuja ja niiden määritelmiä CMS-kokeen kontekstissa. Tutkimusta motivoi se, että CMS-kokeen tuottaman datan jettien energiakorjauksissa jettien mauilla on suuri merkitys. Jetin maulla tarkoitetaan yleensä sen kvarkin tai gluonin tyyppiä, josta jetti on peräisin. Merkitys voi kuitenkin vaihdella hiukan eri yhteyksissä, kuten b-jettien tutkimuksessa.

Työ keskittyy jettimakujen tutkimukseen protoni–protoni-törmäyksien simulaatioissa. Simulaatioiden rakenteen vuoksi jettimaku määritellään algoritmien avulla. Makujen tutkimus alkaa aiemmin hyväksi todetun määritelmän vakauden vertailulla kolmen simulaatio-ohjelmiston välillä. Vakaudella tarkoitetaan sitä, että kunkin jettimaun fysikaaliset ominaisuudet pysyvät samoina kolmessa eri törmäystyypissä (referenssi-prosessit).

Simulaatio-ohjelmien väliset vakausominaisuudet todetaan hyviksi, mutta samalla havaitaan, että tilastollisesti liian monen jetin maun määritys epäonnistuu. Tähän ongelmaan haetaan ratkaisua kehittämällä parannettuja maun määritelmiä. Kaksi uutta määritelmää osoittautuu toimiviksi.

Makututkimuksessa kerättyä tietoa sovelletaan edelleen top-kvarkkeja tuottavien törmäysten tutkimiseen. Jettejä esiintyy erityisen paljon tällaisissa törmäyksissä, minkä vuoksi niiden tuntemus on tärkeää. Ilmenee, että jettimakujen ominaisuudet ovat jokseenkin samanlaisia kuin referenssi-prosesseissa. Jotkin poikkeavuudet kuitenkin vaativat jatkotutkimusta. Työn päätteeksi suoritetaan simulaatiopohjainen top-kvarkin massan määritys. Tämä tarjoaa hyvää ymmärrystä käytännön hankaluuksista, joita esiintyy top-kvarkin massan mittauksessa.

Asiasanat:	LHC, CMS, Jetti, Jettimaku, Top-kvarkin massa
Kieli:	Englanti

Acknowledgements

First of all I would like to thank Asst. Prof. Mikko Voutilainen for the opportunity to work with jet physics at the CMS experiment. This has proven to be a very interesting time to work in one of the greatest science projects in the world. I would also like to thank Prof. Mikko Alava for supervision and guidance through the technicalities related to this thesis.

Thanks are due to my co-workers at the University of Helsinki for making this time enjoyable. Special thanks go to Santeri Laurila for the effort of making corrections to this text. I have to thank also all my classmates through the years at the Aalto University. Without a good atmosphere the coursework would have felt a lot more stressing.

Last of all, thanks go to my friends and my family. The support from my parents was essential for getting this text finished.

Espoo, March 20, 2016

Hannu Siikonen

Symbols and Abbreviations

Symbols

c	the speed of light
$\hbar$	the reduced Planck constant
σ	cross section
L	instantaneous luminosity
dN/dt	rate of interactions
m	mass
E	energy
$\mathbf{p}$	momentum 3-vector
p	momentum 4-vector
p_L	magnitude of longitudinal momentum
p_T, k_T	magnitude of transverse momentum
E_T	transverse energy
y	rapidity
η	pseudorapidity
ϕ	azimuthal angle
ΔR	distance in the (y, ϕ)-plane
$\sqrt{s}$	center-of-mass collision energy
$\mathrm{\mu}$	muon
$\mathrm{\gamma}$	photon
Z	Z boson
W	W boson

Abbreviations

CERN	The European Organization for Nuclear Research
LHC	Large Hadron Collider
LEP	Large Electron Positron collider
ECAL	Electromagnetic Calorimeter
HCAL	Hadronic Calorimeter

ATLAS	A Toroidal LHC ApparatuS
CMS	Compact Muon Solenoid
CMSSW	CMS SoftWare
PF	Particle Flow algorithm
SM	Standard Model
QCD	Quantum ChromoDynamics
HEP	High Energy Physics
MC	Monte Carlo
GPMC	General Purpose Monte Carlo event generator
LHE(F)	Les Houches accord Event (File)
PDF	Parton Distribution Function
LO	Leading Order
NLO	Next to Leading Order
ISR	Initial State Radiation
FSR	Final State Radiation
MPI	Multiple Parton Interactions
MET	Missing Transverse Energy
QGL	Quark-Gluon Likelihood
JES	Jet Energy Scale
JEC	Jet Energy Corrections
event	a single proton-proton collision event
standard candle events	dijet, γ+jet and Z$\mu\mu$+jet events
LO parton	a LO hard process outgoing parton
FS parton	a partonic final state parton (before hadronization)
LO definition	a new flavor definition that utilizes the LO partons (derived from the physics definition)
CLO definition	like the LO definition, but uses corrected parton momenta
FS definition	a new flavor definition that utilizes the FS partons (derived from the algorithmic definition)
ROOT	a C++ library designed for data analysis at CERN
Pythia 6	a FORTRAN Monte Carlo event generator
Pythia 8	a C++ follower for Pythia 6
Herwig++	another event generator, not as popular as Pythia

Contents

Chapter 1

Introduction

The discovery of a Higgs boson is certainly one of the greatest highlights so far for the Large Hadron Collider (LHC). This last missing piece of the standard model (SM) of particle physics was theorized decades ago. The significance of the discovery can be attributed to the fact that it completed the SM. Even if the finding of the Higgs particle has gained the greatest media coverage, it is only one of the important results from the first run of the LHC.

Currently a majority of the personnel working for CERN (the European organization for nuclear research) has something to do with the LHC. Modern particle physics experiments require great efforts in engineering, data analysis and simulation.

The LHC entered its first long shutdown period in early 2013. During the shutdown multiple upgrades were applied to the accelerator complex and its detectors. In 2015 the second run of the LHC commenced, stretching until the end of 2018. During this run the collision energy is incremented significantly from those of the previous run. Several upgrades and runs have been planned to be performed even after the current run. There is a great interest for the new physics results beyond the SM, which will possibly be observed during this LHC run. Supersymmetry is one of the most popular hypothetical theories, but the experiments have not verified it.

The most significant LHC run 2 results are likely to be produced when the run approaches its end. Then the maximum amount of data is available and detector calibrations have been enhanced. In the meantime, simulations provide a good method for the preparation of the analysis tools. In this work the main focus is given to the simulations and analysis of collision events at the LHC. The simulations are done in the context of the CMS (Compact Muon Solenoid) detector.

Within the limits of the SM, the top quark and the Higgs boson are in a way the most interesting elementary particles. The top quark has different characteristics than the other quarks due to its large mass. Top is the most massive known elementary particle, leaving the Higgs boson with approximately one fourth smaller a mass. The Higgs boson and top quark masses can be used together to make predictions of the stability of the universe. The current mass measurements say that the universe should be metastable but the prediction is vulnerable to the mass measurement error. Thus an excellent accuracy for the measurement of both of these masses is highly desirable. Between the most recent measurement results there have been differences of the order of 1 GeV in the top quark mass.

An important motivation for this thesis is to initiate the construction of an analysis environment that determines the top quark mass. This will be continued as a PhD project that aims to an extremely precise measurement of the top quark mass. With its data the upcoming LHC run will bring possibilities for a significantly improved mass measurement, which fits in well with these plans. While the collision energies are increased the collision events become even more complicated. Thus a carefully planned analysis is essential with the data from the new LHC run.

A most important physical observable in modern collider physics is a *jet*, a collimated spray of particles. The formation of jets is caused by the nature of the strong interaction. For instance in collision events involving top quarks, jets have an especially important role. Studying clusters of particles instead of single particles adds a significant complication to the measurements. However, accurate physical studies are still feasible, as long as the jets are considered carefully.

One way for characterizing jets is labeling them according to the strongly interacting elementary particle that was their origin. This determines a so called *jet flavor*. A full understanding of the strong interaction and quantum chromodynamics (QCD) is still lacking. Thus at the present simulations are the only tool that allows us to study the nature of the jets closely.

Another central objective of this thesis is to study closely the properties and definitions of jet flavors. A deeper understanding of the jets is the key for reducing systematic errors in the measurement of the top quark mass. These studies are done in what we call *the standard candle events*. In the context of this thesis these refer to three collision event types with distinct and detectable properties. They form a stepping stone for studying the universal properties of jets. Since a great amount of simulation data is involved, the flavor definitions are given in an algorithmic form. Only a well-performing definition allows rigorous analysis.

The flavor studies are initiated with a study of the robustness of a previously favored jet flavor definition. A robustness comparison is done between three simulation software distributions: Pythia 6, Pythia 8 and Herwig++. This serves as a continuation to a previous master's thesis [1], which studied the robustness properties only using Pythia 8. Such an extension is essential, since an agreement between different software packages hints of a physical generality of the results. After the robustness studies a thorough inspection of the problems with the flavor definitions follows.

In addition to the central simulation software, also some software tools are needed for the processing of the generated particle-level data. To handle the multitude of software packages a massive software environment was developed for the uses of this work.

The structure of this thesis is as follows. In chapter 2 a short overview to the most relevant theoretical concepts is given. The third chapter continues with a more detailed description of the LHC and the context of the simulations that are done in this work. In the fourth chapter the logical structure of Monte Carlo (MC) collision event generators is introduced. Also a brief description of the relevant software packages and their differences follows. Chapter 5 continues by describing the processing and storing of simulated particle data. The processing involves the practice of clustering particles into jets and intricate algorithmic handling of the collision events. Finally, in chapters 6 and 7 the results along with a discussion are provided. Chapter 8 concludes the work, followed by the references and the appendices.

Chapter 2

Theoretical background: particle physics and jets

This chapter serves as an introduction to the most important physical phenomena in jet production. In short, jets are a manifestation of the physics described by the SM. In high-energy particle collisions color confinement and the asymptotic freedom of quarks cause the production of particle cascades, called jets. They are an important tool for studying quarks, which can only be observed indirectly in the nature. In collider physics jets are particularly important, since much of the data produced by detectors are processed as jets.

The chapter starts by going through practical conventions and relevant quantities in collider physics. Then, a compact traversal from the basics of particle physics to jet physics is done. To conclude, the full mechanism of jet production in proton-proton collisions and the constitution of jets are reviewed. The text is based on fundamental textbook references [2–4], and practical knowledge gained while working in the CMS project at CERN.

2.1 Conventions and quantities at the LHC

Natural units are used implicitly everywhere to reduce the complexity of physical expressions. In particle physics the relevant choices are fixing the natural constants so that $c = \hbar = 1$. These choices provide a numerical connection between the units of time, distance, mass and all the derived quantities. Thus velocities appear dimensionless, and the values of mass, energy and momentum all have the same units. The latter three are typically viewed in the multiples of eV and e.g. masses are often found to be in the MeV or GeV scale. By referring to the unit choices one can return to SI-units. However, it is common to skip this and to present results directly in natural units.

The **cross section** σ is an important quantity in the physics of particle collisions. A loose definition can be given as

$$\sigma = \frac{\#\text{interactions (collisions) in a time unit}}{\#\text{incoming particles per area in a time unit}}. \tag{2.1}$$

The cross sections are thus given in the units of area and expressed in barns – typically in femto- or picobarns ($1\,\mathrm{b} = 100\,\mathrm{fm}^2$). At the LHC the accurate formulation for σ would take into account also the composition of the colliding proton bunches.

Proton-proton collisions can be elastic or inelastic. In the former case the physics of a collision is relatively simple. On the other hand, the latter case can be categorized into diffractive and non-diffractive collisions. At the LHC non-diffractive collisions are of the biggest interest. They involve physics processes that probe the structure of matter deeply. All the processes mentioned contribute to the total cross section. Due to the form of the definition in Eq. (2.1), the cross sections of all occurring sub-processes sum up to a total cross section σ_{tot}.

Luminosity L is a very useful quantity that is closely paired up with the cross sections. In simplified terms instantaneous luminosity can be defined as

$$L = \sigma_{tot}^{-1} \frac{dN}{dt}. \tag{2.2}$$

That is, the rate of interactions (dN/dt) normalized with the total cross section (σ_{tot}). In the collisions of proton-bunches at the LHC the expressions for L and σ can be given in more involved forms. However, the underlying ideas do not change.

Integration of the instantaneous luminosity L over the *data gathering time* gives another important quantity: **integrated luminosity** L_{int}. Using L_{int} is the most common way of expressing the amount of recorded collision event data. For studying a specific sub-process, a multiplication between L_{int} and the corresponding cross section approximates the amount of relevant data. In addition, theoretical and simulated cross sections can be compared with experimental ones to gain a better understanding of the collisions.

The choice of a **coordinate system** is arbitrary, but it should be mutual for all analyses. In the CMS experiment the cylindrical symmetry of the detector is used as a motivation for this choice. The z-axis lies in the beam direction, the x-axis points towards the center of the LHC ring and the y-axis points upwards from the ground. The direction of momentum in the xy-plane is indicated by an azimuthal angle ϕ. In most analyses one can expect that the azimuthal direction makes in average no physical difference. Any anomalies in the ϕ-direction can be used to observe difficulties in detector calibration or insufficient sample sizes.

In accelerator physics the momentum component in the beam direction (p_L) is often substituted with **rapidity**, y. It is defined as

$$y = \frac{E + p_L}{E - p_L}, \tag{2.3}$$

where E is the energy of a particle or a jet. Translations in rapidity (Δy) are Lorentz-invariant with respect to boosts in the beam direction. Additionally, the differential Lorentz-invariant cross section $d\sigma/dy$ should be approximately constant in proton-proton collisions. That is, a plot of particle frequencies as a function of y should give a flat profile. This prediction is valid in some range $|y| < Y$.

A **radial distance** ΔR is mostly used to study the differences in the directions of particles or jets. It is defined using the above-mentioned two measures for direction: $\Delta R = \sqrt{\Delta\phi^2 + \Delta y^2}$. In simplified terms for instance a jet can be defined as a cone with a certain axis (y, ϕ) and a radius R.

Pseudorapidity η is an important experimental quantity that is often used instead of rapidity. Rapidity carries the inconvenience that it requires knowledge of the mass of a given particle. In high-energy physics it is often justified to assume that the momentum is much larger than the rest mass of a particle. Thus Eq. (2.3) simplifies to

$$\eta = \frac{|\mathbf{p}| + p_L}{|\mathbf{p}| - p_L} = -\ln\tan(\theta/2)\,. \tag{2.4}$$

Here $\mathbf{p}$ is the total momentum of the particle or jet and θ the polar angle of the momentum direction. There has been some controversy between the convenience of pseudorapidity and the problems caused by using it instead of rapidity [5].

Transverse momentum and **transverse energy** are specific quantities for studying the properties of particles and jets orthogonally to the beam direction. These are defined correspondingly as

$$p_T = k_T = \sqrt{p_x^2 + p_y^2} \tag{2.5}$$

$$E_T = \sqrt{p_T^2 + m^2}. \tag{2.6}$$

In high energy physics (HEP) the energies are usually so large that the rest masses are ignored. Hence, E_T and p_T are often used interchangeably, which can cause confusion. When it comes to p_T, a double-differential invariant cross section can be presented as a quantity proportional to $d^2\sigma/dydp_T^2$. Thus the motivation for using p_T is closely related to that of using y (and η).

Some common words have ambiguous special meanings in collider physics. **Hardness** refers to high p_T values of a particle or a jet, but can be understood more generally as a high energy. Correspondingly, **softness** implies small p_T values. Furthermore, at the LHC an **event** means a single collision of two protons. Stretching the definition further, an event can be considered to mean only **hard interactions** between protons. Hard interactions are handled more closely in Section 2.3. The exact meaning of the word event is typically connected to the context.

The center of mass energy of the colliding protons at the LHC is denoted with $\sqrt{s}$. Here s is a corresponding **Mandelstam variable** defined as $s = (p_1 + p_2)^2$. The symbols p_1 and p_2 refer to the four-momenta of the colliding protons. One should be aware of this notation as it is commonly used in the relevant publications. The LHC run 1 finished using $\sqrt{s} = 8\,\text{TeV}$ and the run 2 began with $\sqrt{s} = 13\,\text{TeV}$.

The elementary particles gluons and quarks are together referred to as **partons**. These are the elementary particles that are subject to the strong interaction. The properties of partons will be clarified in the next section.

2.2 The Standard Model

The SM is the indispensable foundation of modern particle physics. It consists of experimentally confirmed physical theories that fit well together and are accepted by most physicists. Beyond the standard model there are many competing theories that could be used to explain various phenomena. The SM can be extended with such theories, e.g. with supersymmetry. Intricate experimental tests are needed to get further proof of the validity or non-validity of these kinds of extensions.

The essence of the SM is in particles and interactions. There are three types of interactions in the SM: the strong, the weak and the electromagnetic interaction. These interactions are mediated by the respective gauge bosons: gluons, W or Z bosons, and photons. The gauge bosons of interactions can be thought of as ripples in the corresponding fields. The electric and the weak interactions are unified in the electroweak theory.

Gravity is completely absent from the standard model for several reasons. First and foremost there is a lack of a solid and experimentally confirmed quantum formulation of gravity. Moreover, gravity plays usually a very small role on the particle level. Thus particle-level events can be understood even if the contribution of gravity is neglected.

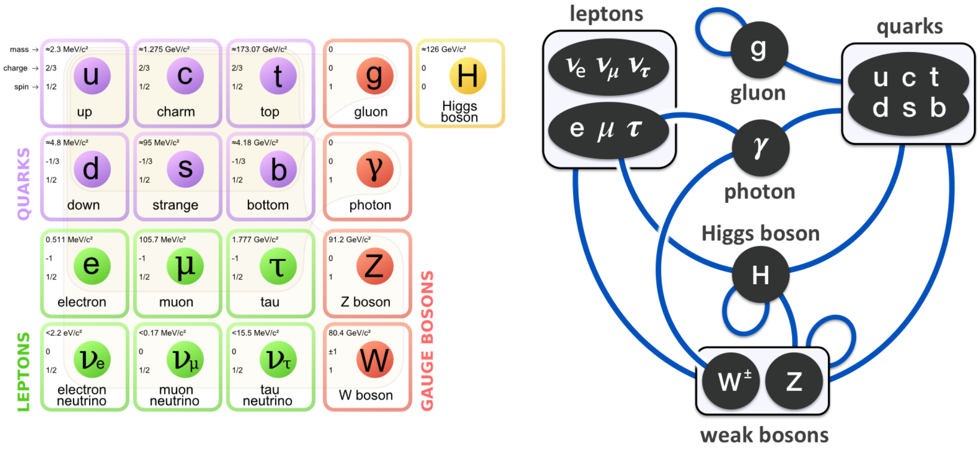

mass →
charge →
spin →
≈2.3 MeV/c² 2/3 1/2 u up
≈1.275 GeV/c² 2/3 1/2 c charm
≈173.07 GeV/c² 2/3 1/2 t top
0 0 1 g gluon
≈126 GeV/c² 0 0 H Higgs boson
≈4.8 MeV/c² -1/3 1/2 d down
≈95 MeV/c² -1/3 1/2 s strange
≈4.18 GeV/c² -1/3 1/2 b bottom
0 0 1 γ photon
0.511 MeV/c² -1 1/2 e electron
105.7 MeV/c² -1 1/2 μ muon
1.777 GeV/c² -1 1/2 τ tau
91.2 GeV/c² 0 1 Z Z boson
<2.2 eV/c² 0 1/2 νe electron neutrino
<0.17 MeV/c² 0 1/2 νμ muon neutrino
<15.5 MeV/c² 0 1/2 ντ tau neutrino
80.4 GeV/c² ±1 1 W W boson
QUARKS
LEPTONS
GAUGE BOSONS
leptons
ν_e ν_μ ν_τ
e μ τ
g
gluon
quarks
u c t
d s b
γ
photon
Higgs boson
H
$W^\pm$
Z
weak bosons

Figure 2.1: The particles and couplings in the standard model. [6, 7]

In Fig. 2.1 the forces and couplings of the elementary particles in the SM are shown. The ordinary **matter** consists of elementary fermions (half integer spin) as opposed to the gauge bosons (integer spin) which convey the interactions. There are two important groups of matter particles: quarks and leptons. An obvious symmetry between these two is present: both have three columns and two rows. The columns are called generations. According to experimental results there should not be any unknown generations left. A deeper understanding of the reason for this is lacking.

Each generation consists of two particles that are coupled through the weak interaction by an exchange of a W boson. For leptons there is an obvious asymmetry between these pairs. The neutrinos have very small masses and no electrical charge, which makes a notable separation to the charged leptons. On the other hand, all quarks are charged and their masses are closer to each other than those of the leptons. As a special case, the masses of the first quark generation are small and almost equal to each other.

All quarks and leptons interact through the weak interaction, as indicated above. The weak interaction mixes quark generations, meaning that processes involving weak gauge bosons can lead to a change of generation. Also the lepton generations are mixed in a similar manner. This mixing is most importantly observed as **neutrino oscillations**, which have been confirmed in recent measurements. Due to neutrino oscillations the neutrino masses have to be non-zero. This is an important result, since the neutrino masses are so small that their values are still unknown.

Neutrinos are in general a difficult object to study. Due to their zero electrical charge they are the only elementary fermions that do not interact electromagnetically. This complicates measurements considerably.

The quarks are the only elementary fermions that interact through the strong interaction. This is the most important difference between quarks and leptons. Being a subject to the strong force means that quarks have always a color charge. The options are red, green or blue and the corresponding anti-colors. Thus there are actually 2×6 unique quarks in each generation. Gluons, on the other hand, carry two color charges. Due to this, there are in total eight different gluons.

A peculiar property of the color charge is that only **color singlets** can manifest themselves in nature. A singlet is a state containing all three colors or anti-colors (baryon), or a color and the respective anti-color (meson). This behavior of the color charges is called color confinement.

If the momenta of quarks or gluons within a singlet differ by much, the property of asymptotic freedom manifests. The force holding the singlet together becomes stronger and stronger when the distance between the partons grows. If the separation grows large enough, a quark–anti-quark pair can be produced in-between. This reduces the distances between quarks, diminishing the forces between them. Owing to the production of new quarks, separation into new color singlets is possible.

The recently discovered Higgs particle is the most curious one in the current SM. It is the only elementary scalar boson (spin 0) in the model. The finding of a Higgs particle was important, since it indicates the presence of the Higgs field. The properties of this Higgs boson might open a door for the study of new physics in the future.

The Higgs field itself is an important part of the SM. Because of this field the theory of the SM makes sense as a whole. It is the only known field that has a non-zero value at its energy minimum. Due to this the rest masses of the W and Z bosons are non-zero. This allows also the non-zero rest masses of quarks and leptons.

To conclude, some attention is given to the relationship between masses and energies. In particle physics it is often preferred to use the terms **mass** and **invariant/rest mass** interchangeably. This leaves the notion of relativistic mass in the Einstein formula $E = mc^2$ into a deprecated position. Nevertheless, the Einstein formula is convenient also for understanding rest masses. This is the case for instance for protons, for which the rest mass is many magnitudes higher than the sum of the masses of the three quark constituents (uud). A majority of this rest mass is covered by the energies of internal gluons and quark–anti-quark pairs. These are due to the interaction between the three **valence quarks**. The properties of protons are studied more closely in the following chapter.

2.3 Proton-proton collisions

The LHC is a machine that produces proton collisions. For instance as compared to electron-positron colliders, hadron colliders generate very complicated collision events. Electrons and positrons are according to the current understanding pointlike fundamental particles. On the contrary, protons are not pointlike since they are composite particles constituting from quarks. Thus, the strong interaction plays an important role in hadron collisions.

To study the proton-proton interactions it is necessary to understand the inner structure of protons well. A proton is made up of two **up** (u) quarks and a **down** (d) quark. This is not the full story: gluons and quark–anti-quark pairs emerging from the Dirac sea are also present. The mass of a proton emphasizes the role of these sea quarks and gluons. The u and d quarks have the respective masses of 2.3 MeV and 4.8 MeV but the proton mass is approximately 1 GeV.

At high energies there are many ways in which protons can interact. One example is diffractive elastic scattering. This is observed particularly in the forward direction, where the proton trajectories remain almost unchanged. In physics studies hard interactions between protons are usually of the greatest interest. This means events in which the partons within the proton participate to the collision process. Fig. 2.2 shows an example of this kind of a collision. One parton from each proton is brought to the main interaction, leaving additional remnants from both protons. The most energetic collision of this kind in an event is commonly called the **hard interaction** or **hard process**. It is possible that also the proton (beam) remnants interact similarly with smaller energies. These additional inter-proton interactions are referred to as **multiple parton interactions** (MPI). Even without MPI the beam remnants have a non-trivial time development, caused by color confinement.

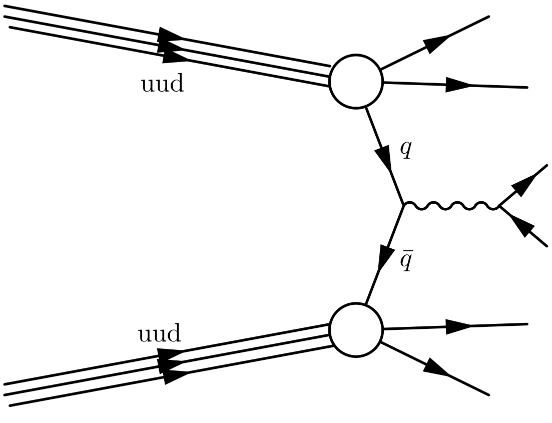


Figure 2.2: An example of a proton-proton interaction.

The **parton distribution functions** (PDF) are an important tool for studying the proton structure quantitatively. These give the probabilities for different parton types to emerge in an interaction as a function of x and Q^2. The variable x is the fraction of the original proton momentum carried by the parton going to the hard interaction. On the other hand, Q is the energy transfer in the proton-proton interaction.

At the present the PDFs are an inevitable source of uncertainties in simulations. The relative uncertainties of the PDFs are greatest when x is small. This implies that **soft interactions** (small energy exchange) are the hardest ones to model accurately. The most important PDF producers are the CTEQ group and the MRS/MRST/MSTW group.

2.4 Hadronization and jets

In consequence to the properties of the strong interaction the behavior of high-energy quarks is non-trivial. It is much more straightforward to make predictions for instance for the electromagnetic interaction. Photons are coupled only to particle–anti-particle pairs. On the contrary, gluons can couple to quark–anti-quark pairs and also to other gluons. This is due to the fact that in contrast to the zero electric charge photons, gluons carry color charge.

The gluon-gluon self-interactions make the force field structures difficult to study in an analytical form. Nevertheless with a numerical **lattice QCD** method it has been shown that asymptotic freedom follows from the QCD theory. Lattice QCD is a promising tool but it cannot be currently used in HEP simulations. Dynamically evolving large QCD systems are too challenging even for modern computers.

As described in the previous sections, partons with large energies may cause the production of quark–anti-quark pairs. When the initial parton energy has been split up between sufficiently many partons, **hadronization** occurs. In hadronization the multitude of original and produced quarks is distributed into color singlets. Theoretically this process is challenging. In simulations this is handled with well-motivated but theoretically not completely solid models. Thus high quality predictions can be made, even without a complete theoretical model.

Due to the strong interaction and hadronization, high-energy partons will thus produce sprays of particles. These sprays are observed at the detector, pointing in the direction of the original parton momentum. Since the process is highly probabilistic in its nature, the exact observed particle content can vary much. In understanding the original HEP process only the properties of the initial partons matter. The particles arriving to the detector are just an image of the original process. Thus we arrive to the concept of a **jet**. Ideally, a jet is a cluster of particles heading in a common direction and originating from a single high-energy parton. A sum over the momenta of the cluster of observed particles gives a direct handle to the original parton. The larger the original parton momentum is, the better collimated is the resulting jet. This is a result that follows from basic relativistic kinematics.

In practice real jets have many problems that are not considered in the ideal definition. One origin of issues is the potential overlap of jets at the detector. It is not trivial to distinguish whether the particles within an apparent jet are originated from one or multiple partons.

For a small amount of collision events it would be possible to cluster particles into jets by hand. Nonetheless, usually very high statistics are needed for physics analyses. Thus the jet clustering needs to be performed using algorithms. This results into an additional level of possible problems. Various issues may be left unseen due to the automatic algorithmic form of the clustering. All these potential sources of errors are dealt with by excluding events with suspicious physical properties from the analyses.

Since a jet is ideally thought to be originated from a single parton, a **flavor** can be assigned to each jet. The flavor of a jet is the flavor of the quark or gluon from which the jet originated. When it comes to jet flavors, the point of interest varies according to the type of analysis that is made. A generic study can be performed by comparing quark and gluon jets. Gluon jets tend to be wider and more numerous in their particle content than quark jets. These basic ideas lead to simulation-assisted methods for distinguishing quark jets from gluon jets.

In a more sophisticated study one might wish to make a separation between jets originating from different quark flavors. The two heaviest flavors are the most special ones. The **top** quark (t) has a mass heavier than the W boson. Because of this, the main decay channel for t is through the weak interaction. This occurs in such a short time period that hard interactions have no time to manifest. Thus the t quark does not strictly speaking obey color confinement and it does not form jets. The t quark decays with a small error margin almost always to a b quark and a W boson. Therefore, b jets are left as the quark jets with the heaviest flavor. Jets of the b flavor are of a special interest in many analyses. Therefore, special b-tagging algorithms have been developed.

The other jet flavors are of less interest and separating them from each other is difficult. The lightest quark flavors, u and d, are mostly considered indistinguishable in the jet context. Their masses are so small and so close to each other that they produce very similar jets. Often all the four lightest quark flavors (d, u, s and c) are considered together, as the separation of flavors is challenging. In simulations these can be separated, but the context of the detectors should always be taken into account.

2.5 Relevant particles at the LHC

In a large-scale data analysis there is a risk of losing the touch to reality. This can be avoided by creating physically intuitive representations for the numerical data. Here, a practical way to keep in touch with reality is to study the particles observed at the detector. According to the standard model various leptons and a multitude of hadrons are produced in the collisions. Thus the particle profile is quite different from the mundane protons, neutrons and electrons.

The lifetimes of most of the particles are so short that they are never seen at the detector level. A common convention in the CMS project is to consider particles with lifetimes greater than 10 mm/c stable. Since the particles move at varying speeds in varying directions this is simply an approximation. This causes no great errors since the lifetimes of different particles vary by several magnitudes.

The CMS detector cannot generally distinguish between various hadron types. Thus, simulations are a convenient and the only possible method for studying the particle content. The plots in Fig. 2.3 present the energy fractions of particle types present in jets. We see that there are many normally rare particle types present.

These plots give a short introduction to the kind of analysis that is done in this work. They have been generated by the software machinery that has been developed for the uses of this thesis. Two event generators are used to give a view of a physical property. A comparison between event generators hints of the possible generator-related uncertainties. The particle spectra produced by the generators are relatively similar. However, with a closer look there are differences of the order of a %-unit in some of the categories.

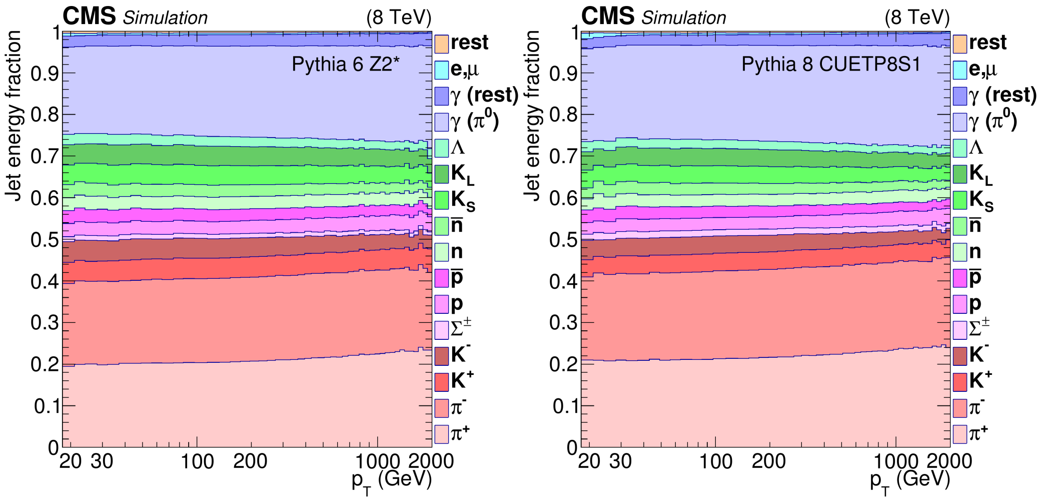


Figure 2.3: Jet particle energy fractions in Pythia 6 and Pythia 8.

The particles in the plots are sorted so that they can be grouped to the categories distinguished by the CMS detector. From the bottom to the top these are **charged hadrons** (ch), **neutral hadrons** (nh), **photons** (γ), **electrons** (e) and **muons** (μ). In the total jet energy the leptons play an insignificant role. However, outside of the jets *isolated* electrons and muons are an important tool of analysis. In addition to being a part of the physical analysis they are used as signals of certain event types.

The charged hadrons are most commonly pions ($\mathrm{u\bar{d}}$, $\mathrm{\bar{u}d}$) and the next place is taken by kaons ($\mathrm{u\bar{s}}$, $\mathrm{\bar{u}s}$). Mesons are typically lighter than baryons. This explains why the former two play a much more important role than protons (uud, $\mathrm{\bar{u}\bar{u}\bar{d}}$) and sigmas (uus,$\mathrm{\bar{u}\bar{u}\bar{s}}$, dds, $\mathrm{\bar{d}\bar{d}\bar{s}}$). Replacing a u or a d with an s in any particle results in a less common particle. For the massive c and b quark flavors the lifetimes are so short that none are observed.

When it comes to neutral hadrons the situation is quite similar. The greatest portion of neutral hadrons is covered by pions ($(\mathrm{u\bar{u}} - \mathrm{d\bar{d}})/\sqrt{2}$). All these decay to photons before reaching the detector. There is only a small portion of photons originating from other sources. Kaons (a mixture of u and $\mathrm{\bar{s}}$ or $\mathrm{\bar{u}}$ and s) have the second place in analogy to charged hadrons. Again after the mesons come the baryons: neutrons (udd,$\mathrm{\bar{u}\bar{d}\bar{d}}$) and lambdas (uds,$\mathrm{\bar{u}\bar{d}\bar{s}}$).

We observe that the possible jet constituents are not as numerous as one could think. Mesons are more abundant than baryons and the light quark flavors are more common than the massive ones. However, here only the standard type QCD events were studied. If more exotic events are chosen, the averaged jet contents can behave differently.

Chapter 3

The LHC and the CMS experiment

The context of the CMS detector at the LHC is an important part of this work. This is in spite of the fact that the results of the work are almost completely based on theory and simulations. Generic analysis could be done without the CMS context but this would be less beneficial for the whole project. Additionally, often the tuning of the simulations is done considering certain collision energies and detectors. Thus, if the full context was not considered the simulations could be mis-calibrated.

In this chapter we review the structure of the LHC and then the operation of the CMS detector. Then a further view is given of the acquisition of event data. Moreover, the steps of processing detector signals into particle data are considered. Finally we take a look at the common software structures used at the CMS experiment. The software is developed by a large number of scientists. It is used in most of the official analyses and thus it makes up an essential part of the CMS project. The software is not utilized in this work, since it restricts the scope of experimental studies severely. The chapter is based on the references [8, 9].

3.1 Production of proton collisions at the LHC

The foundation of CERN was envisioned after the second world war. The core purpose was to promote nuclear research for peaceful use in Europe. After some planning and negotiations the organization was officially established in 1954. The central facilities of CERN were to be built at the *Meyrin site* on the border of France and Switzerland, just outside Geneva.

After its establishment CERN has housed various important physical experiments and inspired many important innovations. For instance the **world wide web** was originally started at CERN. The impact of the www on science has been at least as great as it has been on the rest of the world. These days the nature of research is very international and fast global communication is essential.

At the moment the LHC is the main point of interest at CERN. Nevertheless there are always unrelated experiments and planning of future projects going on. Some of the CERN projects and experiments have lived their time and been terminated. However, the use of many experiments has been prolonged as much as possible. Moreover, parts of the previous experiments are often recycled to new projects, if feasible. The LHC was placed in the 27 km long circular tunnel that was previously occupied by the Large Electron-Positron collider (LEP). Additionally, the LHC uses multiple preceding accelerators as its pre-accelerators.

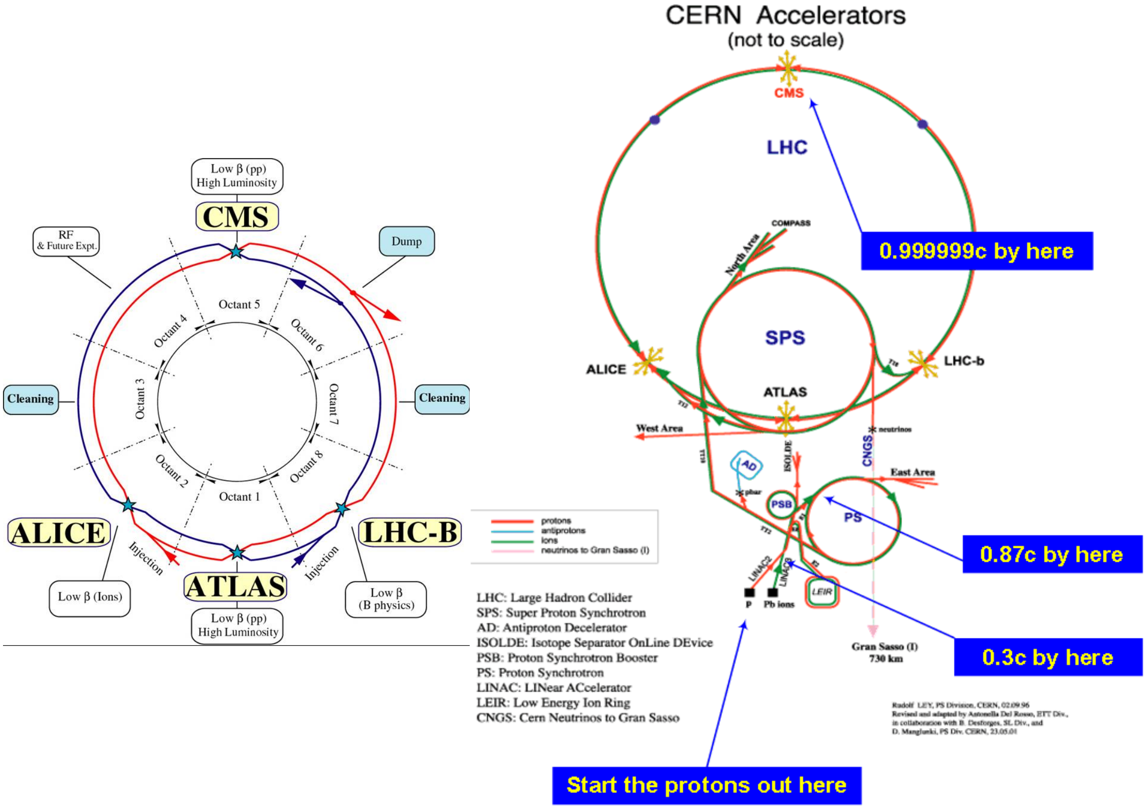

Low β (pp)
High Luminosity
CMS
RF
& Future Expt.
Dump
Octant 5
Octant 4
Octant 6
Cleaning
Octant 3
Octant 7
Cleaning
Octant 2
Octant 8
Octant 1
ALICE
LHC-B
Injection
Injection
ATLAS
Low β (Ions)
Low β (pp)
High Luminosity
Low β
(B physics)
CERN Accelerators
(not to scale)
CMS
LHC
COMPASS
North Area
0.999999c by here
SPS
ALICE
LHC-b
ATLAS
West Area
neutrinos
CNGS
ISOLDE
AD
East Area
pbar
PSB
PS
0.87c by here
LINAC2
LINAC3
LEIR
P
Pb ions
protons
antiprotons
ions
neutrinos to Gran Sasso (I)
LHC: Large Hadron Collider
SPS: Super Proton Synchrotron
AD: Antiproton Decelerator
ISOLDE: Isotope Separator OnLine DEvice
PSB: Proton Synchrotron Booster
PS: Proton Synchrotron
LINAC: LINear ACcelerator
LEIR: Low Energy Ion Ring
CNGS: Cern Neutrinos to Gran Sasso
Gran Sasso (I)
730 km
0.3c by here
Rudolf LEY, PS Division, CERN, 02.09.96
Revised and adapted by Antonella Del Rosso, ETT Div.,
in collaboration with B. Desforges, SL Div., and
D. Manglunki, PS Div. CERN, 23.05.01
Start the protons out here

Figure 3.1: The LHC structure (left) and its placement with respect to the pre-accelerators. [10]

The main purpose of the LHC is to produce collision events between protons or lead ions. The structure of the LHC including pre-accelerators is shown in the Fig. 3.1. The lead ions or protons are at first injected by linear accelerators in the circular Proton Synchrotron Booster (PSB). After this follows the even larger circular rings of the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS). Finally the particles are injected into the two LHC pipes, in which they rotate in opposite directions.

The LHC itself consists of eight octants. This structure follows from the design of the circular ring: it is not an actual circle but a polygon. It consists of alternating linear parts and magnet clusters that bend the trajectory into the correct direction. At some of the mid-points of the octants the tubes are crossed, allowing the particle beams to collide. It is noteworthy that already after the linear accelerators the particle velocities are getting close to the speed of light. Thus the LHC is greatly ruled by relativistic kinematics, making the design arduous. The high speeds and energies explain the need for a high quality vacuum in the LHC tubes.

Making single particles collide at speeds close to the speed of light would be very challenging. Therefore particles in the LHC are accelerated in **bunches**. These bunches are guided by the magnets that are placed into various parts of the LHC tunnels. Collision events are produced at the crossing points of the LHC tubes in **bunch crossings**. The time interval between separate bunch crossings was initially 50 ns but it has been recently upgraded to 25 ns.

Particle collisions are produced by focusing the bunches, using quadrupole magnets analogously to focusing lenses. If the particle beams are not focused at the time of a potential bunch crossing the amount of particle collisions is very small. It is important to maintain a high quality of the particle bunches and this requires strong magnets. The magnets used at the LHC operate at temperatures near to absolute zero.

The LHC itself is just a supplier of particle collisions. In order to benefit maximally of the LHC, multiple experiments are placed on its ring. There are four major experiments that are each located in their own octants at the LHC. Of these ATLAS (A Toroidal LHC ApparatuS) and CMS are the largest ones. They are designed for general purpose physics studies. Their primary focus is on proton-proton collisions, but also the lead ion runs are utilized for some studies.

ATLAS is placed on the Meyrin site - the main campus of CERN. The CMS detector is in the opposite end of the LHC in a relatively isolated location. However, most of the people working with these experiments do their work somewhere else than at the detector. Some are placed at CERN, but most of the research is done in universities all over the world.

There are many similarities between CMS and ATLAS. The general idea is to have separate groups studying the same phenomena to reduce detector-related errors. Moreover, there are great physical differences between these detectors, as well as the physics goals of the corresponding experiments. Because of their differences the two experiments excel in different tasks.

The two smaller detectors at the LHC are ALICE (A Large Ion Collider Experiment) and LHCb. ALICE has its main focus on lead ion collisions and it houses complicated physics studies. The LHCb on the other hand concentrates on b-physics, i.e. physics of particles containing the b quark. LHCb has recently found indications of a **pentaquark**, that is, a composite particle consisting of five quarks [11].

The smallest experiments at the LHC are paired up with the sites of the larger ones. The MOEDAL detector was recently placed at the LHCb site to observe hypothetical magnetic monopoles. TOTEM has forward region detectors placed at the CMS detector site and LHCf has correspondingly forward detectors at the ATLAS site.

The LHC and its detectors have been designed so that they can be upgraded step by step to reach higher collision energies and luminosities. The highlight of the first run was the finding of the Higgs boson. Recently in 2015 the second run began, after the first long shutdown and upgrade period ended. Some hints of potential new findings have already been seen [12]. It is possible that this run will result in significant physical findings.

3.2 The CMS detector

The ATLAS and CMS detectors are relatively similar but there are also distinct physical differences. As the name CMS[1] indicates, CMS is spatially more compact than ATLAS. At the same time it is the heavier of the two, totaling a weight around 14,000 tonnes. Furthermore, a distinct property of the CMS is the powerful solenoid magnet, also indicated by the abbreviation CMS. This work is done in the context of the CMS experiment, so from here on we only concentrate on it.

The CMS detector was installed into a cave previously occupied by a detector of the LEP collider. The detector is placed approximately 100 m underground. It is 22 m in length and 15 m in its diameter. The philosophy of the detector is such that the closer the equipment required by a measurement is to the beam pipe, the smaller and more precise it is. As one moves outwards in the radial direction, the detector layers catch particles that are harder and harder to observe. The measurements in the outer layers have typically a more coarse resolution.

In Fig. 3.2 a 3D presentation of the CMS detector is given. Often only the circular cross-section of the detector is presented, but a 3D model gives a more complete picture. An important distinction has to be made between the **barrel region** of the detector and the **endcaps**. These have the same basic structure, but the cylindrical structure of the detector is anyhow distinctively visible in the measurements. In many studies data from the barrel region are preferred so that the results are as uniform as possible. Within the barrel the tracks of charged particles are most effectively bent by the magnetic field. This property gives a better distinction between charged and chargeless particles.

It is instructive to study the structure of the CMS detector layer-by-layer. The innermost part of the detector is the **silicon tracker**. In radial direction it covers a zone ranging from around 0.2 m to 1.2 m. It is the most accurate information provider in the detector for the spatial and temporal positions of charged particles. The tracker consists of two important parts: the pixel detector and strip detectors.

[1] Compact Muon Solenoid

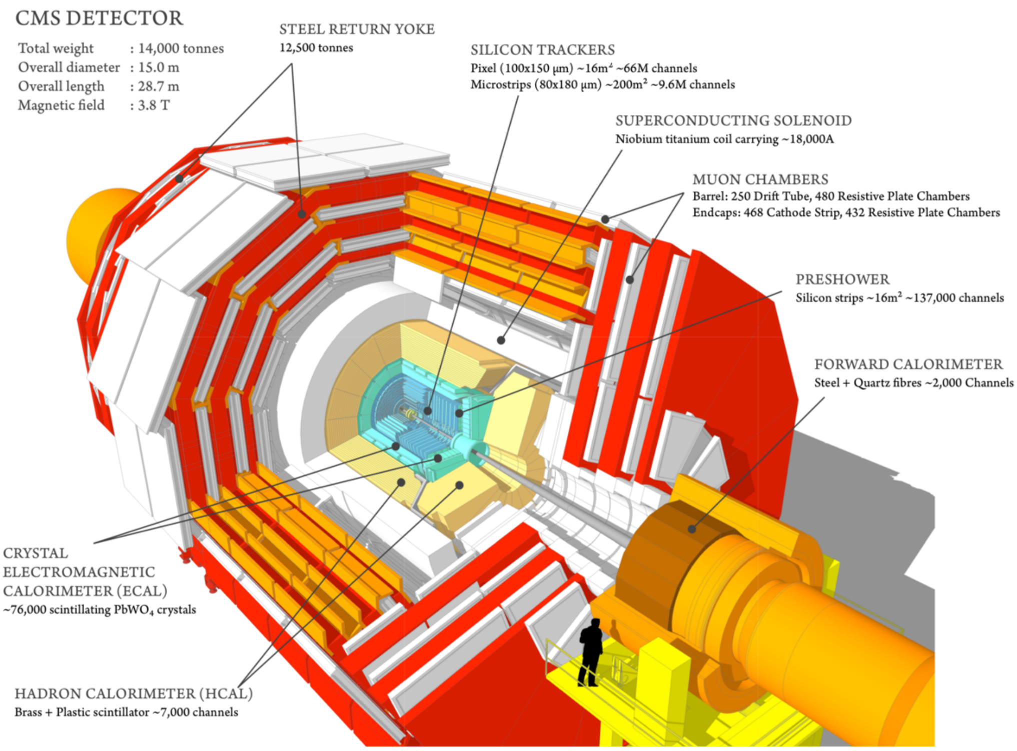

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 3.2: A full 3D model of the CMS detector. [13]

Currently the pixel detector consists of three layers of $100 \times 150\,\mu\mathrm{m}^2$ **pixels**. It is the closest part of the detector to the **interaction point** of the proton bunches. It brings the best measurement resolution as close to the collision as possible. A **phase I** upgrade is soon to be implemented to the pixel detector, adding another layer to it. This will bring the detector even closer to the interaction point.

The strip detector measures particle signals from silicon strips, which is very efficient when combined to the pixel detector data. Without the pixel detector a strip detector would be much less useful. The strip detector consists of four parts that follow the cylindrical geometry of the detector.

The next part of the detector after the tracker is the **electromagnetic calorimeter (ECAL)**. Its main purpose can be stated as stopping electrons and photons and measuring their energy. Also other electromagnetically interacting particles may leave traces to ECAL, but this phenomenon is relatively small. The important difference between photons and electrons is that the tracker does not observe photons. The ECAL consists of $PbWO_4$ crystals that cover an approximate region of $|\eta| < 3$. In the barrel region the front face of an ECAL crystal is $22 \times 22\,\mathrm{mm}^2$ and the radial length 23 cm. The endcap measures are almost the same.

The operation of ECAL is based on scintillation light produced in the crystals and discerned by photodiodes ($5 \times 5\,\mathrm{mm}^2$). Based on the crystal size and the operation mode of ECAL the resolution cannot be as good as in the tracker. The scintillation light cascades to the diodes in a spread-out pattern.

The following layer is the **hadronic calorimeter (HCAL)** that concentrates on stopping hadrons and capturing their energies. In contrast to ECAL it consists of consecutive layers of plastic scintillators and brass. With all of its parts taken into account, HCAL reaches up to $|\eta| < 5.2$. As opposed to ECAL, the particle material arriving to HCAL can be sorted rather poorly. There are various types of hadrons, but only charged and neutral hadrons are distinguished, according to the tracker and ECAL signals. The HCAL reaches radially from 1.77 m to 2.95 m. It is organized into towers of detector elements; the exact shape of these varies according to the placement in the detector.

Outside the HCAL is the massive solenoidal **magnet** of the CMS. It produces a magnetic field of 3.8 T within the detector. Outside the solenoid the field patterns are captured by a large iron yoke. It allows a magnetic field of 2 T in the outermost parts of the detector.

The final detector layer of the CMS are the **muon chambers**. These are integrated with the iron yoke, which provides the chambers with a good-quality magnetic field. The main purpose of the gaseous muon chambers is finding muons and measuring their momenta. Unlike the previous detector layers, the chambers do not try to stop the muons, which simply fly through. Apart from muons, most of the known particles are captured by the CMS detector. The only known outgoing particle type not detected are the weakly interacting neutrinos. These are observed indirectly as missing transverse energy (MET) in the measurements. Contradicting its name, MET is not an actual energy but a 2D momentum vector in the transverse plane.

The structure of the CMS detector hints of the nature of the experimental studies in particle physics. A large priority is given to electrons, muons and photons. Hadrons are detected in a more generic fashion. This is largely due to the fact that the hadrons constitute jets. The exact hadronic structure is not very important. On the other hand, photons and charged leptons are important signals for rare and interesting event types. Guessing the underlying process only based on jets is difficult. On the contrary, a given amount of leptons or photons provides very clear event signals.

3.3 From signals to event data

The technical description of the CMS detector leaves us with a bunch of signals produced in various parts of the detector. A considerable effort is made in putting these signals together to reconstruct the collision events. The potential overlap of events in time and space requires very accurate signals from each detector layer.

Even if the bunch crossings occur every 25 ns, this does not cause additional difficulties, thanks to the high quality of the detector hardware. On the other hand, the spatial overlap is a real problem. In each bunch crossing, on average 20 or more significant proton-proton interactions occur.

Inaccuracies and other phenomena related to the overlap of collisions are treated under the title of **pileup**. To handle the spatial overlap of collisions, a very effective reconstruction of the particle tracks is required. Thus the particles can be connected to **primary vertices**. This makes it possible to handle the particles from different collisions separately. In each event the primary vertex is the point where the primary proton-proton interaction occurs.

Another important prerequisite in the CMS data acquisition are the **trigger** systems. Triggers are used for selecting the most interesting collisions from the multitude of events. If an early-stage selection was not done, the required storage capacity would be overwhelming. A coarse selection is done at the **level 1 (L1) trigger**. It consists of low-level hardware that makes a very generic type of an event selection.

In a classical detector design the L1 trigger is followed by L2 and L3 triggers. In the CMS these have been replaced by the **high level trigger (HLT)**. As an advantage to the classical design, the HLT is more freely programmable. It does a quick reconstruction of the events and applies a more complicated event selection. The event selection at the triggers has to be performed very quickly, which restricts the complexity of the processing. The fraction of the data that survives the triggers enters a long progression of further processing and analysis.

An important part of the data analysis chain is the **Particle Flow (PF)** algorithm. It is here that the particles are labeled as electrons, muons, photons and neutral or charged hadrons. This kind of data sorting facilitates further processing of the particle data. For instance jet analysis and the study of missing transverse energy can be performed for PF particle data. The ideology of the PF algorithm is that it takes a maximal advantage of each part of the detector. Signals arriving from the subdetectors are combined to obtain a full picture of a collision event. It is not said that it is the ideal algorithm, but it has performed very well, so far. A new PF-based algorithm, PUPPI, has been under development for some time. However, it is not yet as mature as PF.

An accurate event reconstruction includes many challenges. One example of this is that in the PF algorithm all charged hadrons are labeled as π^+ or π^- and neutral hadrons as K_L^0. As we can recall from Fig. 2.3 from the previous chapter the charged pions cover most of the charged hadrons. On the other hand K_L^0 and K_S^0 have the same masses and together they cover a large portion of the neutral hadrons. However, there are also other hadrons present, which are thus mishandled. The reason for these choices is the fact that many hadrons cannot be distinguished from each other at the detector level. Only with an assigned value of mass, the energies and momenta of the particles can be connected properly. Since certain mass values are used, the most frequent ones are the best choices. In addition, the error in the energy-momentum relations is significant only for particles with small energies.

In some cases one can think that the PF algorithm could be optimized even more. However, the greatest shortcomings of the algorithm are related to the non-optimal properties of the detector. For instance, especially the neutral hadron energy is drained already by the ECAL. Because of this the neutral hadron energy needs to be weighted upwards. On the other hand, a very energetic photon can in principle deposit some of its energy into HCAL. These phenomena are known, but usually they can only be handled statistically. The complete picture of an event is always an approximation, based on a multitude of calibrations.

3.4 CMS Software

A final essential link in the analysis chain is the CMS software (CMSSW). It is a massive software compilation that consists of a C++ skeleton complemented by PYTHON scripts and some legacy FORTRAN code. Unlike some sensitive analysis material, CMSSW is completely available at github, see Ref. [14]. Even if the code is available, the software is usually run at CERN servers. This allows the utilization of a large network of computing resources. The central philosophy of CMSSW is to provide a common interface to a plethora of software packages related to HEP. It has the ability to perform every step from simulations to data processing. The user can quite freely choose the software packages to be used. The analyses can be run either for simulation results or for real data.

In a complicated multi-step physics analysis a software structure like CMSSW is a logical choice. It is more simple to leave the handling of the program interfaces to the software experts. Thus an average user can concentrate on making their own analyses. This usually concerns only a small part of the whole CMSSW.

Aside from the listed good features there are also some significant disadvantages in CMSSW. Because of the massive size of the software and its multi-purpose ideology, it accommodates a bureaucratic system. Few people have a good understanding of all the stages of an analysis process. There are also conflicting needs and requirements presented by various groups in the CMS experiment. Making a desired change to the software can thus be infeasible or at least very slow. Thus experimentation with new methods can at times be very challenging. Also for instance the official simulation samples are provided at a rather slow pace from the dedicated generator group.

From the given perspective it is easy to see that working outside of CMSSW has its benefits. Therefore, in this work it was decided to use the software packages outside of the whole CMSSW context. The final and standardized CMS analyses, however, have to be made with CMSSW. Nevertheless, for now a separate implementation allows us to experiment more.

Chapter 4

Methods: Monte Carlo event generation

In this chapter we review the full process of modern Monte Carlo event generation. Historically, the steps of the simulation might not have been as distinct as they are today. An increased modularity is in part due to the transition from FORTRAN to C++ and class structures. This allows software packages to concentrate only on certain parts of the simulation process. However, the most important packages are general purpose Monte Carlo event generators (GPMC), which can do all the simulation phases. If necessary, these can be supplemented with other software packages.

The initial proton-proton collision is usually modeled with a **hard process**. In the simulation context this means a high-p_T leading order (LO) or next-to-leading order (NLO) process that defines the general structure of the event. The hard process is supplemented by **parton showers** to make the event structure complete and realistic. The beam remnants are handled in **multiple interactions**. The transition from parton level to hadron level is difficult to model. This is executed in **hadronization**. After hadronization the last step is to make some of the particles decay and to add a detector simulation if necessary. This work focuses on using the GPMCs Pythia 6, Pythia 8 and Herwig++. The use of supplementary software packages is reviewed to gain a more comprehensive picture of all of the possibilities.

The following text relies heavily on references provided by the event generator groups. Especially the Pythia 6 manual [15] is very thorough and instructive. It functions as an excellent supplementary material for the brief Pythia 8 [16, 17] and Herwig++ [18] manuals. In addition, the Les Houches guide [19] gives a good insight in the differences of the common event generators. Finally, the paper provided by the CTEQ group [20] handles the event generator topics on a purely theoretical level.

4.1 Simulation phases

In the following the steps of event generation are described, mostly in the chronological order.

4.1.1 Hard Process

The hard process is the central piece of most event generation schemes. It serves as a simplified picture of the proton-proton collision. In leading order (LO) simulations it is modeled as a $2 \to 2$ process with two incoming partons. The incoming partons are thought to be originated one from each of the colliding protons. Thus the flavors of these partons are obtained from proton PDFs.

The central idea of the hard process is that the collision of the two partons is the **hardest** part of the proton-proton interaction. That is, greatest p_T values are observable in this process. Executing the hard process leaves the system in a rather complicated state. There are the end products of the hard process itself, which usually require non-trivial QCD-based handling. On the other hand the hard process itself is just a skeleton and a better model is achieved by adding **parton showers** to it. Finally, there are the partonic remnants of the colliding protons. Also these need a thorough QCD-based processing so that a final state can be reached.

At this point it might seem that the hard process causes only confusion. It is, however, a necessary and effective way to begin the generation of a collision event. A great benefit related to the hard process is that it has a great impact on the structure of the whole event. Thus by selecting a certain type or certain types of hard processes, one can easily focus on simulating a chosen set of events. An especially convenient feature is that the hard process can be used for early-stage pruning of simulated events. This allows a significant speed-up for the simulations. The hard process is only a small core of the total interaction and decay tree of the particles.

A good example of early-stage event selection is $\hat{p}_T$ event weighting. The *hat* indicates that we are speaking of a quantity in the final state of the hard process. In practice $\hat{p}_T$ is the transverse momentum of the outgoing hard process particles in their center of mass coordinates. Modern event generators usually have a zero center-of-mass momentum in the hard process, so no changes of coordinates are necessary. The weighting is used to give a $(\hat{p}_T/p_T^{ref})^{4.5}$ bias weight for all events. Since the outgoing hard process partons are usually precursors of jets, the p_T values of the hardest jets are typically correlated with $\hat{p}_T$. This kind of a selection leads to the production of more high-p_T jet events than observed in nature. The selection mimics the jet p_T dependent event selection done at the CMS detector.

Typically the GPMCs (Pythia, Herwig, etc.) are equipped with LO event generation tools for the most important processes. Nevertheless, there are also special MC generators that are only focused on generating the hard process. These are usually designed so that a general-purpose event generator can carry on the simulation after the hard process. The hard process data is transmitted in the form of Les Houches Accord Event files (LHE). This makes the simulation process more complicated, as the two event generators need to communicate interactively through ASCII files. In the LHE context the hard process is called the **signal event**. Hard process types are analogous to the *signals* used by the detectors for saving and labeling data. This emphasizes the impact of the hard process on the total generated events.

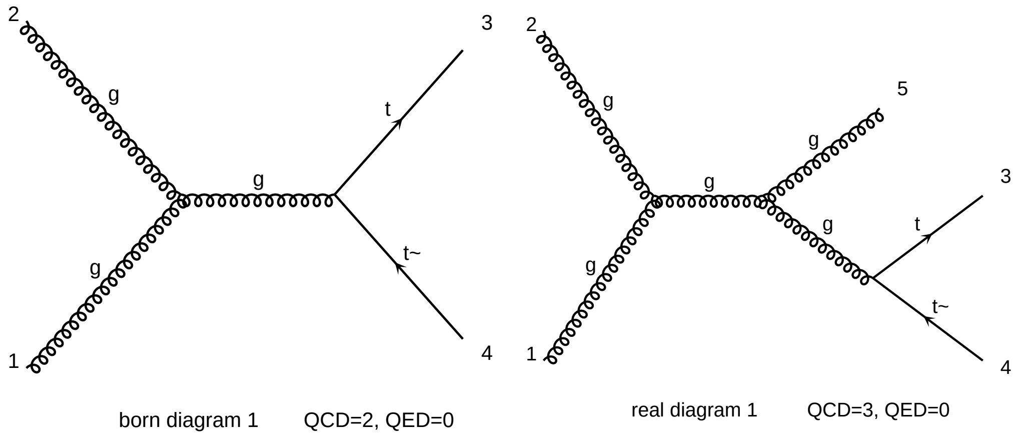


Figure 4.1: Examples of LO (left) and NLO (right) Feynman graphs in the $\mathrm{gg} \rightarrow \mathrm{t\bar{t}}$ process (created by MadGraph).

Typical generators used at the CMS for signal process generation include MadGraph [21] and PowHeg [22]. These are popular since they can be used to create a large variety of processes. Remarkably, these allow also the generation of NLO hard processes. In practice NLO production adds one vertex to the Feynman graph of the event studied. Typically this leads to a hard process final state with three particles. In Fig. 4.1 examples of generic LO and NLO hard processes are shown.

Of the two signal process generators MadGraph is more favorable in the way that given an initial and a final state it can be used to create any LO or NLO hard process. In addition, NLO creation in PowHeg may occasionally produce negative event weights. Without going into further detail [23], this is not a physically reasonable behavior. The general hard process generation is done by defining initial and final states. The Feynman graphs with a maximal number of QCD vertices and a minimal number of QED vertices are generated. The particle types at each phase can be chosen very strictly or in a generic manner. The amount of vertices is determined depending on whether LO or NLO event generation is in question.

Event generation on the NLO-level has become possible only in the recent years. There is a strong theoretical motivation to move into NLO generation, especially as the LHC energies are getting higher. With higher collision energies there will be more and more jets present. As the outgoing products of the hard process become jets, it would be optimal to have as many outgoing particles as possible available. However, the view is not in reality as simple as one could think. Parton showers and multiple interactions are a method for producing the whole remaining event. The NLO parton showers are troublesome as compared to the LO case.

Since there are various problems that can be related to NLO production, in this work only the standard LO generation is used. An NLO extension would be desirable. However, given the presented problems, it is not necessarily a source of better results. On the other hand, next-to-NLO (NNLO) generation techniques are currently being developed. These methods are even further from reliable operation than the NLO generators.

In general one should pay close attention to the assumptions and choices made by using hard process event generation. These days it is relatively easy to use an event generator. It is much more complicated to truly understand the purpose of the various levels of the simulation. Basically the hard process is the root and the reason for the whole structure of an event simulation.

To conclude, some attention is given to an event generation mode that is labeled in Pythia as Soft QCD. The hard process event generation is focused on events with hard parton interactions. That is, events in which QCD processes occur between the colliding protons. However, in recent times also the more generic production of Soft QCD has come along. This allows the production of all kinds of collision events including e.g. diffractive collisions. In certain physics analyses these kinds of events are essential. However, the focus of this work is in jet physics and thus only the Hard QCD type of events are considered. As a curiosity, typically also Soft QCD is based on an event structure that begins with a hard process. In this case the p_T values of the hard process are minimized, allowing a different kind of event generation.

4.1.2 Initial and final state radiation

The so-called *parton showers* are an essential part of modern collision event generation. They are typically sorted into two categories: initial and final state radiation (ISR and FSR). These serve as correction terms for the otherwise very plain hard process. They allow a better accuracy than pure analytical calculations. Without the showers one can only reach a definite order (e.g. NLO). In the event generation cycle the two radiation modes are applied right *after* the hard process. The radiation is initiated by outgoing gluons and photons that are added to the initial and the final state of the hard process. The naming conventions vary to some degree – ISR and FSR can be referred to as initial and final state showers. In Fig. 4.2 the basic ISR and FSR structure is presented. The split of parton showers into ISR and FSR is not fundamental, but it is a very convenient choice for the simulation machinery.

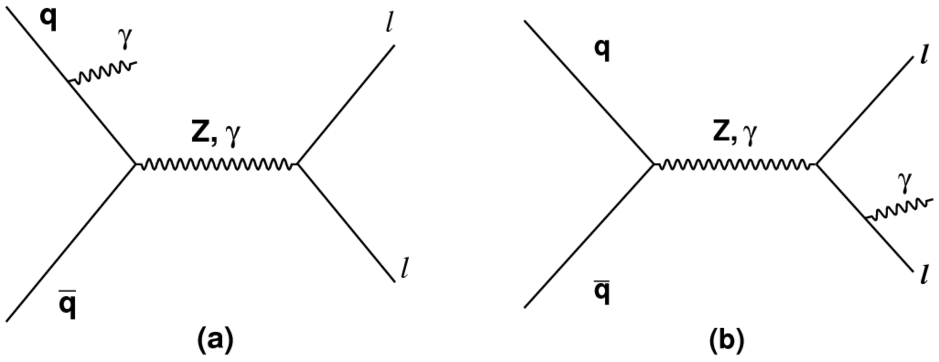


Figure 4.2: Simple ISR (a) and FSR (b) example scenarios. [24]

The initial and final state showers can become very complicated, but the underlying mechanism is relatively simple. Starting from either the initial or final state of the hard process branching is applied. The branching is essentially of the types q → qg, q → qγ, g → gg and g → q$\bar{\text{q}}$. Further branching of photons can be left out in the simple cases. The branching mechanism is somewhat similar to the use of parton distributions in the proton-proton interactions. For each potential branching $a \to bc$ a fraction z and an evolution scale Q^2 is defined. The fraction z describes how the momentum of a is split between b and c. On the other hand the virtuality scale Q^2 describes approximately the time-ordering of the branching processes.

As the showers are generated after the hard process, there are very distinct differences in ISR and FSR. FSR can be viewed as a normal parton system evolution through branching. The FSR particles have timelike virtuality, $m^2 \geq E^2 - |\mathbf{p}|^2$ and a positive virtuality scale. Hence, FSR is also referred to as **timelike showers**.

On the other hand, ISR needs a more complicated handling and is not understood as well as FSR. The initial state radiation has a spacelike virtuality with respect to the hard process: $m^2 \leq E^2 - |\mathbf{p}|^2$ and Q^2 is negative. Starting from the initial state branching is performed backwards in time. This continues until a pre-defined scale Q^2 is reached. Due to the nature of ISR a natural terminology of **spacelike showers** arises.

A most important fact about ISR and FSR is that they change the meaning of LO and NLO completely. Both LO and NLO hard processes are simply skeletons of events in the HEP context. The difference between LO and NLO is dramatic, but not even NLO reaches the complexity of a complete event. When showering is added, numerous branchings are performed and the final states of LO and NLO are not that different. Because of the complexity of the NLO matrix elements LO can be more favorable than NLO.

With the showers comes also the disadvantage that all the calculations are more and more approximate. With respect to the more or less accurate hard process matrix element calculations the parton showers are mere approximations. However, relatively good results can be reached with an insightful use of parton showers.

Considering the parton showers there are various ways to define the evolution parameter Q^2. There are numerous theoretical views on the different definitions and their pros and cons. It is not instructive to go through the details here, but it is acknowledged that the theoretical basis is solid [19]. An older definition used a simple virtuality ordering, with Q^2 proportional to the squared invariant mass m^2. In the modern generators especially p_T-ordering and angular ordering are popular. In p_T-ordering branching is performed with descending p_T values. The process of angular ordering is more involved. In short, the word *angular* refers to angles between the daughter particles and the initial particle of a branching. The evolution starts with large angles between the particles, then proceeding to smaller angles. An additional important remark is that the evolution scales Q^2 are different for ISR and FSR.

The benefit in using this kind of z- and Q^2-values is that they work well with probability distributions. Thus similar formulations are seen within the hard process, showering, as well as in multiple interactions. In Monte Carlo simulations it is of essence to have the system evolution probabilities available. We will not go further into MC theory but the main idea is to have a method for studying systems with a high-dimensional phase space. Covering all the possible final states and integrating analytically over them is not always possible. However, with sufficient sampling the MC *integrals* approximate the analytical results. Using a random number generator, random samples of the total phase space are taken. When sufficient sampling is performed, a view of the whole process is obtained. A notable benefit of the MC techniques is that they resemble very much the intake of data at the detectors. Thus the simulated and gathered data samples are analogous.

4.1.3 Multiple interactions

Multiple parton interactions (MPI) are a distinct part of the simulation chain and cannot be overlooked. The core of MPI is evolving the remnants of the protons in a physically motivated way. Due to color confinement it is clear that one cannot just forget the remaining partons within the protons. The process of MPI is not fully understood and therefore it is a sort of a nuisance in the simulation chain. In the end the handling of MPI is somewhat similar to that of ISR and FSR and thus it should be given the status it deserves.

The focus of MPI is in treating the possible additional interactions between the beam remnants. In more sophisticated models also the effect of ISR and FSR needs to be taken into account. Thus for the beam remnants the question is what kind of interactions are going to take place. In the most simple case no additional interactions would occur and the beam remnants would be transferred directly to the hadronization phase. The usual MPI models depend on a few parameters, which can be tuned to give results that agree with experimental results.

4.1.4 Hadronization

Together with the showering and multiple interactions, hadronization is a very important part of GPMCs. In some sources hadronization is called fragmentation. It is the process of formation of colorless parton clusters from the parton matter. The partons turn into hadrons, which are observable at the detector. The partonic evolution scheme until the end of the hadronization can be considered somewhat unphysical, as the behavior of quarks cannot really be observed. For hadronization there is still no solid physical theory. However, there are various models that give a fairly good approximation of the process. All in all, the good success of MC event generators suggests that the parton-level implementation of the simulations has a good physical motivation.

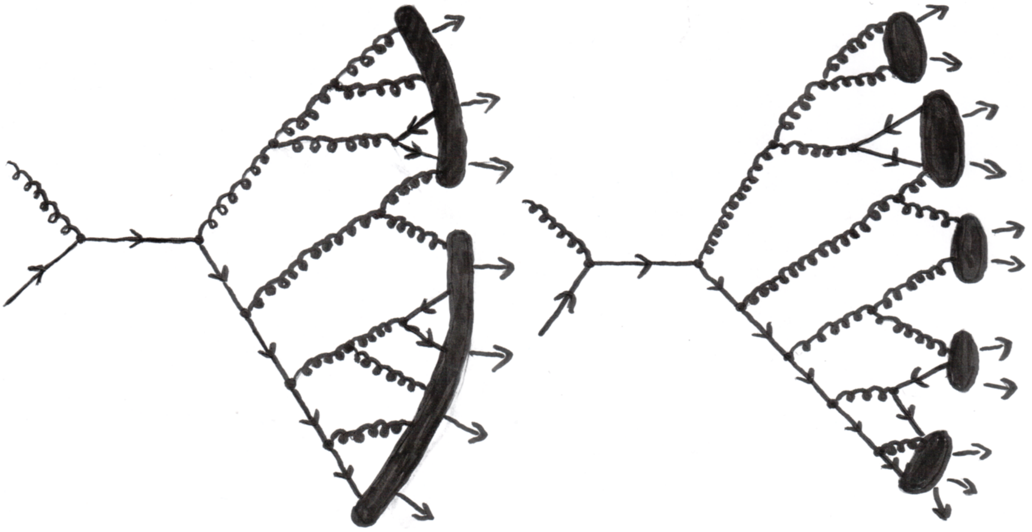


Figure 4.3: An illustration of string (left) and cluster (right) fragmentation.

The most interesting and successful hadronization models presently are the Lund string model and the cluster model. An illustration of these models is given in Fig. 4.3. The great challenge of hadronization is to collect the produced partons in a sensible way into lumps. The partons within a lump should be colorless and fairly close to each other in the phase space.

The string model is motivated by an observation made with help of lattice QCD simulations. The strong force grows linearly as a function of distance - thus giving a physical motivation for color confinement. When the distance between partons grows enough it is energetically favorable for a quark–anti-quark pair to appear.

Thus the partonic system can be thought to consist of such strings, into which more quarks can develop. As the partons travel quickly away from the proton interaction point, these strings eventually break up into smaller pieces. The development is pushed onward so that finally a state with hadronic color singlets is reached.

The cluster model takes a slightly different approach. It is based on a theoretical result according to which the partons are arranged into colorless clusters. This requires that their evolution scale (Q^2) becomes small enough, compared to that of the hard process. Thus the post-showering partons are organized into clusters, which are further evolved into hadrons.

Hence there is a physical motivation for both of the models, but there is no reason to prefer either one of them. The lattice-QCD simulations are built on solid theoretical principles, but currently they can be applied only in simple processes. High-energy collisions provide an unstable environment of QCD interactions, in which theoretical considerations fall short. The self-coupling of gluons makes calculations chaotic. Thus the hadronization models can be appreciated, as they provide results of good quality in a complicated system.

4.1.5 Decay

In comparison to the previous stages of the simulation the decay processes are relatively simple. After the completion of the hadronization process there is no need to return to the parton level view of events. The post-hadronization particles are quickly traveling into all longitudinal and transverse directions. Thus, further mutual interactions are not anymore significant. On the contrary one has to take into account the lifetimes of various particles. For some particles the lifetime is so small that they are never observed at the detector level. Other particles may be unstable, but their decay occurs on average only after reaching the detector. In the CMS simulations, usually particles with a lifetime under $10\,\mathrm{mm}/c$ are considered unstable.

The decay processes are fairly simple. Only for a few particle types the mean lifetime corresponds to the detector dimensions. It is more typical that the lifetime is very long or very short as compared to the detector length scales. Even for particles that might or might not reach the detector, it is effortless to simulate the Poisson process of decay. One simply needs to take into account the decay modes of each particle. In all aspects the decay process is one of the simplest simulation stages. Nevertheless, it should not be neglected. Only due to the decay processes one reaches a realistic final state in the simulation.

4.1.6 Detector simulation

As a last part of the simulation chain, the simulation of detector effects is worth considering. In this work it has been decided to leave the detector simulation out for a couple of reasons. Since CMSSW is not utilized, simulation of a detector would require a significant effort. On the other hand, it can be argued that the present studies can well be made without a proper detector simulation. However, it is important to recognize the consequences of neglecting the detector simulation.

Event generation can be motivated by various reasons. A typical way to use simulations is to compare simulated PF data to the simulation *truth*. That is, all the detector properties and signal processing are applied to simulated collision event data. This gives valuable information for understanding the data gathered by the real detector. However, a full detector simulation is not very useful in the consideration of theoretical aspects.

If the detector effects are left out we get a sort of an idealized picture of the detector data. This is a useful view of events when considering the theoretical aspects of jet physics. Studies of the simulation truth can help in giving a better understanding of the behavior of jets. Moreover, from the point of view of jet physics analyses the detector simulation is not crucial. After calibrations and corrections to the recorded data, the difference between ideal simulated data and detector data is reduced.

In the CMS project as well as most of the other great collider projects, detector simulation is performed with GEANT (GEometry ANd Tracking) [25]. The most important parts of a detector simulation are the implementation of detector geometry and materials. While moving through space the particles interact with the detector and can slow down and decay further. Furthermore the final detector response has to be simulated at this phase. Thus taking into account the physical properties of the various detector parts, a final realistic picture is obtained. The detector response is only implemented statistically, but this coincides reasonably well with the randomness of particle-level events.

4.2 Event generation software

We have now gone through the full process of event generation in the CMS experiment. Additionally, some light has already been given on some of the relevant software releases. Now we go through the role of all of the most important software distributions in the present context.

4.2.1 Pythia 6 and Pythia 8

The most popular GPMC is Pythia. Despite its popularity there is still a version conflict going on within this software. The last FORTRAN version, Pythia 6 [15] became very well established. It has been in the use of the projects at the LHC until the most recent changes. The reason for the popularity of Pythia 6 is that after a long use the users have become very familiar with this software. The use of the new C++ adaptation, Pythia 8 [17], is well motivated as it includes upgraded physics processes. However, change to the newer version has been slow. Also in this work the old Pythia 6 is used in addition to Pythia 8 to be able to see the possible differences. In principle the physics in the versions 6 and 8 of Pythia are quite similar. Nevertheless, as we later see there have been significant changes during the version gap.

To understand the context of the event generators better it is useful to have a full historical view of Pythia. At its time in the year 1997 the release of Pythia 6 was a big step forward. The previous version, Pythia 5 was merged with the latest version of Jetset to provide a more complete package than before. In their previous versions both of these software packages have had a colorful history. Both computers and particle accelerators have progressed much while they have been developed.

The release of Pythia 6 was well-timed so that the software became an important tool in the beginning of the LHC era. There was a pressure to move forwards to a more modern C++ version, which induced the development of Pythia 7. The development of this version resulted in the creation of ThePEG, which provides the basic structures for a further implementation of event generators. The philosophy of ThePEG is that the event generators should provide all the physics; ThePEG takes care of the infrastructure. However, it was later decided that using a common basis for different event generators would not work. Thus, the table was cleared and Pythia 8 was rewritten from the beginning. Now, in the beginning of 2016 Pythia 8.2 is finally reaching a status above its predecessor, Pythia 6.4.

4.2.2 Herwig++

Herwig [18] is another GPMC that has been running for a long time. It has faced many similar problems and changes during its history as Pythia. Also for this software the transfer from FORTRAN to C++ has been a great challenge. For various reasons Herwig has not been as popular as Pythia at the LHC. One of these is the fact that only in the end of 2015 the C++ version Herwig++ 3.0 finally fully superseded the previous FORTRAN version, HERWIG 6. This new C++ version received the title Herwig 7. In this work the last version preceding Herwig++ 3.0, Herwig++ 2.7.1, is be used.

The histories of Pythia and Herwig are closely intertwined. Herwig++ ended up using ThePEG as its basis. Even if ThePEG was supposed to be a generic base for any custom generator, Herwig++ soon became its most important partner. At the time of the release of Herwig 7 also ThePEG 2 was released.

As it comes to Pythia and Herwig one needs to be aware of the possible major differences between these two GPMCs. The troublesome thing in simulation software is that the differences might always be due to bugs in the programs. One can only make the code as good as possible and look for physical discrepancies. If a bug manifests as small changes in the physical results, it can be very hard to identify.

Aside of bugs there are some fundamental differences between these generators. As the biggest ones come the hadronization model and parton shower evolution. Pythia uses p_T ordered showers, as Herwig resorts to angular ordering. On the other hand Pythia uses the Lund string model for hadronization, as Herwig utilizes the cluster model. Thus comparison between Pythia and Herwig functions as a good source for error estimates. This kind of a comparison can be used for both bug tracking and physics comparison.

4.2.3 Other event generators

There is a multitude of event generators available. Anyhow, it is not arbitrary that Pythia and Herwig have been selected into this work. GPMCs are relatively rare and therefore the choices are quite restricted. On the other hand, to keep up with the upgrades of the LHC, the generators need active and strong development. Only a few projects have the sufficient resources for this.

A new GPMC, Sherpa [26], has recently been under development. Unlike Pythia and Herwig it has only a relatively short history behind it. It has been collected from a group of pieces of software to cover the full process of event generation. Sherpa presents good new opportunities for event generators. As the inter-generator comparison is the only reasonable method for error evaluation, this is very valuable. However, Sherpa is not yet as established as the two other software packages. Thus it was not included into the scope of this work.

4.2.4 Event generator tunes and LHAPDF

One last remark of the event generators has to be made concerning **tuning**. In practice this means setting the values of a group of variables so that they best correspond to the experimental results. Making good tunes is very intricate business. The tunes are often closely connected to the parton distributions in use. Additionally, the use of different PDFs in the hard process and the rest of the event usually requires new tunes. This can create complications for instance when using MadGraph.

The software package LHAPDF [27] is the commonly established source for PDFs. Usually the event generators have some PDFs internally included, but LHAPDF allows access to any PDFs. It is used also in this work for the PDF interface. Tunes are provided by the event generator groups but also for instance by the CMS collaboration. In this work we aim to use the CMS tunes and the corresponding PDFs. For simplicity the same PDF is used for the hard process and the ISR-FSR-MPI phase.

Chapter 5

Methods: Analysis of particle level data

The event generators are used for creating simulated particle data, but nothing more. Complicated analysis software is necessary for obtaining results from the raw data. Since this work does not utilize the CMSSW there is some freedom in the implementations. This also necessitates the implementation of some functionalities normally provided by the CMSSW.

Jet clustering is the most important part of the analysis process. This serves as a basis for all further analyses. As this work is concerned with jet flavors, it is important to have a thorough look on jet flavor definitions. The jet clustering process is strongly tied together with the flavor tagging. A special case of flavor tagging is b-tagging, which plays a large role in the CMS experiment. Finally, the quark-gluon variables provide a useful set of variables to compare different types of jets.

Since various event types are studied, the event selection parameters are of high importance. These are often related to the physical reality of the detector and the ability to distinguish certain events. The choice of parameters is to some degree arbitrary. Good choices take into account the energy scales and the objectives of the measurement.

As a conclusion the measurement of the top mass is studied. This is a non-trivial task, as the t quark needs to be reconstructed from the final-state particle information.

5.1 Jet clustering

Jet clustering is the backbone of all analyses and thus it requires a special emphasis. Ideally a clustering algorithm should be very robust. The jet structures should stay constant under small perturbations of the provided event data. For instance, changes in the particle list ordering should have no impact on the results. On the other hand a jet clustering algorithm should have a good computational performance. Since the clustering procedure is frequently repeated, it would otherwise increase the required computation time severely.

One necessary limitation for the current jet sorting algorithms is the definition of a maximal jet radius in the (y, ϕ)-plane. Thus wide jets are split into several jets by the algorithms. However, it is usually difficult to distinguish a wide jet from two overlapping jets. One possible approach is to discard such ambiguous events.

At the moment the most popular jet sorting algorithm is the **anti-k_T** algorithm [28]. This is presently used in the CMS analyses as well as in this work. An efficient implementation of the algorithm is found in FastJet [29], which is the leading software package for jet clustering. The implementation of anti-k_T found in FastJet has most of the features mentioned above. The greatest drawback of the obtained jets is that they are not always representations of single partons. This can result from wide jets or other similar scenarios in which the algorithmical logic is not ideal.

To understand the properties of the obtained jets better it is useful to review the anti-k_T clustering algorithm. A trial jet that can receive more particles is called a **pseudojet**. The algorithm involves two distance measures for entities, which are particles or pseudojets. The distance between two entities i and j is characterized by

$$d_{ij} = \min(k_{Ti}^{-2}, k_{Tj}^{-2}) \frac{\Delta R_{ij}^2}{R^2}, \tag{5.1}$$

where k_T are the corresponding transverse momenta, ΔR_{ij} the (y, ϕ)-plane distance and R the jet radius. On the other hand the *transversity* of an entity i is characterized by

$$d_{iB} = k_{Ti}^{-2}. \tag{5.2}$$

At all stages of the clustering all the distance measures from Eqs. (5.1) and (5.2) are gathered into a single list. The smallest of these distances is taken into further inspection. If a d_{ij} type of distance is smallest, the elements i and j are combined into a new pseudojet. If a d_{iB} type of distance is smallest, the pseudojet i is declared as a final jet. The *anti* in the name of this algorithm refers to the negative exponent of the k_T values. The anti-k_T algorithm is usually preferred over other k_T algorithms with different exponents.

Concerning the jet clustering of simulation data, the particles given to the algorithm have to be selected carefully. Neutrinos should not be taken into account, as they are not observed by the detector. However, the inclusion of neutrinos does not have a great impact on the results. As shown in Fig. 2.3, the contribution of neutrinos is very small to the total energy of ideal jets. All charged leptons can be taken into jet clustering, excluding the analyses where the leptons are used as signals or indicators. This means that charged leptons from the hard process are typically excluded from jets.

5.2 Event selection

In this work we study four different event types. The event generator specific recipes for generating the events are reviewed in Appendix A. We use the term **standard candle event** for three of these event types. These are the γ+jet events, the Zμμ+jet events and dijet events. The expression is well motivated at the CMS, as these events are easily distinguished based on the detector signals. Therefore these kind events are excellent for instance in studies of the jet flavor. The fourth event type is one that involves the production of $\mathrm{t\bar{t}}$. The definition of this event type is intricate, but it is one of the most effective ways to access the top quark mass. Due to the large mass of the top quark, all events involved with a t quark are more complex than the standard candle processes.

In Pythia and Herwig these events are defined by the hard process. Thus a great amount of effort is saved since we can concentrate on simulating directly the events of interest. However, there are often also other kinds of physical selections (**cuts**) that one wishes to do for the events. On one hand, it is useful to do similar cuts that are done for the data gathered by the CMS detector. Even if we know that the hard process is of the desired type, it can manifest in an ambiguous form at the detector. On the other hand, quality cuts are often necessary. Events with overlapping jets and problems with flavor tagging are not useful in further analysis. Also for instance too large $|\eta|$-values or too small p_T-values give reason for rejecting events from the analyses. At the detector such values imply a lower quality of the data gathered.

5.2.1 Dijet events

Dijet events are distinguished by two dominating jet signals. These jets originate from hard process partons and there can be additional jets from ISR, FSR and MPI. The hard process in dijet events is a fully QCD-based process: the two outgoing particles can be practically any two partons. Nevertheless, the QCD processes of t, $\mathrm{\bar{t}}$ and $\mathrm{t\bar{t}}$ production are excluded. These processes differ greatly from the more generic partonic 2 → 2 processes. The Feynman graphs of the dijet hard processes are shown in Fig. 5.1.

The p_T-values of the jets give a useful tool for selecting events in which the radiative effects are weak as compared to the dijet signal. We want to choose events for which the p_T values of the two leading jets are much greater than that of a potential third jet. This is conveyed through the parameter

$$\alpha = \frac{2p_{T3}}{p_{T1} + p_{T2}}. \tag{5.3}$$

In this work we require $\alpha < 0.3$. Typically it is also required that the two leading jets are of a back-to-back nature. This is seen from having a large enough azimuthal difference between the jets, here we use $|\Delta\phi| > 2.8$. It is also beneficial to have

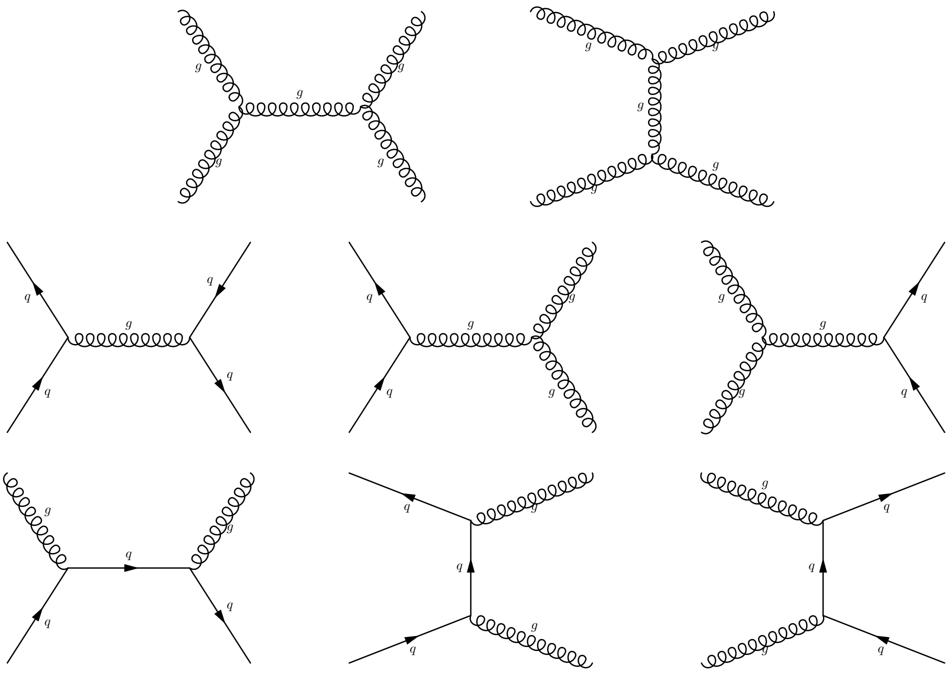


Figure 5.1: LO Feynman graphs of dijet production.

p_T- and η-cuts for the two jets. We use only events with two leading jets for which $p_T > 30\,\mathrm{GeV}$ and $|\eta| < 2.5$.

In addition to the presence of two distinct jets it is implicitly required that there are no strong leptonic signals present. In an explicit definition this means that there are no **isolated leptons** in the event picture. That is, charged leptons which have an energy fraction greater than a (e.g. $a = 0.9$) of the total energy deposit within a radius R (e.g. $R = 0.3$) from the lepton. In simulations this requirement is secured by the choice of the hard process type.

5.2.2 γ+jet events

The γ+jet events are such that there is a strong jet signal and a strong photon signal present. An isolation condition could be constructed for the photon, but here we are contented by the selection of the hard process type. In γ+jet events the LO hard process has one QCD vertex and one QED vertex. The Feynman graphs for this process are presented in Fig. 5.2. This event type is much more specific than the QCD processes used for dijet events.

Quality control is done in a similar manner as for dijets, except that the photon takes the place of the other jet. In this case it is customary to define α using only

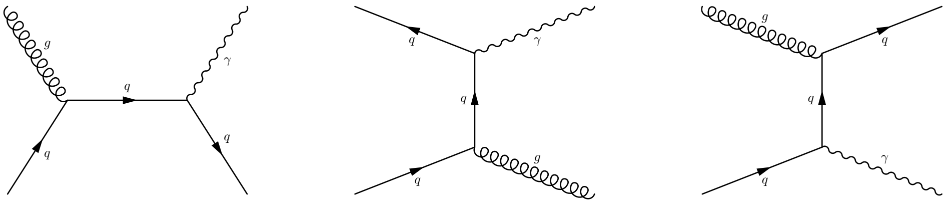


Figure 5.2: LO Feynman graphs of γ+jet production.

the photon and the possible next to leading jet transverse momentum:

$$\alpha = \frac{p_{T2}}{p_{T\gamma}}. \tag{5.4}$$

The back-to-back- and p_T- and η-cuts are implemented analogously to dijet events for the photon and the leading jet.

5.2.3 Zμμ+jet events

The Zμμ+jet events are otherwise similar to γ+jet events, but a Z boson is emitted instead of a photon. The Z boson is not observed at the detector, but instead its decay products can be directly observed if it decays into charged leptons. The Feynman graphs for this process are obtained by interchanging γ's in Fig. 5.2 with Z bosons and adding the decay to muons. The resulting graphs are given in Fig. 5.3.

In this work the studies are restricted to Z decaying into two muons. It would be possible to study also the Z-decay into two electrons, which is an important target of studies at the LHC. However, the small differences between Zμμ+jet and Zee+jet are not of interest here. On the contrary, the decay of Z to τ-leptons has special characteristics, as τ decays before reaching the detector.

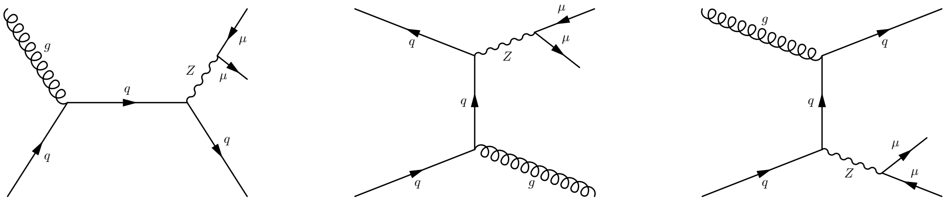


Figure 5.3: LO Feynman graphs of Zμμ+jet production.

At the CMS detector the muon signals are clear and distinguishable, owing to the large muon chambers. Using the two final state muons the original Z boson can be reconstructed. With the reconstructed Z most of the further event selection is similar to that in γ+jet events. However, it is useful to place limits for the Z mass, noting that $m_z = 91.19\,\mathrm{GeV}$ and the Breit-Wigner width is $\Gamma = 2.5\,\mathrm{GeV}$. We use a loose constraint 70 GeV $< m_Z <$ 110 GeV. The decay of virtual Z bosons could be mixed with the decay of virtual photons without this cut.

5.2.4 ttbar lepton+jet events

The ttbar ($\mathrm{t\bar{t}}$) lepton+jet processes are in some ways different to the ones listed above. The process type used here provides a good method for accessing the top quark mass. As the top mass is greater than the sum of b quark and W boson masses, the top quark decays most of the time through this channel. The weak decay occurs more quickly than any strong interactions and thus t is seen effectively as a free particle. This is in contrast to the other quark flavors, for which it is not possible to reconstruct mass values kinematically. For the top quark the case is the opposite. It is also possible to study events with a single top quark. However, the presence of two top quarks at a time provides better methods for mass calculations.

The basis of all the $\mathrm{t\bar{t}}$-processes are the basic QCD processes (see Fig. 5.1) with a $\mathrm{q\bar{q}}$ final state. The final state top quarks decay to b, $\mathrm{\bar{b}}$, $\mathrm{W^+}$ and $\mathrm{W^-}$. Thus at the detector level two b-flavored jets and the W boson decay products are observed. For the analysis purposes it is essential to tag correctly the b-flavored jets. This is further discussed in the next section.

The W bosons can decay either into a pair of quarks or into a charged lepton and its neutrino. In the $\mathrm{t\bar{t}}$ lepton+jet events exactly one decay of both of these types occurs. Thus there will be two additional light-flavored jets and a distinct charged lepton in the decay products. To get a complete picture of these events the top decay processes are presented in Fig. 5.4. The decay of the W bosons is left out, since the lepton/quark decays are interchangeable in these events.

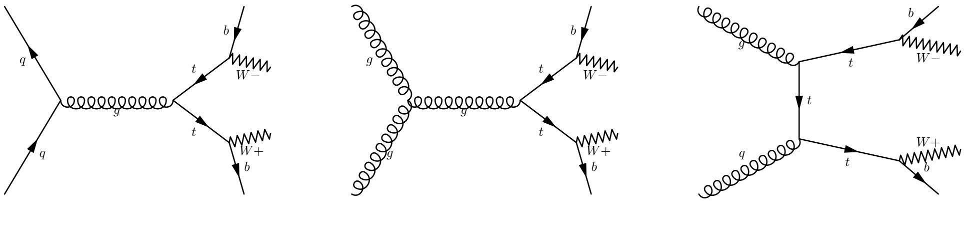


Figure 5.4: LO Feynman graphs of $\mathrm{t\bar{t}}$ jet production. In the lepton+jet case one W decays to leptons and one to quarks.

Practically all top-related processes are more or less complex. In the $\mathrm{t\bar{t}}$ lepton+jet events at the detector the other t quark can be reconstructed only approximately. Some of the initial top quark momentum is carried off by a neutrino, for which calculations are approximate. Only the MET (missing transverse energy) can be accessed, and it does not give a full truth of the neutrino momentum. This also explains why events with both W bosons decaying into leptons are not favored. On the other hand, events with only quark-decays have an inconveniently large number of final-state jets. The more jets, the more potential combinations have to be checked in a mass reconstruction.

The charged lepton serves as a good indicator for these events. An isolation condition can be applied for it. For electrons 90% of the energy within $\Delta R < 0.3$ must originate from the electron. For muons the corresponding conditions are 88% and $\Delta R < 0.4$. We require the presence of exactly one charged lepton. It can be checked from the simulation truth that this really is the lepton originating from the W boson. The general p_T- and $|\eta|$-requirements can also be applied to the four leading jets. This is even given that the light-flavored W boson jets are not always included in the four leading jets.

5.3 Flavor Definitions

This section introduces various algorithmical jet flavor definitions. First, an introduction to the existing definitions is given. Then, a closer listing is given of the flavor definitions used in this work.

5.3.1 Background

Simulations give good insight for flavor tagging. As the real measurements give only indirect information of the flavor structure, this is of a high importance. The hard process is the source of the hardest partons (LO partons) and thus it is desirable to concentrate on these partons in flavor tagging. However, some difficulties in using the full parton history are created by the event generators. For instance, the momentum changes caused by ISR and FSR are not always propagated back to the hard process.

To escape the bias in the momentum values within the partonic history, some algorithms use only final-state partons (FS partons) for flavor tagging. That is, the partons given just before hadronization. This causes other problems, as there can be multiple FS partons within a jet. The target of flavor tagging is to find the one original parton that produced a jet. The original parton is often found in the history of the FS partons.

Another problem with flavor tagging is the mechanism of connecting partons with jets. In principle this could be done using the total event history. However, it would be computationally intensive. It would also be difficult to implement such a study into the structures of the CMSSW. Therefore, the classic shortcut is matching the partonic momentum directions with those of the jets.

For jets with $R = 0.5$ a positive match has been given for a parton with $\Delta R < 0.3$ to the jet axis. Thus through momentum comparison also cases with excessive amounts of ISR and FSR bias are discarded. When used with the LO partons, these steps describe the **physics definition** of jet flavors. The ΔR-limit of 0.3 is mainly motivated by preventing partons from being matched with multiple jets. From the modern perspective, a ΔR-limit is not a very good way to perform this.

In a previous Master's thesis the robustness of the jet flavor definitions was studied [1]. This was done with Pythia 8 in the standard candle events, described in section 5.2. Here by **robustness** we mean that each jet flavor should have a characteristic physical behavior. Three different tagging algorithms were used: the **physics definition**, the **QCD-aware definition** and the **hadronic definition**. The QCD-aware algorithm is a modern attempt to find an improved flavor tagging method by the ATLAS collaboration. The hadronic definition is a modern flavor definition by the CMS collaboration.

It is necessary to remark that the flavor definitions can vary much according to their physical uses. In Ref. [1] it turned out that the old physics definition was still the best choice in terms of robustness of jets. These results were in part due to the fact that the hadronic jet definition concentrates on the purposes of **b-tagging**. That is, on finding jets containing a b-hadron. The QCD-aware definition is still under development, so it might get improved.

Efficient b-tagging is important, since a bottom hadron is often a good indicator of special events, e.g. the $\mathrm{t}\bar{\mathrm{t}}$ lepton+jet events. In the jet energy corrections (JEC) and the jet energy scale (JES), however, the distinction between quark and gluon jets is of the greatest importance. From the point of view of b-tagging, such a parton-level distinction is not that important. Therefore, we see that the purposes of JES require a distinct flavor definition.

The physics definition relies on the ideal situation where a jet corresponds clearly to one high-energy parton. If there is no match or conflicting matches for a single jet, the jet is labeled with no flavor. For instance at $\sqrt{s} = 8\,\mathrm{TeV}$ and $R = 0.5$ the no-flavor cases are almost always caused by a missing parton match.

The physics definition has been mostly in use with the FORTRAN-era event generators, especially with Pythia 6. The parton-jet correspondence can be disturbed by various factors. FSR gives a physical kick to the jet, disturbing it from the original parton direction. Additionally, the momentum values can be convoluted by various simulation techniques. The greatest shortcoming of the physics definition is that it relies on plain momentum values. The ΔR difference for pairing is somewhat artificial.

The goal of the flavor studies done in this work is extending the analysis of Ref. [1] in all ways. We begin the studies using the physics flavor definition. The enhanced analysis code allows a wide comparison between Pythia 6, Pythia 8 and Herwig++. Motivated by the results, experimental flavor definitions are developed and tried out. It becomes apparent that the physics definition is not anymore sufficient. Alternative definitions are described in the following subsection.

5.3.2 A listing of old and new definitions

An important point of study is the effect of the ΔR limit in flavor tagging. The effect of varying this limit is studied in this work. Furthermore, we are led to a new approach using **ghost partons**. This method is utilized in the hadronic definition. Flavor tagging can be performed using a jet clustering algorithm by appending the partons as **ghosts** to the particle list. The momentum of a ghost is multiplied by a small coefficient, e.g. 10^{-21}. Thus ghosts have no effect on the jet energy but they are still matched with the proper direction of momentum. A ghost parton approach escapes the artificial definition of a ΔR limit, as the partons are sorted uniquely into jets.

In this work we introduce the **LO flavor** definition. This upgrades the physics definition to the use of LO parton ghosts in the flavor tagging. Additionally, it discards any special handling of b quarks. The CMS community still uses the old physics definition so a distinct naming is favorable.

A further upgrade of the LO flavor definition is introduced as the momentum-corrected LO, the **CLO flavor** definition. It is motivated by the fact that the simulation process does not always update ISR and FSR effects to the hard process partons. This definition looks at the outgoing hard process particles and seeks their final parton descendants. That is, partons in the end of the partonic evolution stage (ISR+FSR+MPI). An artificial version of the outgoing hard process partons is constructed using the final-state partons. A total four-momentum is obtained for each LO parton by summing over their descendant partons. Thus we get effective corrections for the LO parton momenta. From here on the definition proceeds similarly as the LO flavor definition.

The final partonic state should have momentum values that are physically in agreement with the final state of the event. The greatest risks of the CLO definition is that the parent-child relations are not always handled properly by the generator. Thus the four-momentum summation can include or exclude some partons erroneously. A useful by-product of the CLO definition is a corrected four-momentum value for the hard process partons.

The **hadronic definition** utilizes FS partons preceding hadronization, and certain hadrons of interest as ghost particles. The hadrons of interest are such that they encompass a b or a c quark and are not in an excited state. Here with an excited state we mean a b- or a c-hadron that has a corresponding b- or c-hadron as its daughter. If a b-hadron ghost is clustered within a jet, the jet is labeled as a b jet. If there are no b-hadrons but a c-hadron is found, the jet is labeled as a c-jet. If there are no ghost hadrons within a jet the ghost partons are used. In case there are b- or c-partons within the jet these are given preference similarly as in the hadronic case. Otherwise a light flavor is given to the jet according to the hardest light ghost parton within the jet.

The hadronic definition gives an overly strong preference for hadrons from the jet physics point of view. Thus outside of b quark identification this definition is not suitable for the studies of the physical properties of jets. The **algorithmic**

definition is otherwise similar to the hadronic definition except that only FS partons are used, hadrons are excluded. In addition, this is an older definition and it uses a similar ΔR matching as the physics definition. The hadronic definition has mostly taken the place of the algorithmic definition.

The algorithmic definition is upgraded for the uses of this thesis similarly as the physics definition. We discard any preference of b quarks and do the flavor tagging using FS partons as ghosts. The new definition is called the final state flavor definition, i.e. the **FS flavor** definition. This definition is better for jet physics studies than the hadronic one, as it studies directly the parton level and does not prefer b quarks. However, the definition is still not free of problems.

In addition to the **LO**, **CLO** and **FS** definitions one additional new definition is presented. We call it the **historic definition**. This definition requires no additional particles or partons for matching purposes. Instead, for each final-state particle an additional status flag is given. It indicates the outgoing LO parton that is the ancestor of each particle. In practice the flag has to include information of both the flavor and the parton index, as there might be many particles of the same flavor present. In case no such ancestor is found or there is a conflict of ancestors, the particles are given no flavor. Finally, when the jets are sorted the jet constituents in the final state are looped over. A sum over E_T values is taken for each candidate flavor. The highest E_T sum determines the flavor of a jet.

The historic definition is motivated strongly by the partonic origin of jets. It propagates the original flavor directly into the final state of an event. If a jet corresponds to a single initial outgoing parton the hadrons within it should have a corresponding flavor history. This definition can be easily complemented by an E-purity requirement, labeling impure jets with no flavor. That is, it can be required that for a successful flavor tagging at least a certain percentage of the total jet energy must share the same origin. Thus for instance overlapping jets are conveniently excluded. The greatest risk within this definition is the hadronization process. The parent relations might not be physically rigorous when the partons are transformed into hadrons.

Finally, different definitions can be used in a hybrid mode. The LO flavor definition is often a desirable starting point for flavor tagging. However, it produces typically some amount of unmatched jets. One surprisingly promising method is a simultaneous combination of the LO and CLO flavor tagging. Another hybrid mode uses first LO tagging and then fills in the unmatched jets with the FS definition. The physical interpretations of each of these definitions require more consideration.

Thus, we have inspected old and new jet flavor definitions. A deeper motivation for the new definitions is given along with the results of this work. In Table 5.1 a brief summary of the relevant flavor definitions is given. All the flavor definitions used in this work but the physics definition have not been previously in a wide use to the knowledge of the author. However, no dramatic changes have been made, as most of the new definitions try to enhance old ideas. The old naming methods have been ambiguous, so we use descriptive names for the new definitions.

Table 5.1: Most important definitions with brief descriptions.

Definition name	Used here	Particles	Matching	b prioritized
Physics -	X	LO partons	ΔR	Optional
LO -	X	LO partons	Ghosts	No
CLO -	X	Corrected LO	Ghosts	No
Algorithmic -		FS partons	ΔR	Yes
FS -	X	FS partons	Ghosts	No
Hadronic -		Hadrons, FS partons	Ghosts	Yes
QCD aware -		FS partons	Algorithm	N/A
Historic -	X	Stable particles	History	No
LO & CLO -	X	Simultaneous LO & CLO	Ghosts	No
LO & FS -	X	First LO, then FS	Ghosts	No

5.4 Quark/Gluon likelihood

In the jet flavor studies of this work the distinction between quark and gluon jets is one main point of focus. This is the most important distinction in jet energy corrections. The differences between quark and gluon jets are the most elementary way to categorize jet flavors. Between the light flavored jets (u, d, s) the differences are generally small. However, the c and b jets differ to some degree of the other quark flavors. Gluon jets have on average wider cones than quark jets. The quark/gluon likelihood (QGL) parameters are designed to catch this and some other characteristics of the jets. A sophisticated analysis of the jet properties also allows the definition a quark/gluon likelihood. Only a probabilistic determination is possible, since the QGL parameters follow peaked distributions. The motivations for the QGL parameters are further discussed in Ref. [30].

The simplest QGL parameter is the number of detected particles in a jet. The discreteness of this variable motivates a careful consideration of what is counted as a particle. Since the detectors cannot see particles with small energies, a cut that discards particles with small p_T values is motivated in simulations. This makes the results better compatible with experimental observations. In order to have internally consistent results, the same cuts are applied also to the other two QGL variables. A larger particle multiplicity for gluon jets is theoretically motivated and it also agrees with the characteristic of wider cones.

A second QGL variable is directly related to the jet shape. This is the inner radius of the jet, σ_2. As stated before, the gluon jets tend to be wider and thus they have in average larger σ_2 values. The value of σ_2 can be found through the matrix representation of the jet cone in the $\eta - \phi$ -plane. The elements of the jet ellipse matrix M are found by summations over the jet constituents i:

$$M = \frac{1}{\sum_i p_{T,i}^2} \sum_i p_{T,i}^2 \begin{bmatrix} \Delta\eta_i^2 & -\Delta\eta_i\Delta\phi_i \\ -\Delta\eta_i\Delta\phi_i & \Delta\phi_i^2 \end{bmatrix}. \tag{5.5}$$

The differences of each jet constituent i are taken with respect to the jet axis. This matrix represents an ellipse, which favors jet constituents with high p_T values. The minor and major jet axes are found as the eigenvalues of the matrix M.

A third QGL parameter is the fragmentation function, defined by the the jet constituents i:

$$p_{TD} = \frac{\sqrt{\sum_i p_{T,i}^2}}{\sum_i p_{T,i}}. \tag{5.6}$$

This is also variable that characterizes the amount of particles and the p_T distribution of a jet. According to its name it gives an understanding of the fragmentation of the jet p_T between its constituents. For one particle, $p_{TD} = 1$. On the other hand, in the limit of infinitely many jet constituents p_{TD} goes to zero.

It is customary to calculate the QGL probabilities based on the distributions of the presented three variables. In this work instead of using the distributions for calculations they are studied qualitatively. Through a good graphical representation the comparison of the distributions is convenient. This introduces a method for studying the robustness of the flavor definition between samples and event generators.

5.5 Calculation of the top quark mass

The flavor studies lead us to the determination of the top quark mass. Jet flavor studies are an essential point of interest in the top quark events. Good b-tagging methods are essential for making a successful analysis. On the other hand in the $\mathrm{t\bar{t}}$ lepton+jet events there are two other significant quark jets present. Efficient flavor tagging makes it easier to find the correct quark jets. In general accurate jet energy scales are necessary for accurate top quark mass measurements.

In order to get a better understanding of the top quark events, this work is concluded with examples of top quark mass measurements. Here we use only a simple and intuitive approach for measuring the mass. Since we are using simulations, the measurement gives ideally the mass used by the simulation truth. Also more sophisticated methods for the measurement could be used, but they are left outside the scope of this work. For instance a **kinematic fit** uses various statistical properties to make the mass measurement more accurate.

The detector receives signals of two b jets, two quark jets and additional jets of arbitrary flavor. The official CMS b jet tagging algorithm is very efficient. While processing real detector data, the presence of two b jet signals is an essential indicator for the $\mathrm{t\bar{t}}$ events. Thus it is assumed that the flavors of the b jets are known. The other jet flavors are unknown in this simulated mass measurement. Also these flavors can be generally defined using the combined LO & FS flavor definition.

In the measurement of the top mass, it is essential that the jets correspond accurately to certain partons. In practice this requires some luck in the jet clustering phase. The clustering process is mechanical and it will give no direct indications of success or failure in terms of the partons. In the simulations it is motivated to discard an event directly if the flavor tagging fails. Usually this is an indicator of a problem that would manifest in a latter part of the analysis.

The analysis of the leptonic W boson involves less data than that of the quark W boson. The only data available are the four-momentum of the lepton and the MET. The charged lepton is required to be isolated from other particles indicating its origin outside the jets. It is not possible to determine a z-component for the missing energy, as the uncertainties in the beam direction momentum are too large. On the other hand even the missing E_T vector has large uncertainties, as it is only found from the sum of all transverse momenta. The missing transverse energy is only a 2D momentum value, which is missing a z-component. Ideally this would include only one neutrino, but in reality also background neutrinos can be included. For instance if the W boson decays initially into a τ lepton, the missing transverse momentum includes the effect of multiple neutrinos.

It is possible to obtain the missing z-component by forcing the momentum sum of MET and the charged lepton to the W mass shell. This approach has several possible sources of errors. One of these are the errors between the ideal neutrino transverse component and MET. Another is that of forcing the sum directly onto the mass-shell, as the Breit-Wigner peak width of W is notable (1.5 GeV). An even more problematic source of errors is the determination of the z-component direction. By forcing the mass-shell condition there are two possibilities for the z-component. It can be forced to the one closer to the z-component of the charged lepton, but some erroneous cases are left in the results.

The obtained leptonic W four momentum can be paired up with the two b jets. The proximity to the top quark mass indicates which one of the b jets corresponds to the leptonic W. A reference top quark mass can be fetched from the t mass value that is obtained from the light-flavored jets. It is possible that no match is made or that both the b jets match with the same degree of accuracy. It depends much on the accuracy of the measurements, how complicated the evaluation becomes.

The simplest way to study the quark-W would be to make calculations for the two leading p_T jets, excluding the b jets. However, this leads easily to inconsistency. Especially with the higher and higher collision energies there will often be additional high p_T jets present. A better approach loops over the promising high-p_T (here $p_T > 30\,\text{GeV}$) jets. For these jets only a strict range of W masses is allowed. The surviving W candidates can be paired with the b jets. These are processed by pairing up the t mass with the leptonic one. This procedure can be computationally intensive due to the many possible permutations.

Chapter 6

Results

The previous sections showed that the implementation of a sufficient infrastructure for event generation requires much work. In principle all this has been done already in CMSSW. However, CMSSW is not optimal for running flexible and experimental event generation tests. Thus it was decided to rely on a standalone build of all the necessary packages. Aside of requiring much work, such an implementation gives great insight for the total generation process. The software written for these purposes is more closely described in Appendix B. Conducting the simulation stages (Ch. 4) and the needs of the basic analysis (Ch. 5) are performed by the software.

This chapter starts by a examination of the robustness properties between the three GPMCs. From the results of these considerations some problems with the physics definition are observed. The study continues by inspecting multiple experimental jet flavor definitions. After a thorough study, some useful modifications to the flavor definition are fixed. Finally, we move on to studies of the top quark generation processes. A good look is given for the jet properties in these events. Finally, we perform some elementary mass measurements for the top quark.

6.1 Flavor comparison between event generators

In the previous work of Ref. [1] the physics flavor definition turned out to be the favorable one of the three studied definitions. Thus this section studies this definition carefully. As the studies progress, some weaknesses of the physics definition are observed. In this section the CMS run 1 parameters are used: $\sqrt{s} = 8$ TeV and $R = 0.5$ for the jet radius.

6.1.1 Robustness of the physics definition

The previous flavor study of Ref. [1] considered the robustness of the flavor definitions with only Pythia 8. This study was done within the three standard candle event types that are used also in this work. The next logical and necessary step is to extend the study to multiple event generators. Because of the nature of the errors present in simulations this kind of a study is essential. Only an agreement between multiple event generators suggests some level of universality of the results.

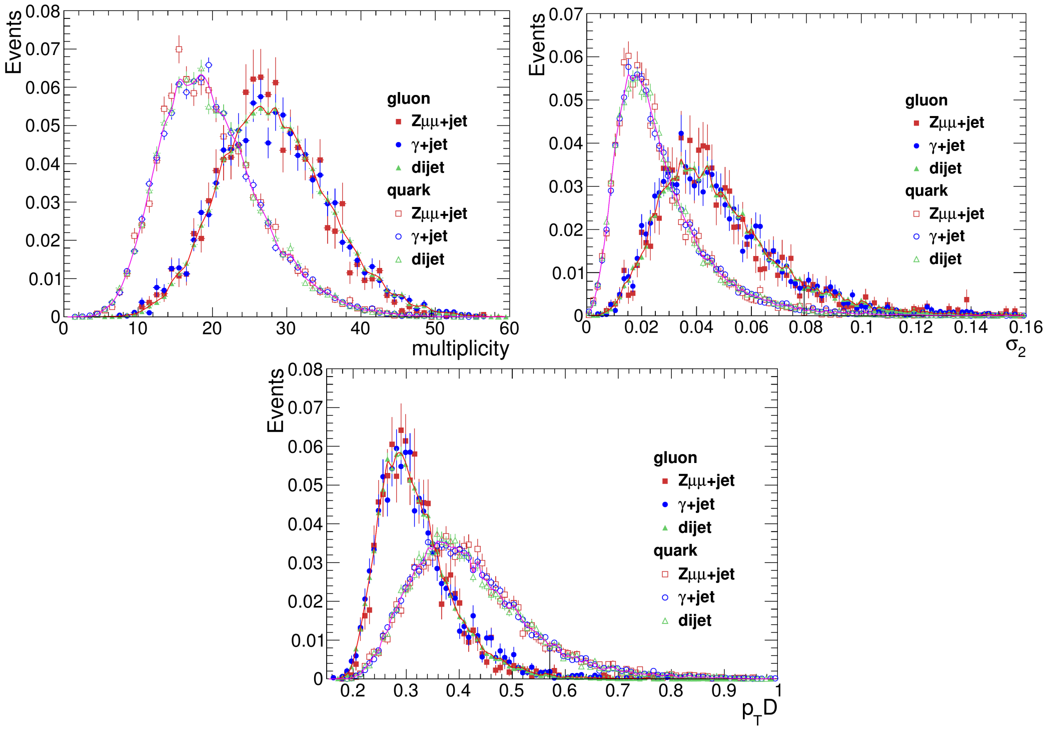


Figure 6.1: A reproduction of the Pythia 8 physics definition robustness results using the standard candle events. The amount of events per bin is scaled by the integral of the histogram.

A logical starting point for the study is the reproduction of the Pythia 8 results. As the same analysis tools are used for all of the event types it is crucial to be able to be able to end up with similar results as the previous study. In Fig. 6.1 the QGL parameters for Pythia 8 are given. The sample size is 10^6 events and a strict p_T cut has been applied. Only a jet p_T range $p_T \in [80, 100]$ GeV is shown, since the QGL behavior is highly dependent on the p_T values. It is seen that the quark and gluon jets are nicely separated, when it comes to the distributions of the QGL parameters. This is exactly what was expected from the physics flavor definition, allowing us to proceed with the studies.

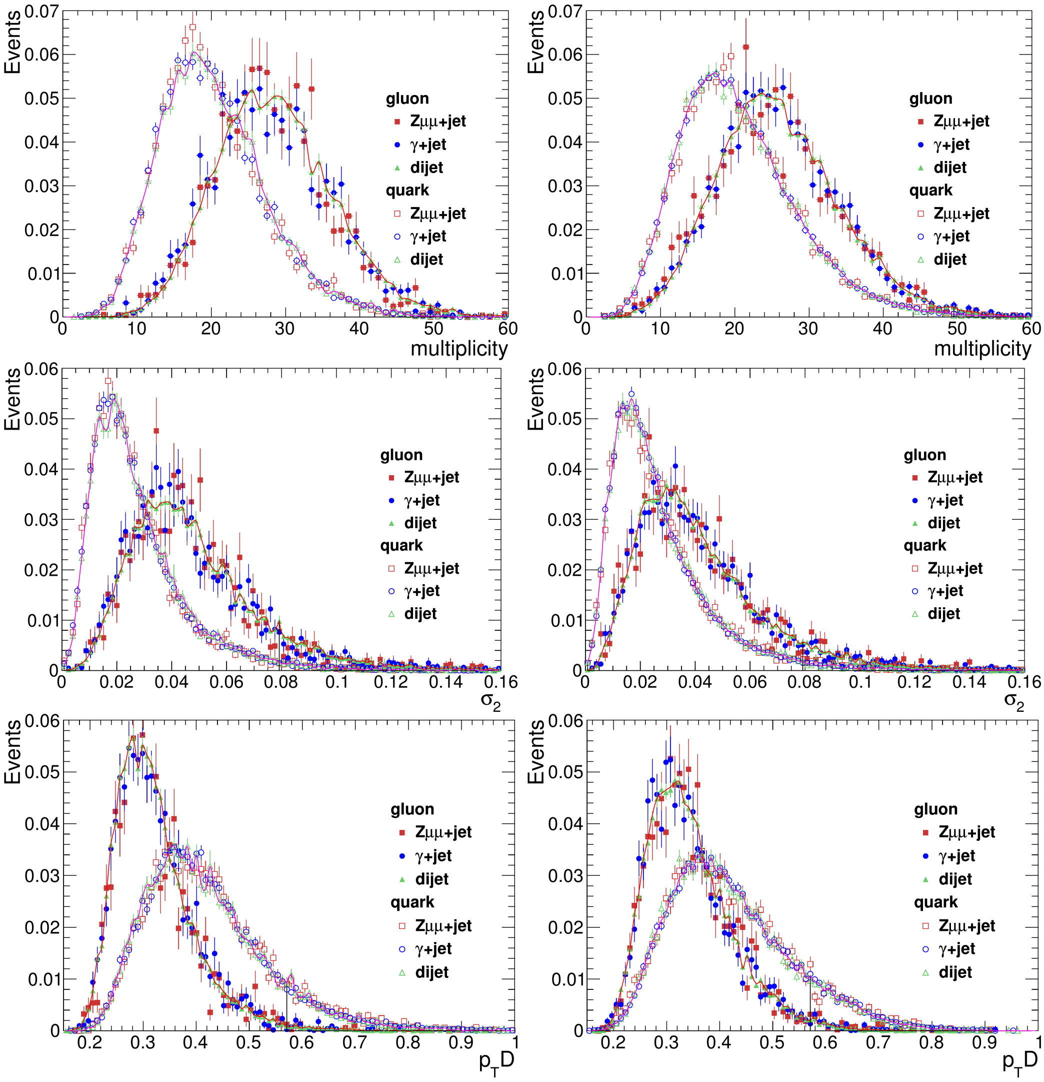


Figure 6.2: Pythia 6 (left column) and Herwig++ (right column) physics definition robustness results using the standard candle events. The amount of events per bin is scaled by the integral of the histogram.

Furthermore, the QGL parameters are given for Pythia 6 and Herwig++ in Fig. 6.2. Qualitatively the agreement seems to be good with the Pythia 8 results. In Fig. 6.3 the results from the standard candle events have been combined into a single plot. This shows that Pythia's are in a good agreement, but Herwig++ disagrees slightly. The disagreement concerns gluon jets, as the agreement between quark jets is phenomenally good. It is important to carry in mind that even the agreement of the two Pythia versions is not trivial, as the software has been completely rewritten.

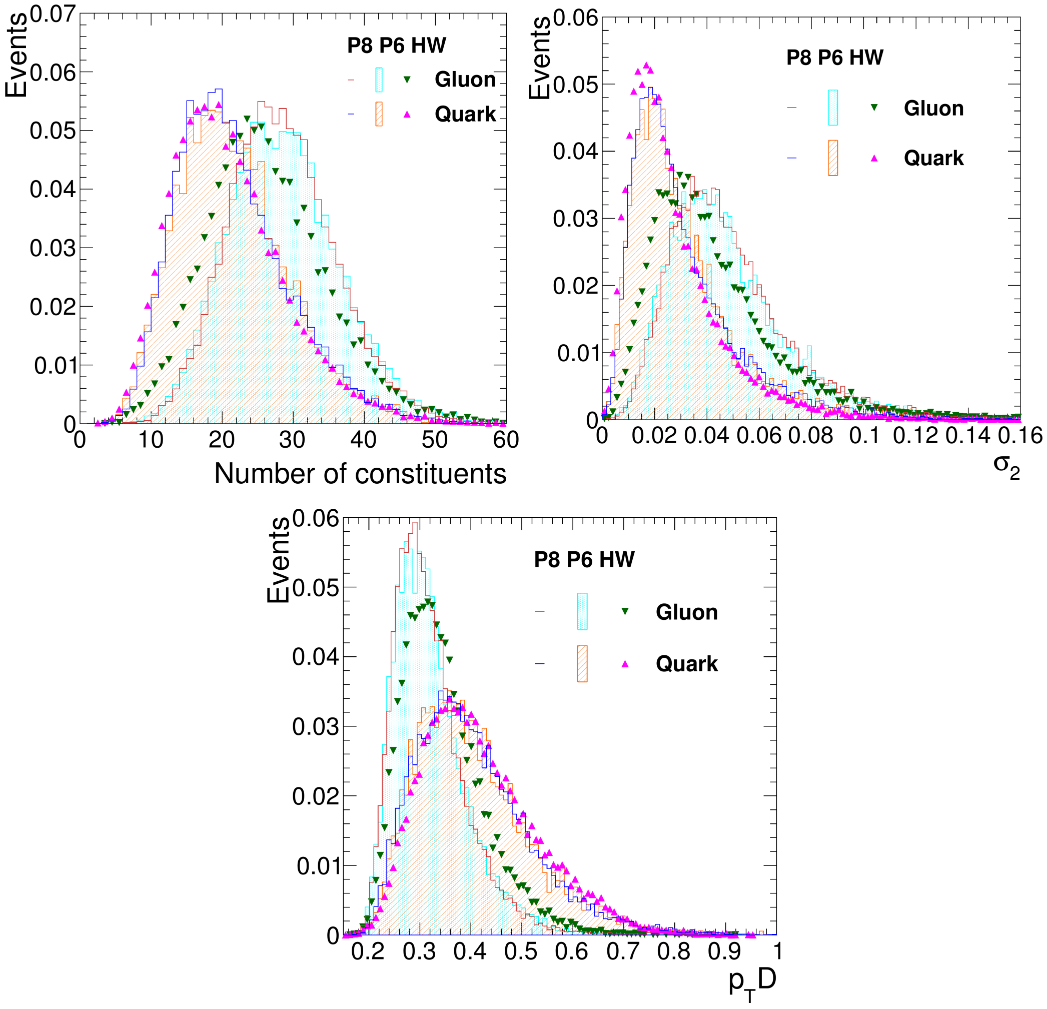


Figure 6.3: A summary of the robustness results using the standard candle events. P6: Pythia 6 , P8: Pythia 8 , HW: Herwig++ . The amount of events per bin is scaled by the integral of the histogram.

6.1.2 Flavor fractions

In a generic comparison of the results, it is useful to study the distribution of jet flavors as a function of p_T and η. This is a basic indicator of differences in the logic of event generators. All differences are possible indicators of physical disparities.

In the uppermost row of Fig. 6.4 the flavor fractions are given for all three GPMCs. The distribution of flavors is somewhat similar for all of them. However, in Pythia 8 there is a large amount of jets with no flavor. For the physics definition this means that no initial parton has been found within $\Delta R < 0.3$ from the jet axis. This is a first indication of the incompleteness of the physics definition.

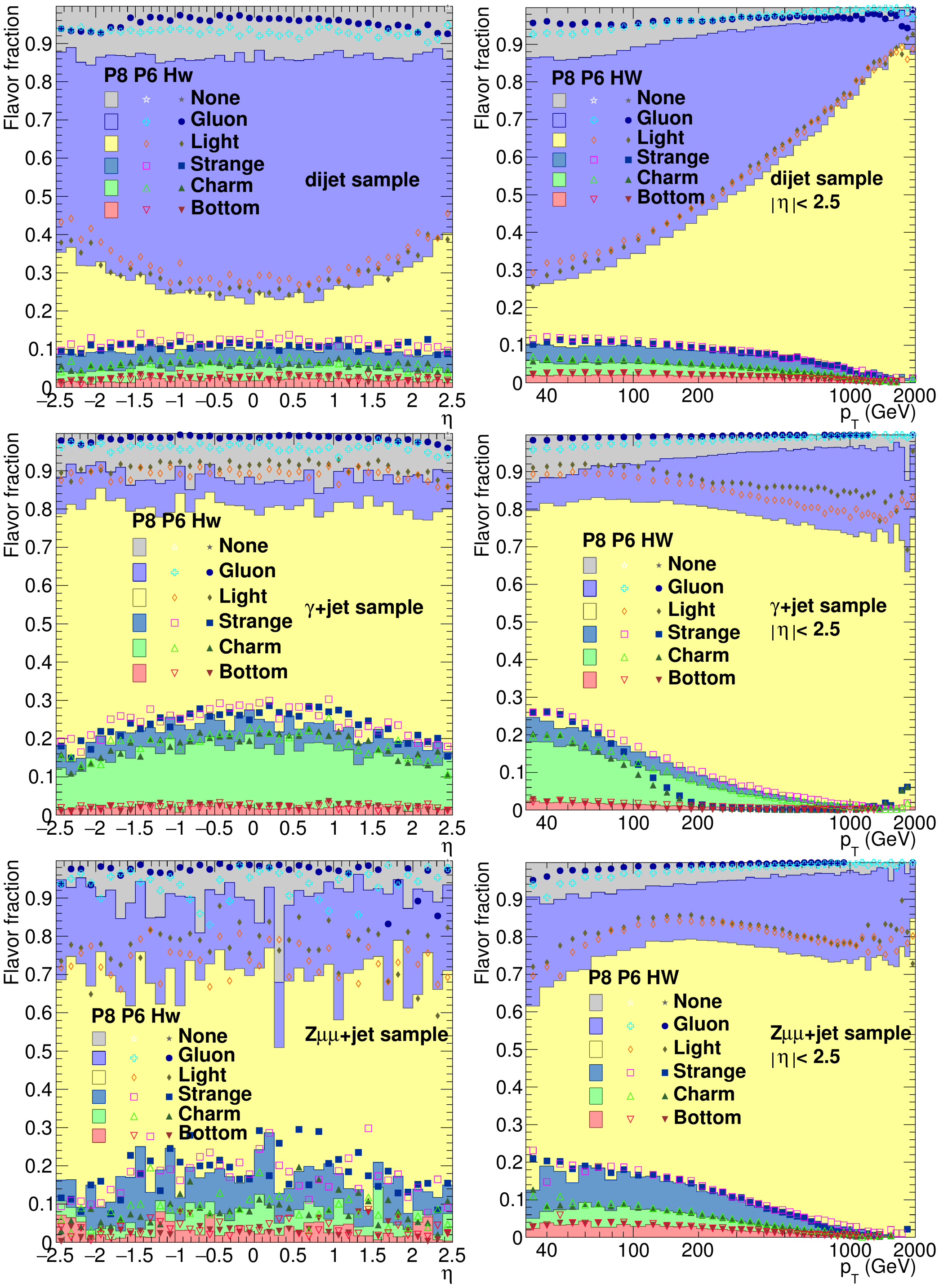


Figure 6.4: Flavor fractions in dijet (first row), γ+jet (second row) and Zμμ+jet (third row) events.

The results in γ+jet and Zμμ+jet events are given in the following rows of Fig. 6.4. There are several aspects that deserve a mention. We observe again a greater rejection rate for Pythia 8 than for the other generators. In the γ+jet events there is also a great difference in the p_T distribution of the jet flavors. This is an unexpected property, as the η distributions agree well. In the Zμμ+jet events the p_T distributions have similar behavior, but the η-distribution is very noisy. This is a direct indication of the failure of the $\hat{p}_T$ reweighting scheme. If there is a large variety of $\hat{p}_T$ values within one η bin, the smaller p_T values dominate the results.

6.1.3 The hard process

The hardest subprocess is a most elementary source for comparison of the results. For the physics definition there is a direct correspondence between the flavors in the hardest process and the final jet flavors. In dijet events there are two outgoing partons in the hardest subprocess. In γ+jet and Zμμ+jet events there is one outgoing parton in addition to the photon or Z boson. Using the same PDF with different generators, there should be statistically no differences in the hard process.

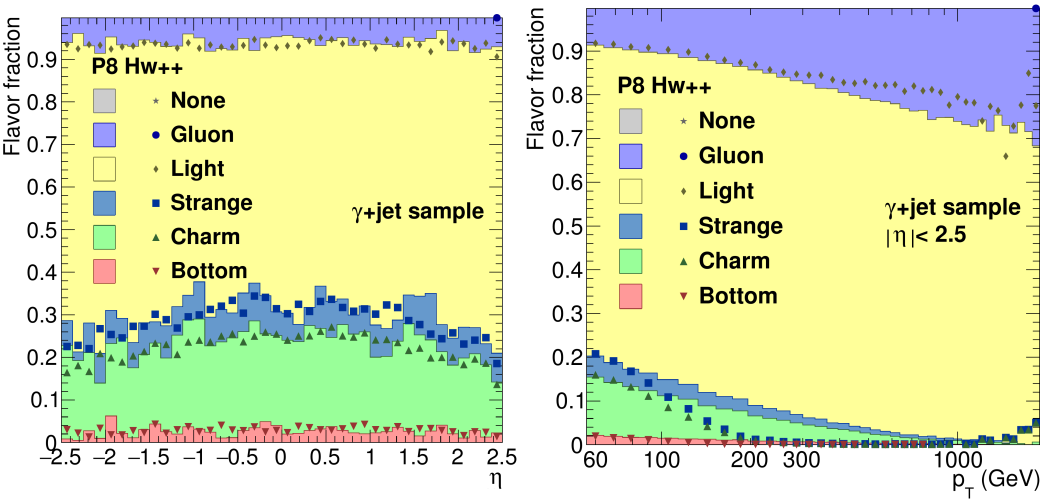


Figure 6.5: Outgoing hard process flavour fractions in γ+jet events for Pythia 8 and Herwig++.

It turns out that in the dijet events and Zμμ+jet events studying the hard process does not reveal much. However, in the γ+jet events an important observation can be made, shown in Fig. 6.5. The differing p_T-behavior of the fractions is visible already here. Thus there is a distinct physical difference between the generators in this case. This means that the γ+jet events need to be handled extremely carefully, when it comes to the jet flavor. Normally this kind of differences could be explained by the usage of different PDFs. However, in these simulations the same PDFs have been used for all generators.

6.1.4 Rejection rates in the physics definition

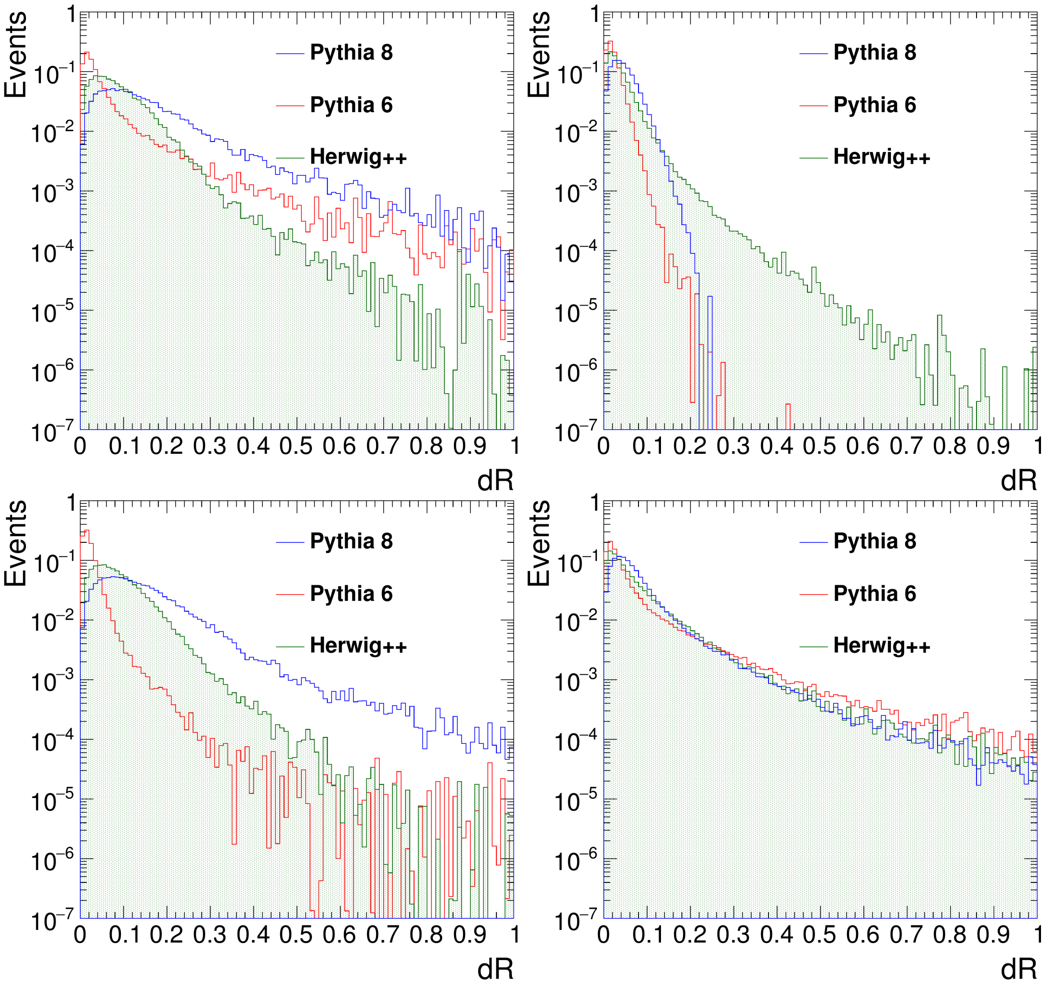


Figure 6.6: Radial distances ΔR ($= dR$ in the figures) between the jet axis and the closest outgoing hard process parton. Upper left: ISR and FSR on, upper right: (ISR and FSR off). Lower left: ISR on (FSR off), right: FSR on (ISR off). The amount of events per bin is scaled by the integral of the histogram.

The high rejection rates in the PYTHIA 8 flavor fractions inspires a thorough study of the rejection phenomenon. In Fig. 6.6 upper left corner the radial distance ΔR between the jet axis and the closest outgoing LO parton is presented. These results are given for the two leading p_T order jets in dijet events. It is seen that the modern GPMCs Pythia 8 and Herwig++ have a spread out profile w.r.t. Pythia 6. At the somewhat arbitrary pairing limit of $R = 0.3$ Pythia 6 and Herwig++ results come close to each other. On the contrary, Pythia 8 LO partons have a systematically larger distance to the jet axis. This gives motivation to find better ways for defining the jet flavor.

It is obvious that Pythia 8 has for one reason or another more differences between the hard process and the final state. The currently used simulation software allows the user to turn off ISR or FSR (or MPI) in the simulations. Thus the effective cause of the spreading out of Pythia 8 results can be studied more closely. In Fig. 6.6 on the upper right corner both ISR and FSR have been turned off. It is seen that now both Pythia versions give a sharp peak, as Herwig has a broader distribution. This encourages us to look for the cause for rejections in ISR and FSR.

Finally, in the Fig. 6.6 lower row the same plot is presented with only FSR or ISR off. This gives a strong suggestion that the main cause for Pythia 8 rejections is ISR. In the Pythia 8 manual [17] it is stated that the handling of FSR and ISR has changed since Pythia 6. The momenta values within the hard process are no longer corrected after applying FSR and ISR. This is due to changes in the logic according to which FSR and ISR are applied. For the timelike FSR this does not do much harm, as there is no causal effect to the hard process. This and the similarity of the plots with FSR on (ISR off) suggest that the FSR curves are physically natural, indicating a FSR-recoil.

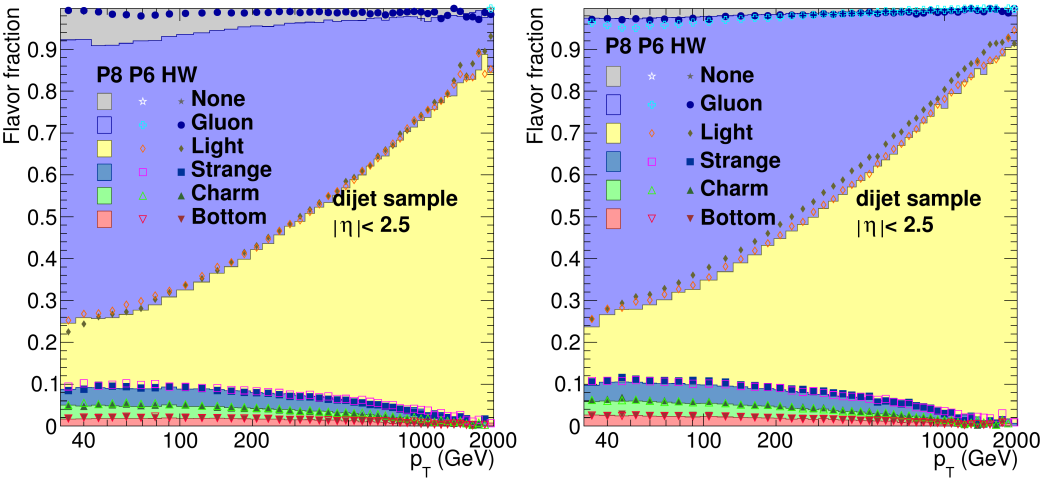


Figure 6.7: Flavour fractions - left: ISR on (FSR off) and FSR on (ISR off).

On the other hand ISR has a causal effect on the hard process. Not correcting the hard process momentum values results in a great bias due to the simulation structure. Because of this the strategies successfully used in Pythia 6 in the CMS project are no longer as viable in Pythia 8. In Fig. 6.7 the flavor fractions are shown with FSR and ISR turned off to enforce the argument.

6.2 Flavor definition studies

In this section the studies are done exclusively with Pythia 8. The motivation for testing new variants of flavor definitions comes from the undesirably large failure rates in flavor tagging. By a good margin the largest rejection rate is observed with Pythia 8. It is thus initially best to concentrate on finding new definitions that improve the Pythia 8 results the most. There is no reason to implement experimental definitions for the other GPMCs without promising Pythia 8 results.

We proceed by studying only dijet events. It is the most generic event type available and thus ideal for jet studies. At this point it is a reasonable assumption that the behavior is relatively similar in the other standard candle events. In these simulations the condition of $\alpha < 0.3$ is used. The α definition uses only the relative energies of the two or three leading jets. Thus also such events are observed in which there is a multitude of small-energy jets present.

Here, we will put into action all the new definitions presented in Section 5.3.2. The progression is done logically so that first we study the simple definitions and then the hybrid definitions. The central results of this section are given in Figs. 6.8-6.13. Figs. 6.8 and 6.10 show the simplest enhancements for the physics definition: the LO and the CLO definition. Both of these have upgraded the physics definition with ghost partons and the CLO definition uses in addition a corrected partonic momentum value.

In Fig. 6.9 the historic flavor definition is presented. This is distinctively different from all the other definitions in use. Furthermore, in Fig. 6.12 the FS definition results are given. Neither the Historic nor the FS definition turn out to be promising, but they give hints for further developments of the definitions.

Finally, in Figs. 6.11 and 6.13 the two hybrid definition results are shown. They combine the LO definition with the CLO definition and the FS definition. These are the most promising two of the flavor definitions studied. In the following, we will go through these results more closely.

6.2.1 New flavor definitions

The simplest method to enhance the physics definition is the use of ghost partons. This does not change the basic principles of the physics definition in any way. However, the arbitrary choice of $\Delta R < 0.3$ is removed. This gives smaller accumulation of unlabeled jets for Pythia 8. The change is visualized in Fig. 6.8, which is a recreation of the first p_T plot in Fig. 6.4. This new definition is called simply the LO definition.

Flavor fractions using new flavor definitions.

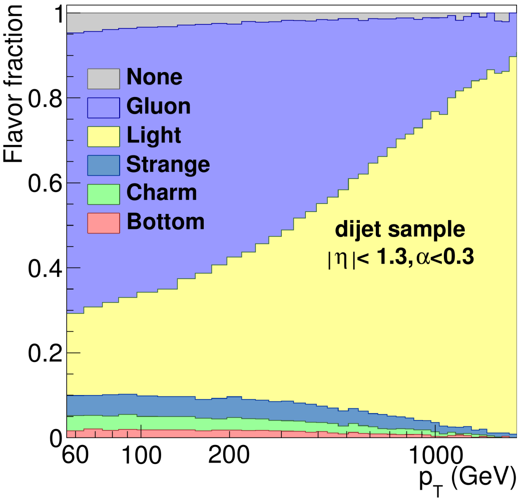


Figure 6.8: LO.

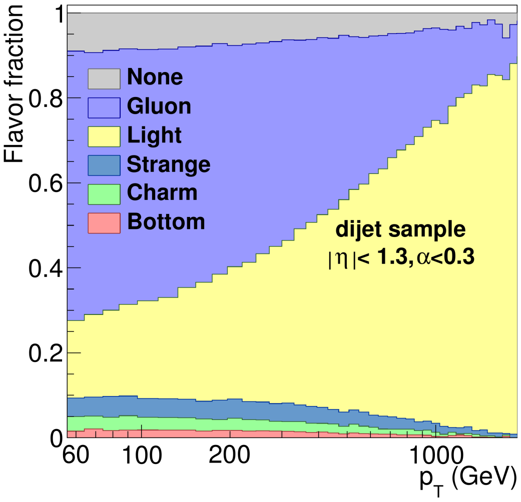


Figure 6.9: Historic.

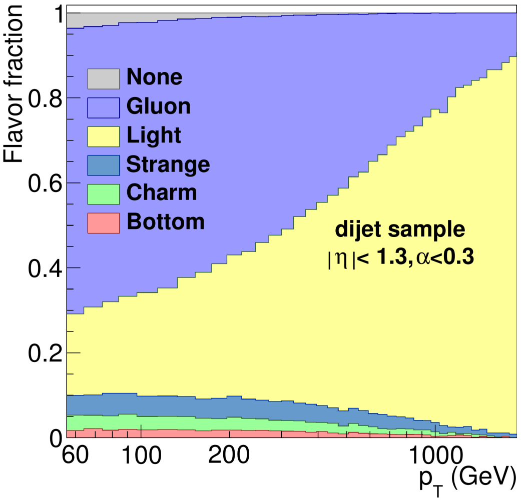


Figure 6.10: CLO.

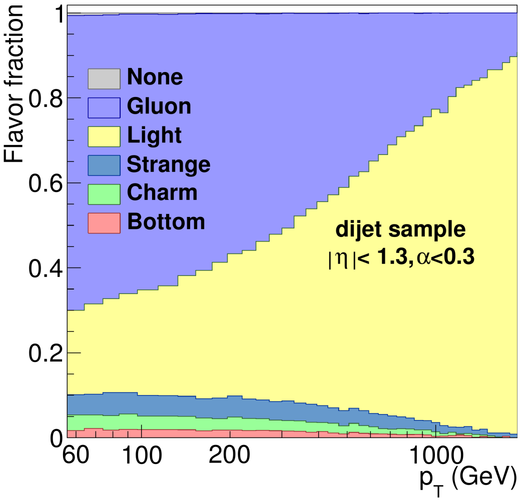


Figure 6.11: LO & CLO hybrid.

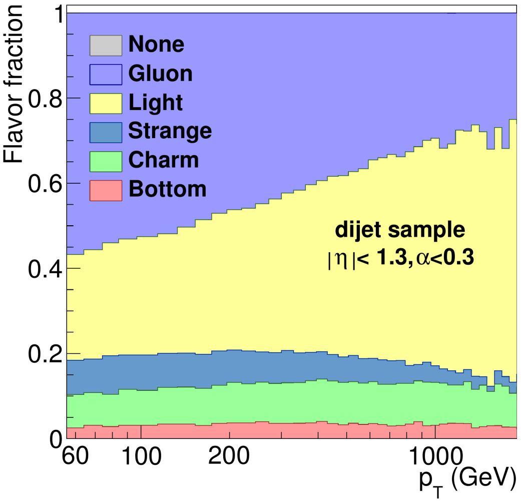


Figure 6.12: FS.

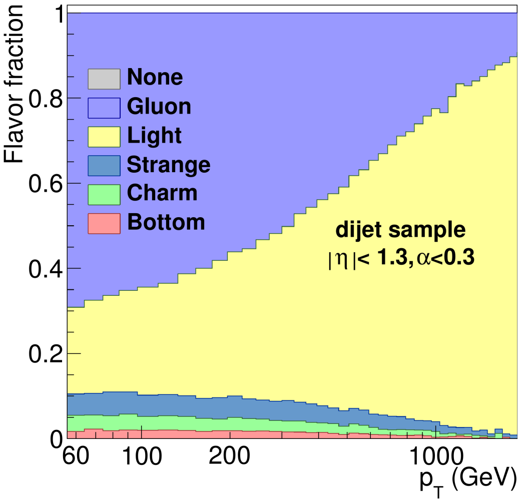


Figure 6.13: LO & FS hybrid.

It should be observed that the LO definition has specific characteristics. The results for $R = 0.5$ jets are mostly similar to physics definition results with $\Delta R < 0.5$. However, due to the distribution of particles and energies in the jet, LO flavor results can be slightly different. Partons from $\Delta R > 0.5$ can land in the flavor definition and partons from $\Delta R < 0.5$ can be excluded from it.

The momentum-corrected LO (CLO) definition tries to take into account the nonphysical ISR-related momentum changes occurring at the parton level. The results are presented in Fig. 6.10. We see that the final result is quite similar to that of the LO definition. However, the results seem to be slightly better than with the LO definition. A further study is needed to show whether the *none*-events are the same as with the LO definition.

The historic definition relies on the partonic history of the constituent particles of jets. A purity condition is given so that at least 70 percent of the jet energy should originate from one parton. The exact value of this condition is somewhat arbitrary, but ideally a very high purity (90 %) is expected from jets. The resulting flavor fractions are given in Fig. 6.9. We observe that the amount of cases with no flavor labeling is very high. The structure of the *none* composition is, however, different to that of Fig. 6.4. Thus the logic of flavor rejection is here somewhat different than with the $\Delta R < 0.3$ condition.

For us to understand this phenomenon a closer look needs to be taken at the event level. For the composition of jets it is ideal to study the directions of particles in the (y, ϕ)-plane. The jets are shown as circles in this plane, with a color coding that implies the jet energy. The two leading (highest energy) jets are given a dark red and red coloring. Consequently, the two following jets are given a magenta and a dark cyan color. All the following jets are drawn as blue circles.

In Fig. 6.14 the final state particles originating from hard process partons are given together with jets and the hard process partons. This is actually the event 0 from the sample handled in the Appendices C and D. These Appendices are introduced more closely in the next subsection. The marker sizes are proportional to the the particle energies. The outgoing hard process gluons are plotted in green, with markers two times as big as normally. The cross symbol indicates the corrected hard process parton momentum. Final state particles are plotted as red squares and blue circles. The former are particles with a historic flavor and the latter particles without a historic flavor. Furthermore, the plot is given in the range $[-\pi, 3\pi/2]$, in order to not lose any information at the continuous ϕ boundary. This plotting format is developed further in Appendix D, to provide a good visualization of the events.

From the given figure we observe that the final state particles originating from the hard process are convoluted all-around the (y, ϕ)-plane. This is very likely consequential to the hadronization process and the logic of parent-child relationships in Pythia 8. The ISR and FSR partons may be marked as the descendants of hard process partons. Furthermore, in the hadronization process the relations can get confused and a multitude of particles ends up as the hard process descendants.

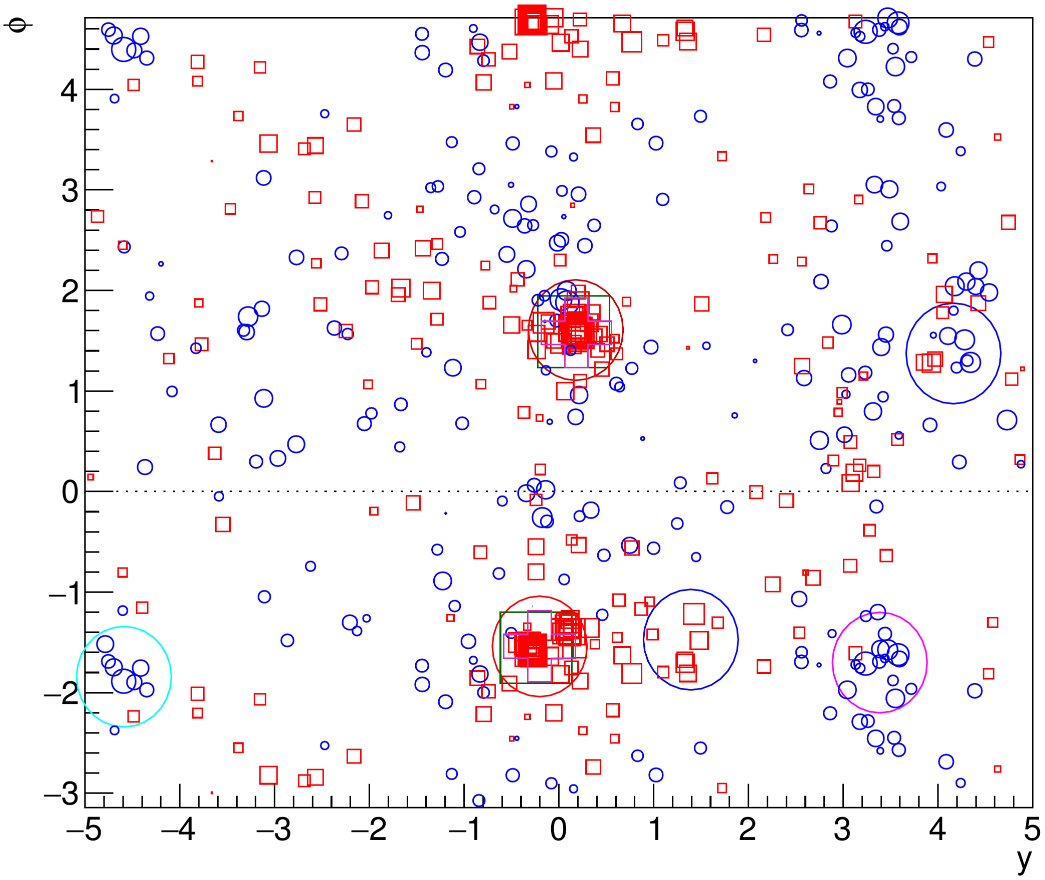


Figure 6.14: The final state particles with a historic flavor: red squares. Other final state particles: blue circles.

Thus this kind of a flavor definition is strongly discouraged. In principle the purity requirements can be lowered, to attain a greater rate of accepted jets. However, this would make the whole process of flavor tagging somewhat arbitrary. According to the event history the descendants of one parton can end up to multiple jets. Forcing a flavor tag according to the history to a single jet is thus not a robust process.

Finally, in Fig. 6.12 the FS (final-state partons) flavor definition results are shown. The FS definition is essentially similar to the hadronic definition, with hadrons and special b-tagging features stripped away. Nevertheless, the plot is still similar to that obtained with the hadronic definition in Ref. [1]. The difference to the LO and the CLO definitions implies an anticipated problem. The FS flavor definition does not observe gluon splitting in which the produced quarks end up within the same jet.

6.2.2 Hybrid flavor definitions

The historic definition did not turn out to be promising. Nevertheless, the LO and CLO definitions gave encouraging results. The core question is whether the events labeled *none* are actually the same in these events. In appendix C a listing of 600 dijet events is studied with these two definitions and the following two hybrid definitions. The listing shows the events in which some of the definitions fails.

It turns out that the overlap of the *none*-cases is surprisingly small in LO and CLO. This gives a motivation to create a hybrid of these definitions. Since both the flavor definitions utilize ghost clustering, these cases can be handled simultaneously. A jet is given a parton flavor if either or both of the original or corrected momentum partons are found within it. If there is no match or a conflict between different original partons a *none*-flavor is given.

The flavor fractions obtained with this hybrid method are presented in Fig. 6.11. It is seen that the *none*-fraction is almost reduced to zero. Most of the unlabeled jets have been turned into gluon jets. Thus we are encouraged to believe that a hybrid flavor definition is the way to proceed. Using a more straightforward definition would have been preferable, but not attainable.

The FS flavor definition used separately did not have the desired properties. However, it can be easily combined with the LO definition. By using the LO definition first and then the FS definition, we can catch most of the gluon-splitting events. On the other hand, the FS definition is excellent for finding the flavor of low-p_T jets. The flavor fractions with the LO & FS hybrid definition are demonstrated in Fig. 6.13. The results are of a high quality and in a good agreement with the other hybrid definition.

Using the LO and the FS definition together is also motivated by the CMSSW structures. Currently the LO partons and the FS partons are provided for analyses purposes. Extensive history information is missing and thus calculating for instance the corrected momentum can be challenging.

To give a better understanding of the problems occurring with each flavor definition a display of problematic events is given in Appendix D. A check of the robustness results would be required in order to achieve a good confidence with the new definitions. However, since the amount of *none*-flavored jets is not great to begin with, the results would not change essentially. The solid basis of the new hybrid definitions lies in the LO definition. This produces similar results as the old physics definition, for which the robustness was already confirmed. Thus we can here omit a further study of the subject by referring to the previous results. The hybrid definitions are the ones with the best performance and each of them has some positive features. In the rest of this work we will use the LO & FS hybrid definition, as this can define also the flavor of all jets in an event.

6.3 Top quark studies

To keep the scope of this work in reasonable limits, only Pythia 8 $t\bar{t}$ events are studied in this section. The simulation machinery has been made ready for studies also with the other two event generators. The significance of using these will become larger when more specific calculations are done. We concentrate on the elementary properties of these events, allowing a more in-depth study in the future. The Pythia 8 sample used here consists of $5 \cdot 10^6$ events, of which many are cut away by quality requirements. For comparison purposes a sample of $5 \cdot 10^6$ dijet events is used.

6.3.1 Jet properties in the ttbar lepton+jet events

Encouraged by the results so far, we begin the $t\bar{t}$ studies by looking at the jet properties. The only way to have a thorough picture of the flavors is to use the LO & FS hybrid flavor definition. This is because there should typically be at least four jets in these kind of events: 2 b jets and 2 light quark jets. On top of these there is a plethora of other jets present. The LO & CLO definition can assign a flavor only for the LO parton jets. Without the help of the FS definition we would lose all information of gluon jets. In Fig. 6.15 the p_T and η profiles are given. The only general cuts applied here is the requirement that there are at least four jets and for the four leading jets $|\eta| < 2.5$ and $p_T > 30\,\text{GeV}$. Additionally, the presence of a single isolated charged lepton is required.

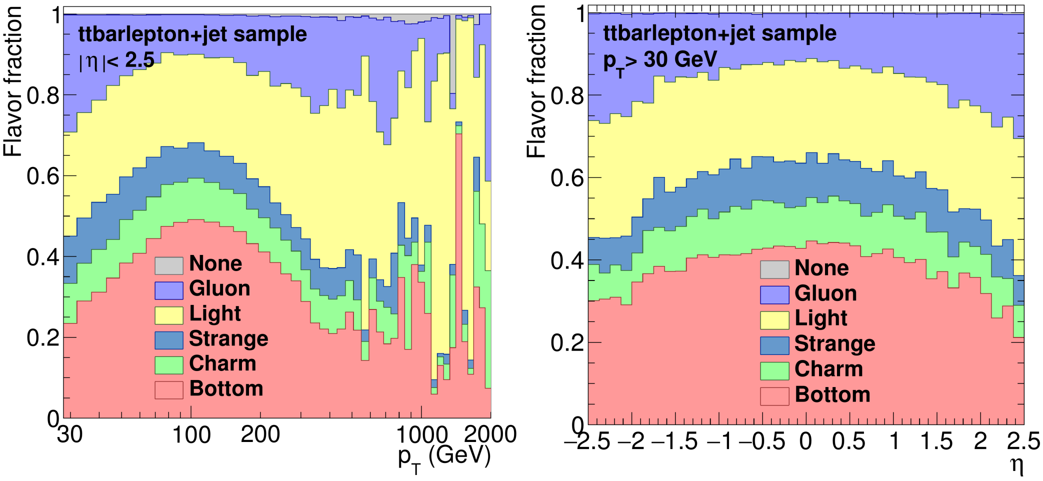


Figure 6.15: Flavour fractions of all jets in ttbar lepton+jet events.

From the given figures some important observations can be made. First of all, we still have the same $\sqrt{s}$ as before, but the events consist on average of more jets.

Because of this, the jets are less energetic. In fact, the statistics for jets already with $p_T > 400\,\mathrm{GeV}$ are rather poor.

We can observe a maximum of the b jet fraction around 100 GeV. This makes sense, as the b jets originate together with a 80 GeV W boson from the 172 GeV t quark. When compared to the standard candle events there is a large abundance of b jets, as one expects for the $\mathrm{t\bar{t}}$-events.

The next thing to study are the robustness results. From all jets it is required $80\,\mathrm{GeV} < p_T < 100\,\mathrm{GeV}$. The small p_T range is motivated by the fact that the plots are highly p_T dependent. In Fig. 6.16 the aforementioned robustness plots are given. Each plot has been smoothened and normalized to have a maximum of 1. As an additional feature of these plots it has to be mentioned that also in the dijet-samples the LO & FS flavor definition has been used. These dijet results correspond very well to those that use only the two leading jets and the LO definition.

It is noteworthy that here some additional cuts are brought for the jets in contrast to the previous robustness plots. We require $p_T > 1\,\mathrm{GeV}$ for photons, $p_T > 3\,\mathrm{GeV}$ for neutral hadrons and $p_T > 0.3\,\mathrm{GeV}$ for charged hadrons. This procedure removes a large amount of the small particle shreds, abundant especially in $\mathrm{t\bar{t}}$ events. In practice this does very little for the σ_2 and p_{TD}, but the number of constituents is almost halved. Even after this procedure the top-events are left with slightly more constituents.

The Fig. 6.16 includes two different versions of robustness plots. One has the minimal amount of cuts used in the flavor plots. The other, the stricter one, requires the event to have good W boson masses and other such signatures of credible $\mathrm{t\bar{t}}$-events. It seems that the results behave a little more smoothly in the less strictly chosen case. No great differences are observable.

As already mentioned, the constituent plots are still slightly deformed in the $\mathrm{t\bar{t}}$ events. These are stretched to the right from the main peak. However, the general behavior in different flavors is such that one could expect. A special note has to be put on the b jets, which are in-between (light) quark and gluon jets.

The p_{TD} values are in general an incredibly good match. In the $\mathrm{t\bar{t}}$-events there are slightly more fluctuations, but this is only natural. There are more jets in the $\mathrm{t\bar{t}}$ events and thus the risk of poor quality jets is naturally higher. Jets that are located side-by-side are a main reason for this.

In the σ_2 plots the agreement is similarly quite good, but there are even more fluctuations. It seems that in the high inner-radius zone there can be some confusion between gluon- and b jets. This would explain the deformations quite well, but cannot be confirmed simply based on these plots.

We see that the robustness agreement between dijet events and $\mathrm{t\bar{t}}$ events is generally good. However, there are some uncertainties present that are not observed in the standard candle events. These might require more attention and some kind of a correcting solution.

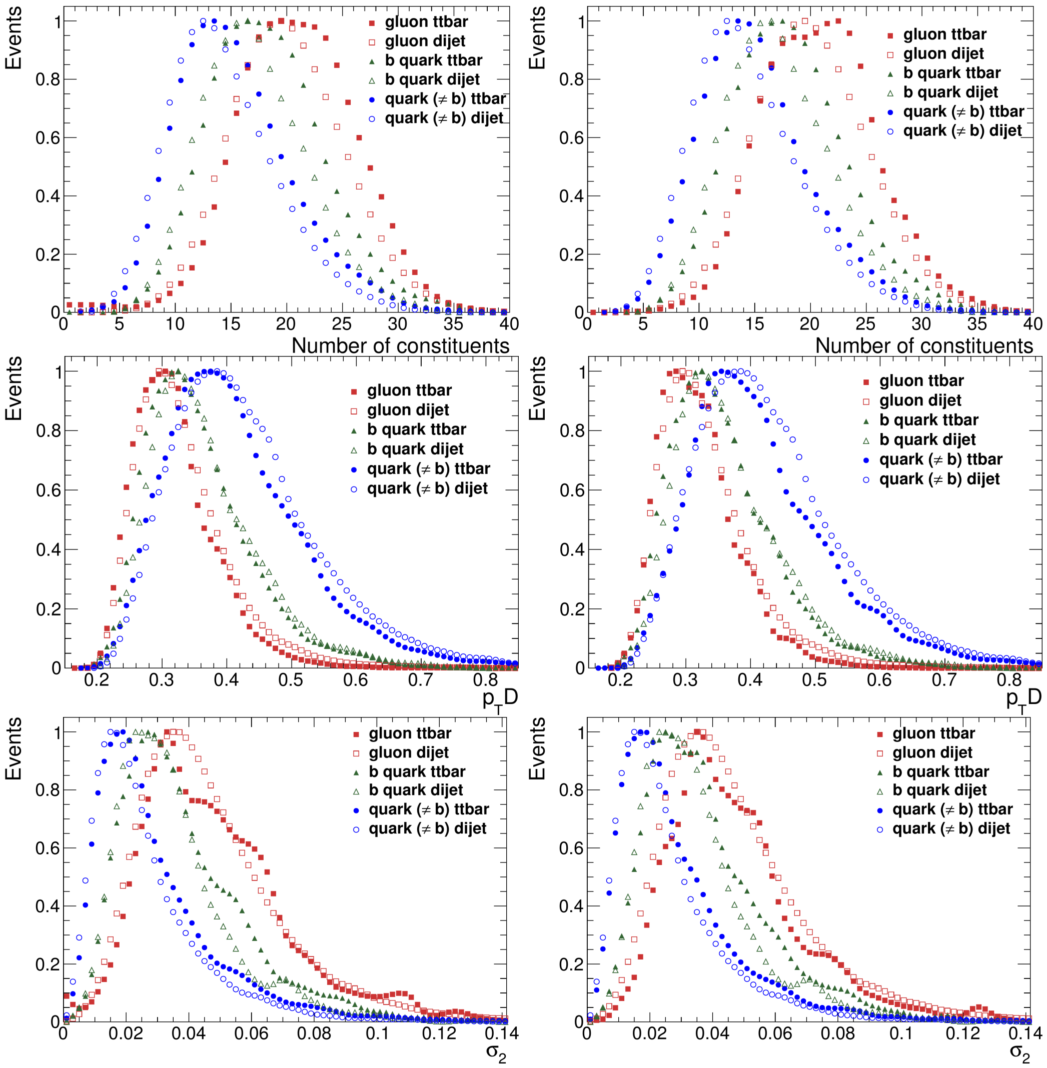


Figure 6.16: Robustness tests, dijet and ttbar lepton+jet events in Pythia 8. Right column: strict cuts for $t\bar{t}$ events, left column: loose cuts.

6.3.2 Elementary top mass measurement

We finish the studies with a handling of the t quark mass. This gives an understanding of the complexity of a measurement of the t quark properties. The mass study tries to emulate the challenges of a real mass measurement. The most important thing is that the missing transverse energy (MET) does not include non-transverse momentum information. In addition, approximating the neutrino that is paired up with the lepton with MET carries some sources of error. Furthermore, MET is calculated using all the available particle information, thus involving accumulated errors.

There are often 1 to 5 soft neutrinos present in the events. Some of the additional neutrinos originate from a τ-decay but the most have their origin in decaying D and B mesons. The effect of these neutrinos is small or to some degree averaged out due to their spatial distribution. Only the neutrino-related errors are present in the simulated MET values. In a way these are more severe than the statistical errors. Handling statistical errors is more trivial.

To obtain a value for the leptonic W boson mass, we need to reconstruct the z momentum of the neutrino. This is done by forcing the MET onto the mass shell of a W boson. Thus there will be some error from the use of MET as the neutrino transverse momentum and some error from using a delta-peak like mass instead of a Breit-Wigner peak. There are also some errors caused by the wrong choice of a MET z-component from the two options. To find the optimal jet permutations, loose constraints are applied to the W boson masses and the equality of the two t quark masses.

In Fig. 6.17 three different W-mass reconstructions are given. One is the best fit to jet results. This is convoluted in a way typical for jets and peaked slightly above the real W boson mass. For the ideal leptonic case (neutrino + charged lepton) we get a sharp Breit-Wigner peak. The reconstructed W boson mass is simply a delta-peak and it is not shown here. Instead, in the given plot the neutrino is reconstructed from MET by setting the momentum z component to zero. This gives a wide spectrum, but as a very interesting feature the mass peak is visible here.

In Fig. 6.18 similar plots for the t quark mass are shown. The jet-based mass is more convoluted, but a shift to the right is not visible anymore, as the energy scale is larger. The ideal leptonic case does not produce a Breit-Wigner-like peak, but the sharpness of this peak is still admirable. In addition, the distribution that uses the reconstructed neutrino (delta-peak W-mass) is presented. This ends up being slightly broader than the distribution obtained for jets. It is another question whether the leptonic reconstruction errors are mostly caused by the choice of the neutrino z-momentum direction or other sources. In the studies above we have obtained a broad view of the various challenges in the calculation of the t-mass.

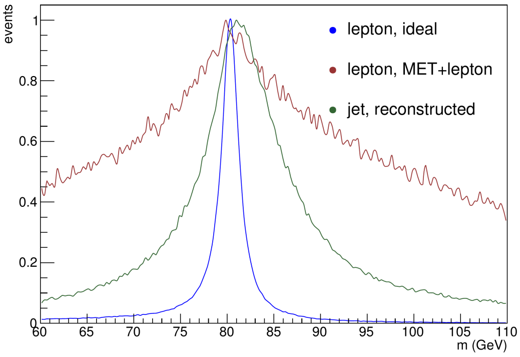


Figure 6.17: W boson mass plots

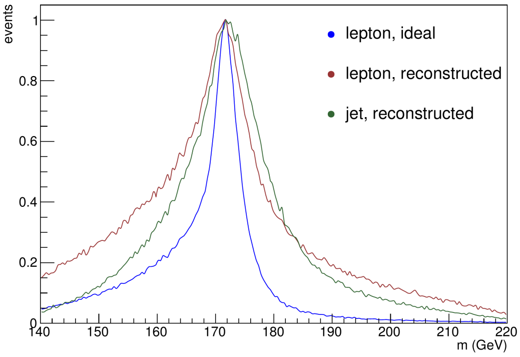


Figure 6.18: t quark mass plots

Chapter 7

Discussion

A large variety of results was presented in this work. In this chapter the most important indications of the findings are analyzed.

The studies started with a reproduction of robustness results of the physics flavor definition with the three GPMCs in use. Following this, a study between jet flavor fractions of the GPMCs pointed out an unnaturally large abundance of *no flavor* in PYTHIA 8. Also other, less significant curious phenomena were observed in the flavor fractions.

A strong focus was given to the shortcoming of the physics definition in PYTHIA 8. A group of new flavor definitions was created to deal with the detected problems. A significant enhancement for the physics definition was found using a couple of new definitions.

The studies were finished with consideration of top quark production. A new flavor definition was utilized to inspect the flavor structure of top-events. Finally, a measurement of the top mass was performed.

7.1 Robustness studies between GPMCs

In the beginning of this study we found out that the robustness condition is fulfilled well in each of the three GPMCs PYTHIA 6, PYTHIA 8 and HERWIG++. With robustness it is meant that the QGL parameters behave similarly for quark and gluon jets in the three *standard candle* event types. The study was done using the physics definition for jet flavor tagging. A common behavior in these event types indicates the generality of the jet properties. This finding is an important extension to a previous study of Ref. [1], which considered only PYTHIA 8.

When the GPMC results were compared with each other, the agreement was also good. Especially the two PYTHIA versions produced results very close to each other. This is essential in avoiding confusion in the ongoing transition from PYTHIA 6 to PYTHIA 8.

Between the PYTHIA and HERWIG results some differences were observed in the gluon jet QGL variable behavior. Quark jet properties were a good match. In terms of the PYTHIA results, the HERWIG gluon jets seem to be slightly more quark jet like. There can be several reasons for this phenomenon. There might be some fundamental underlying differences between the two generators, causing the difference. Furthermore, it is possible that the methodology of flavor tagging in HERWIG++ causes the difference. It could also be that some quark jets are accidentally mis-labeled as gluon jets, causing the difference.

The QGL parameter differences between the gluon jets in HERWIG and PYTHIA were, nevertheless, relatively small. Thus it can be stated that the observed quark and gluon jet properties are likely physically realistic. Yet, the source of the difference in the gluon jet QGL variables should be studied more closely in the future. If this is not caused by the flavor tagging mechanics, it can be used to estimate the errors of gluon jet properties. Since there is a disagreement, it could be that all the generators produce slightly biased results.

Studying the jet flavor fractions as a function of η and p_T revealed somewhat similar results. The abundance of different jet flavors had similar behavior in the different GPMCs. However, there were notable differences between the amount of cases in which no jet flavor was found. Between PYTHIA 6 and HERWIG++ the amount of a *none* flavor tags was somewhat similar. For PYTHIA 8 this fraction was considerably larger, inspiring a closer study of the phenomenon.

Some less dramatic observations were done using the flavor fractions. In the γ+jet events and the Zμμ+jet events the η-profile of the fractions turned out to be noisy. It was estimated that this could be a result from the event weighting, which is simply not distributed uniformly in the η direction. With these plots and this data it is not possible to tell whether this hypothesis is entirely correct or not.

Another feature of the flavor fractions was a distinctively different behavior in γ+jet events. The heavy and light quark flavor jets had a notably differing behavior in HERWIG++ events. The reason for this phenomenon was traced down to the hard process. It turned out that the same difference is present already in the flavor fractions of the outgoing LO partons. In studies of the differences between quark and gluon jets this does not have a great significance. Nevertheless, the observation is worth a closer study. As the next step it should be studied whether the newest versions of HERWIG++ and PYTHIA reproduce the result. It is possible that some development has been made.

As a final step the reason for the large rejection rates in PYTHIA 8 were studied. It was investigated how turning off ISR and/or FSR impacts the different GPMCs. A focus was given to the distance between the jet axis and the closest LO parton. It turned out that when ISR is off, all the generators produce very similar results. Adding ISR tended to broaden the ΔR profiles considerably, most notably in PYTHIA 8. The reason for this was traced down to the updated program logic. The momentum values of the LO partons are not anymore updated after adding ISR in PYTHIA 8. This kind of a change introduces a great motivation to update the jet flavor definitions.

The physics definition worked well during the Pythia 6 era. In the commencing Pythia 8 era we are probing higher energies with new versions of analysis software. Thus it is necessary to keep updating the flavor definitions.

7.2 New flavor definitions

The flavor definition studies were greatly motivated by the shortcomings of the physics definition. A group of new definitions was introduced to repair the observed flaws. The new definitions are mostly updated versions of the older flavor definitions. Special care was taken when selecting names of these definitions. The old definitions tend to have ambiguous titles, like the *algorithmic* and the *physics* definition. Such names carry little information and may cause confusion.

The physics definition was upgraded with small changes to what we call the LO definition. This uses the outgoing LO partons of the hard subprocess for determining the flavor. A further development of the LO definition was made in the momentum-corrected LO definition (the CLO definition). In this definition we tried to calculate a correction for the Pythia 8 outgoing LO parton momenta. Both the LO and CLO definition turned out to be effective upgrades of the physics definition. However, neither of them reached a fully satisfying coverage of the *none* flavors.

To understand this shortcoming the event structures were studied in detail. It turned out that the failing cases of the two new definitions were not fully correlated. Thus a combination of the two new definitions was motivated. The resulting LO & CLO definition reaches a very high flavor tagging efficiency. Its only downside is that only the LO-parton jets can be tagged. This becomes significant for instance in top-quark production, where many ISR and FSR jets are present.

In order to study other possibilities than derivatives of the physics definition, the *historic definition* was introduced. This is an attempt to propagate the outgoing LO parton information to the final state. The jet flavor was to be decided according to the *ancestors* of the final-state particles. Nonetheless, it turned out that the tags of descendants on the final-state level are propagated arbitrarily in Pythia 8. The hypothesis is that the parent-child relations are thoroughly mixed up in the hadronization process. Due to these results the historic definition was discarded, for now.

As a final new flavor definition a modification to the hadronic and algorithmic definitions was attempted. This is called the FS definition, as it uses *final-state*/pre-hadronization partons. Not that surprisingly, it turned out that the FS definition had a similar flavor profile as the hadronic definition. However, the FS definition provided special possibilities for use in combination with the other definitions. A definition using LO & FS was thus initiated. This equips the FS partons for flavor tagging only if the LO tagging fails.

The LO & FS definition has many favorable properties. First of all, this definition can be conveniently implemented in large software installations, such as the CMSSW. In contrast, the calculation of the CLO momentum values could pose a great problem in CMSSW. Secondly, this definition is useful in studying events with ISR and FSR jets. It is rare that any of the LO or radiative jets are tagged with a *none* flavor.

There is a slight preference for using the LO & CLO definition when only LO parton jets are studied. It is not clear how successful the LO & FS definition is tagging these jets in case the LO definition fails. It is possible that due to the use of the FS definition some gluon-splitting cases are missed. However, it is also possible that the FS definition performs better than the CLO definition when the LO definition fails. A closer study of the exact behavior is required in the future. Meanwhile, the use of the LO & FS definition is favorable, when non-LO jets are taken into account. In the scope of this thesis the definition produced good results and it has versatile flavor tagging properties.

7.3 Top quark measurements

In the final part of the thesis the ttbar lepton+jet events were studied. The analysis was equipped with the LO & FS definition, which suits well the study of LO, ISR and FSR jets. The jet flavor fraction profiles attained a very reasonable form. This can be considered as a first success of the jet flavor studies.

The robustness studies using the QGL parameters were more complicated than in the standard candle events. The profiles of the different jet flavors were in general similar in the ttbar lepton+jet events and dijet events. Yet, there were some notable differences. Jets in the $\mathrm{t\bar{t}}$ events had on average more constituents. A solid explanation for this phenomenon could not be found. It might just be a consequence of the GPMC features and the fact how much energy each final-state particle has.

The p_{TD} values were in good agreement, but in σ_2 slight differences were observed. This might be for instance due to the mixing of b jets and gluon jets in the flavor tagging process. There is no certainty of this, and the reason for the observed behavior should be studied more closely.

As a final remark of the QGL parameter studies the features of b jets should be noted. The b jets turned out to be distinctively different to the lighter quark flavors. In the p_{TD} distributions it seems the b jets are even closer to gluon jets than the light flavored jets. This is an important phenomenon to be aware of. It also introduces a surprisingly great risk of mixing between b jets and gluon jets.

In the top mass measurements many of the possible challenges were observed. In the $\mathrm{t\bar{t}}$ lepton+jet events some of the greatest challenges arise from inability to measure the outgoing neutrino momentum. Only the missing transverse momentum is known, but this often includes the momenta of other neutrinos. The neutrino can be reconstructed by forcing the sum of MET and the charged lepton momentum on the W-mass shell. This kind of a procedure carries much uncertainty.

Uncertainties are caused also by the handling of the W boson originating from jets. It seems essential that the b jets are effectively tagged. If these are not known, an even higher amount of possible jet combinations needs to be considered. Even if the two b jets are successfully tagged, it can be challenging to find the jets corresponding to the W boson. It was found out that there is no guarantee that these are included in the four leading-p_T jets.

With the two W bosons found, one still needs to make the correct combinations with the b jets. Only thus the top quarks can be reconstructed. According to the quality of the two W bosons, there can be various challenges in this final combination. If the leptonic W boson is reconstructed with poor quality, one cannot utilize the mass-symmetry between the two t quarks. It appears that the accuracy of the measurement suffers greatly if this symmetry cannot be utilized. There is still much to study before a more complete analysis can be done. Most importantly, the magnitudes of the different error sources need to be understood.

7.4 Future considerations

Much progress was made in the studies of this thesis, but at the same time numerous new questions arose. First of all, many semi-trivial upgrades need to be applied to the studies. This thesis concentrated on working at $\sqrt{s} = 8\,\mathrm{TeV}$. It is necessary to update all the studies to $\sqrt{s} = 13\,\mathrm{TeV}$, as the LHC run 2 is starting to gain momentum. Also some important software updates should be applied. Both Pythia 8 and Herwig++ have received important updates that are not used here.

The update of the energy scale and the software versions could have an important impact on many of the phenomena observed in this work. For instance in $\mathrm{t\bar{t}}$ events an upgrade in the energy scale is welcome, as the jets tended to not be very energetic. On the other hand, the updates could have an impact for instance on the difference between the Herwig++ and Pythia gluons. The small difference in the QGL parameters of gluons is one of the most interesting subjects requiring a closer study.

Another upgrade that is required after this work is the extension of many of the studies of this work to Pythia 6 and Herwig++. Both the flavor studies and the top quark studies were done only with Pythia 8 to probe new ideas. Especially the new flavor definitions should be tried out also on the other generators. Another remark is that the newest CMS tunes should be utilized. In this work the use of a common PDF was emphasized, forcing the use of older tunes. To keep up with the progress, the use of new tunes is favorable.

It is important to pay attention on the naming of the new *LO definition*. This is seen at the latest when NLO and NNLO generators are utilized. With these, the hard process will end up with NLO or NNLO results. This means that the phenomenon of gluon splitting within a jet might be contained within the hard process. Thus, attention needs to be paid when the hard process order is increased. At this point without further studies it is good to make the distinction that LO partons are used in this definition. An important future development is to see whether the definition needs enhancements in NLO event generation.

One of the open questions left is the use of the FS partons in the LO & FS definition. A preferable way to use FS partons could for instance be some kind of a combination of the partons within a jet. Ideally, this would reconstruct particles similar to the LO partons. However, this kind of a development can prove highly difficult to implement. The QCD-aware definition used in the ATLAS collaboration tries to achieve something similar to this. The QCD-aware robustness behavior does not reach the required quality.

In the top mass measurement one important prospect is the kinematic fit. It is not, however, clear whether this is the optimal method for the mass calculation. Because of the structure of the top events, advanced data processing methods are necessary. This is an area where there is no single right or wrong solution. The core of the problem is in retrieving the neutrino information and finding the correct jet combinations. The analysis methods need to be developed gradually. Future developments may also depend on how the use of the flavor definitions will develop.

Chapter 8

Conclusions

This work serves as a broad introduction to jets, jet flavors and their uses in the simulated top quark mass measurement. To begin with, the elemental theoretical concepts were introduced. The most important part of this introduction was the definition of jets in the ideal case. Aside of this, it was emphasized that jets are an essential tool in the analyses of HEP. Following this introduction to jets, the context of the CMS project at the LHC was explained. This is important to keep in mind throughout the studies.

Since this work revolves around general purpose Monte Carlo event generators, a thorough explanation of their internal logic was given. The whole event generation scheme is built around the hard process. This is the most energetic part of the proton-proton interaction, around which the rest of the collision events can be conveniently built. Even with the given level of explanation, only a shallow understanding of the GPMCs is obtained.

In addition to the GPMC explanation the purpose and logic of various other software packages was explained. One of the most important tools in the analysis phase is jet clustering. The clustering process is closely tied to the modern jet flavor algorithms. Moreover, the clustering algorithms explain some of the shortcomings of the jets found in data.

The result-section of this work was commenced with a presentation of jet flavor robustness results. This was done using the physics flavor definition, which had been previously found to have good properties in terms of robustness. The results were obtained using the three *standard candle* event types with the three GPMCs Pythia 6, Pythia 8 and Herwig++. It was seen that there is a fairly good correspondence of the jet properties between these different samples. This is an encouraging result in the general classification of jet properties.

In the robustness studies it was found that an unnaturally large fraction of Pythia 8 jets are left without a flavor tag. This was investigated in depth and it was found that the cause for the problem was ISR. Apparently, in Pythia 8 the effect of ISR is not corrected to the hard process partons. This gave a strong motivation to update the physics definition.

In the following part of the studies some experimental flavor definitions were tried out. It was found that no single definition produces a sufficiently low none-flavor rate together with physically satisfying results. However, some combined definitions performed very well. In the core of both of the hybrid definitions was the upgraded physics definition - i.e. the LO definition. Of these hybrids, the LO & CLO definition and the LO & FS definition the latter turned out to be slightly more favorable. It is easier to implement and more versatile in its uses. Thus, with the finding of working upgrades to the physics definition, the flavor studies were a success.

In the final part of the thesis we moved on to study top quark production. The new LO & FS flavor definition allowed comprehensive flavor studies in these events. It was observed that there is a distinctive flavor distribution and that the average jet energy is relatively low. Thus as this analysis is upgraded to $\sqrt{s} = 13\,\mathrm{TeV}$, better statistics at the higher jet energies can be expected. The robustness results of flavors between dijet events and $\mathrm{t\bar{t}}$ events were encouraging. There were some differences, but in general the jet behavior in these events was similar. In these studies an important distinction was made between the lighter flavored jets and b jets.

Finally, we arrived to the measurement of the top mass. Some of the difficulties of using real detector data were recreated. The neutrino paired up with the charged lepton has to be constructed only by using MET. This means that there is some error in the transverse part of the neutrino momentum and that the z-component has to be completely reconstructed. Additionally, when it comes to the t quark going into light-flavored jets, a multitude of combinations has to be considered. This study gives a good basis on further developments of a precise mass measurement during the LHC run 2.

When arriving to the end of the studies of this thesis it became clear that many things still remain to be done. Since the work consists of an intensive use of software, the newer software versions need to be tried out. With the use of simulation software for studies comes an endless variation of parameter values and tunes.

The most important achievement of this thesis can be considered to be the progress with the flavor definitions. We started by demonstrating problems with the existing definitions. Then we moved on to propose various new definitions. The study was finished by a practical demonstration of the use of a new flavor definition. This forms a complete arc of problem-identification and solving.

Bibliography

[1] A. Abhishek, “Jet Flavor Studies at CMS”, Master’s thesis, Indian Institute of Technology Roorkee, India, 2015. https://cernbox.cern.ch/index.php/s/c6214da1ca09d359735f4acbebd1fe74.

[2] I. J. R. Aitchison and A. J. G. Hey, “Gauge Theories in Particle Physics: a practical introduction”, volume 1-2. Taylor & Francis group, 4th edition, 2013.

[3] M. E. Peskin and D. V. Schroeder, “An Introduction to Quantum Field Theory”. Addison-Wesley Publishing Co., 1st edition, 1995.

[4] F. Halzen and A. D. Martin, “Quarks and Leptons”. John Wiley & sons, 1st edition, 1984.

[5] V. A. Schegelsky et al., “A note on rapidity distributions at the LHC”, arXiv:1010.2051.

[6] MissMJ via Wikimedia, “Standard Model of Elementary Particles”. https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg, 2014. [Online; accessed 11-December-2015], License CC-BY 3.0, No changes to the original have been made. https://creativecommons.org/licenses/by/3.0.

[7] Eric Drexler via Wikimedia, “Elementary Particle interactions in the Standard Model”. https://commons.wikimedia.org/wiki/File:Elementary_particle_interactions_in_the_Standard_Model.png, 2014. [Online; accessed 11-December-2015], License CC0 1.0.

[8] CMS Collaboration, “The CMS experiment at the CERN LHC”, *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[9] CMS Collaboration, “Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET”, Technical Report CMS-PAS-PFT-09-001, CERN, 2009. Geneva, (Apr, 2009).

[10] CERN, “LHC machine outreach”. http://lhc-machine-outreach.web.cern.ch/lhc-machine-outreach/lhc_in_pictures.htm, 2015. [Online; accessed 20-December-2015], CERN copyright http://copyright.web.cern.ch/.

[11] LHCb Collaboration, “Observation of J/ψp Resonances Consistent with Pentaquark States in $\Lambda_b^0 \to$ J/ψK$^-$p Decays”, *Phys. Rev. Lett.* **115** (2015) 072001, doi:10.1103/PhysRevLett.115.072001, arXiv:1507.03414.

[12] T. A. collaboration, “Search for resonances decaying to photon pairs in 3.2 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector”,.

[13] CMS Collaboration, “CMS detector design”. http://cms.web.cern.ch/news/cms-detector-design, 2011. [Online; accessed 21-December-2015], CERN copyright http://copyright.web.cern.ch/.

[14] CMS Collaboration, “CMS Software”. https://github.com/cms-sw/cmssw, 2015. [Online; accessed 22-December-2015].

[15] T. Sjostrand, S. Mrenna, and P. Z. Skands, “PYTHIA 6.4 Physics and Manual”, *JHEP* **05** (2006) 026, doi:10.1088/1126-6708/2006/05/026, arXiv:hep-ph/0603175.

[16] T. Sjostrand, S. Mrenna, and P. Z. Skands, “A Brief Introduction to PYTHIA 8.1”, *Comput. Phys. Commun.* **178** (2008) 852–867, doi:10.1016/j.cpc.2008.01.036, arXiv:0710.3820.

[17] T. Sjostrand et al., “An Introduction to PYTHIA 8.2”, *Comput. Phys. Commun.* **191** (2015) 159–177, doi:10.1016/j.cpc.2015.01.024, arXiv:1410.3012.

[18] M. Bahr et al., “Herwig++ Physics and Manual”, *Eur. Phys. J.* **C58** (2008) 639–707, doi:10.1140/epjc/s10052-008-0798-9, arXiv:0803.0883.

[19] M. A. Dobbs et al., “Les Houches guidebook to Monte Carlo generators for hadron collider physics”, in *Physics at TeV colliders. Proceedings, Workshop, Les Houches, France, May 26-June 3, 2003*, pp. 411–459. 2004. arXiv:hep-ph/0403045.

[20] CTEQ Collaboration, “Handbook of perturbative QCD; Version 1.1: September 1994”, *Submitted to: Rev. Mod. Phys.* (1994).

[21] J. Alwall et al., “MadGraph 5 : Going Beyond”, *JHEP* **06** (2011) 128, doi:10.1007/JHEP06(2011)128, arXiv:1106.0522.

[22] C. Oleari, "The POWHEG-BOX", *Nucl. Phys. Proc. Suppl.* **205-206** (2010) 36–41, doi:10.1016/j.nuclphysbps.2010.08.016, arXiv:1007.3893.

[23] S. Alioli et al., "Vector boson plus one jet production in POWHEG", *JHEP* **01** (2011) 095, doi:10.1007/JHEP01(2011)095, arXiv:1009.5594.

[24] A. V. Ferapontov, "Search for particles decaying into $Z\gamma$ at D0", in *Proceedings, 34th International Conference on High Energy Physics (ICHEP 2008)*. 2008. arXiv:0810.1751. License CC-BY 4.0, No changes to the original have been made. https://creativecommons.org/licenses/by/4.0.

[25] GEANT4 Collaboration, "GEANT4: A Simulation toolkit", *Nucl. Instrum. Meth.* **A506** (2003) 250–303, doi:10.1016/S0168-9002(03)01368-8.

[26] T. Gleisberg et al., "Event generation with SHERPA 1.1", *JHEP* **02** (2009) 007, doi:10.1088/1126-6708/2009/02/007, arXiv:0811.4622.

[27] A. Buckley et al., "LHAPDF6: parton density access in the LHC precision era", *Eur. Phys. J.* **C75** (2015) 132, doi:10.1140/epjc/s10052-015-3318-8, arXiv:1412.7420.

[28] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-kt jet clustering algorithm", *JHEP* **04** (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[29] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual", *Eur. Phys. J.* **C72** (2012) 1896, doi:10.1140/epjc/s10052-012-1896-2, arXiv:1111.6097.

[30] CMS Collaboration, "Performance of quark/gluon discrimination in 8 TeV pp data", Technical Report CMS-PAS-JME-13-002, CERN, Geneva, (2013).

[31] F. Rademakers et al., "ROOT Status and Future Developments", *CoRR* **cs.SE/0306078** (2003).

[32] H. Siikonen, "A software environment for jet physics studies". https://github.com/errai-/jetscripts, 2016. [Online; accessed 13-March-2016].

[33] Particle Data Group Collaboration, "Monte Carlo particle numbering scheme". http://pdg.lbl.gov/2015/reviews/rpp2015-rev-monte-carlo-numbering.pdf, 2015. [Online; accessed 18-March-2016].

Appendix A

Event generator settings

The event generator settings consist of a couple of layers of choices. These choices can be found from the code of the github repository of Ref. [32]. To give a better understanding of the used choices the settings are repeated here. The event types, collision energies, PDFs and tunes have to be selected explicitly. In this work the CMS project tunes and the collision energy $\sqrt{s} = 8\,\text{TeV}$ are generally used. The tunes are used with the PDF CTEQ6L1 in order to exclude differences caused by the choice of PDF.

A.1 Pythia 6

In this work the version Pythia 6.4.28 is used. The Pythia 6 settings are stored in common blocks. These are not very informative, so short comments are given for the choices. The settings are given in a pseudocode format. We begin with generic settings.

```
MSTU(21) = 1 // Check for errors
MSTJ(22) = 2 // Unstable particle decay
PARJ(71) = 10 // ctau = 10 mm
MSTP(33) = 0 // no K factors in hard cross sections
MSTP(2) = 1 // which order running alphaS
MSTP(51) = 10042 // Structure function (PDF CTEQ6L1)
MSTP(52) = 2 // LHAPDF
MSTP(142) = 2 // Turn on Pt reweighting
Initialize("cms", "p", "p", 8000) // p-p at 8 TeV
```

In the following, the relevant settings for the CMS Pythia 6 tune Z2* are given.

```
PARP(82) = 1.921 // pt cutoff, multiparton interactions
PARP(89) = 1800. // sqrts for which parp82 is set
PARP(90) = 0.227 // MPI: rescaling power
MSTP(95) = 6 // Color reconnection setParams
```

```
PARP(77) = 1.016 // CR
PARP(78) = 0.538 // CR
PARP(80) = 0.1 // Prob. colored parton from BBR
PARP(83) = 0.356 // MPI matter distribution
PARP(84) = 0.651 // MPI matter distribution
PARP(62) = 1.025 // ISR cutoff
MSTP(91) = 1 // Gaussian primordial KT
MSTP(93) = 10.0 // Primordial KT-max
MSTP(81) = 21 // MPI
MSTP(82) = 4 // MPI model
```

Finally, we proceed to the event type specific settings. The dijet events are obtained with relatively simple standard QCD settings:

```
MSEL = 1 // Standard QCD
CKIN(3) = 25 // Min pthat
CKIN(4) = 3000 // Max pthat
```

For the γ+jet events the standard QCD processes are turned off. Then, specific settings are turned back on:

```
MSEL = 0 // Standard QCD off
MSUB(14) = 1 // Subprocess 1 on
MSUB(29) = 1 // Subprocess 2 on
MSUB(115) = 1 // Subprocess 3 on
CKIN(3) = 10 // Min pthat
CKIN(4) = 3000 // Max pthat
```

The Zμμ+jet events are in principle similar to the γ+jet events, but the required settings are more subtle. We want to include only Z decays to a $\mu^+\mu^-$ pair. In addition, there is a cut for $\hat{m}$, i.e. the virtual Z boson mass. This excludes most of the possible virtual photon processes.

```
MSEL = 13 // y*/Z + f/g on the final state
// Leave only decay to muons on:
MDME(174,1) = 0 // Z decay to d dbar
MDME(175,1) = 0 // Z decay to u ubar
MDME(176,1) = 0 // Z decay to s sbar
MDME(177,1) = 0 // Z decay to c cbar
MDME(178,1) = 0 // Z decay to b bbar
MDME(179,1) = 0 // Z decay to t tbar
MDME(182,1) = 0 // Zee
MDME(183,1) = 0 // Znuenue
MDME(184,1) = 1 // Zmumu
MDME(185,1) = 0 // Znumunumu
MDME(186,1) = 0 // Ztautau
MDME(187,1) = 0 // Znutaunutau
```

```
CKIN(1) = 40 // mhat min
CKIN(2) = -1 // mhat max
CKIN(3) = 15 // pthat min
CKIN(4) = 3000 // pthat max
```

The final event type is that of the $t\bar{t}$ events. It would be preferable to select only events with one W boson decaying into quarks and one decaying into leptons. In PYTHIA 6 this is difficult to accomplish, and not done on the generator level. Thus the event selection can only be made once the whole event has been generated. The settings are simply those of generating pairs of heavy quark flavors.

```
MSEL = 6 // choose top quark
MSUB(81) = 1 // qqbar -> qqbar
MSUB(82) = 1 // gg->qqbar
PMAS(6,1) = 172 // set top mass
CKIN(3) = 25 // pthat min
CKIN(4) = 3000 // pthat max
```

When necessary, ISR, FSR and MPI can be turned off using the following flags:

```
MSTP(61) = 0 // ISR off
MSTP(71) = 0 // FSR off
MSTP(81) = 0 // MPI off
```

A.2 Pythia 8

This work uses the version PYTHIA 8.2.12. In PYTHIA 8 the settings are similar to the ones in PYTHIA 6. As opposed to the less intuitive PYTHIA 6 common blocks, the settings are tuned in a text mode. To begin with, generic type of settings:

```
Next:numberShowInfo = 0
Next:numberShowProcess = 0
Next:numberShowEvent = 0
Next:numberCount = 0
// Allow photon radiation in lepton-pair decays
ParticleDecays:allowPhotonRadiation = on
// Set particles with long enough lifetimes to stable
ParticleDecays:limitTau0=on
ParticleDecays:tau0Max=10.
// Event weighting
PhaseSpace:bias2Selection = on
PhaseSpace:bias2SelectionPow = 4.5
PhaseSpace:bias2SelectionRef = 15.
// CM energy
Beams:eCM = 8000.
```

Also the tune settings have to be set again for all event types. We use the CMS tune CUETP8S1, which pairs up with CTEQ6L1. Recently a newer tune has been adopted by the CMS experiment. In these studies more emphasis is put on having a common PDF than using the newest tune.

```
Tune:preferLHAPDF = 2
PDF:pSet = LHAPDF6:cteq6l1
// CMS UE Tune CUETP8S1-CTEQ6L1
Tune:ee 3
Tune:pp 5
MultipartonInteractions:pT0Ref=2.1006
MultipartonInteractions:ecmPow=0.21057
MultipartonInteractions:expPow=1.6089
MultipartonInteractions:a1=0.00
ColourReconnection:range=3.31257
```

The dijet settings are such that they allow hard QCD events to be produced. A more specific dijet selection is done in later phases of analysis.

```
HardQCD:all = on
PhaseSpace:pTHatMin = 30.
```

In γ+jet events we turn on the specific processes:

```
PromptPhoton:qg2qgamma = on
PromptPhoton:qqbar2ggamma = on
PromptPhoton:gg2ggamma = on
PhaseSpace:pTHatMin = 10.
```

And once again in Zμμ+jet events we choose only the necessary event types.

```
WeakZ0:gmZmode = 2 // Produce only Z0
WeakBosonAndParton:qqbar2gmZg = on
WeakBosonAndParton:qg2gmZq = on
23:onMode
23:7:onMode = on // Z decays only to muons
PhaseSpace:pTHatMin = 20.
PhaseSpace:mHatMin = 75.
```

The top events are handled by a specific process. An additional user-implemented module (*user hook*) has been created. This selects events with one W decaying into quarks and one into leptons at the hard process level. Thus the simulations have been made considerably faster than without the module.

```
Top::gg2ttbar = on
Top::qqbar2ttbar = on
PhaseSpace:pTHatMin = 30.
```

To conclude, the ISR/FSR/MPI off settings are provided:

```
PartonLevel:MPI = off
PartonLevel:ISR = off
PartonLevel:FSR = off
```

A.3 Herwig++

This work utilizes Herwig++ 2.7.1 and ThePEG 1.9.2. Also in Herwig++ the settings follow a similar structure as in the two Pythia versions. The settings are tuned in a tree structure. Generic settings:

```
cd /Herwig/Generators
set LHCGenerator:DebugLevel 3
set LHCGenerator:UseStdout 0
set LHCGenerator:PrintEvent 10
set LHCGenerator:MaxErrors 10000
set LHCGenerator:EventHandler:StatLevel Full
set /Herwig/Decays/DecayHandler:MaxLifeTime 10*mm
set /Herwig/Generators/LHCGenerator:EventHandler:LuminosityFunction:
   Energy 8000.0*GeV
set /Herwig/Shower/Evolver:IntrinsicPtGaussian 2.0*GeV
mkdir /Herwig/Weights
cd /Herwig/Weights
create ThePEG::ReweightMinPT reweightMinPT ReweightMinPT.so
set reweightMinPT:Power 4.5
set reweightMinPT:Scale 10*GeV
insert SimpleQCD:Preweights[0] /Herwig/Weights/reweightMinPT
```

For Herwig++ we use the CMS tune CUETHS1 (PDF: CTEQ6L1).

```
// Energy extrapolation:
set /Herwig/UnderlyingEvent/MPIHandler:EnergyExtrapolation Power
set /Herwig/UnderlyingEvent/MPIHandler:ReferenceScale 7000.*GeV
set /Herwig/UnderlyingEvent/MPIHandler:Power 0.3705288
set /Herwig/UnderlyingEvent/MPIHandler:pTmin0 3.91*GeV
// Colour reconnection settings:
set /Herwig/Hadronization/ColourReconnector:ColourReconnection Yes
set /Herwig/Hadronization/ColourReconnector:ReconnectionProbability
   0.5278926
// Colour Disrupt settings:
set /Herwig/Partons/RemnantDecayer:colourDisrupt 0.6284222
// inverse hadron radius
set /Herwig/UnderlyingEvent/MPIHandler:InvRadius 2.254998
// MPI model settings
set /Herwig/UnderlyingEvent/MPIHandler:softInt Yes
set /Herwig/UnderlyingEvent/MPIHandler:twoComp Yes
set /Herwig/UnderlyingEvent/MPIHandler:DLmode 2
// LHAPDF settings
cd /Herwig/Partons
create ThePEG::LHAPDF customPDF ThePEGLHAPDF.so
set customPDF:PDFName cteq6l1
set customPDF:RemnantHandler /Herwig/Partons/HadronRemnants
set /Herwig/Particles/p+:PDF customPDF
set /Herwig/Particles/pbar-:PDF customPDF
```

The dijet events are selected once again by turning generic QCD events on:

```
set /Herwig/Cuts/JetKtCut:MinKT 30.0*GeV
insert SimpleQCD:MatrixElements[0] MEQCD2to2
```

The photon+jet events are created similarly with a specialized function:

```
set /Herwig/Cuts/JetKtCut:MinKT 10.0*GeV
insert SimpleQCD:MatrixElements[0] MEGammaJet
```

Zμμ+jet events have a similar function; muon decay chosen separately:

```
set /Herwig/Cuts/JetKtCut:MinKT 20.0*GeV
set MEZJet:ZDecay Muon
set MEZJet:GammaZ All
set /Herwig/Cuts/QCDCuts:MHatMin 75.0*GeV
insert SimpleQCD:MatrixElements[0] MEZJet
```

Also the top quark events are produced using special functions:

```
set /Herwig/Cuts/JetKtCut:MinKT 30.0*GeV
set MEHeavyQuark:QuarkType Top
set MEHeavyQuark:Process Pair
insert SimpleQCD:MatrixElements[0] MEHeavyQuark
```

For Herwig++ the options for turning ISR/FSR/MPI off are:

```
set /Herwig/Shower/SplittingGenerator:ISR No
set /Herwig/Shower/SplittingGenerator:FSR No
set /Herwig/Shower/ShowerHandler:MPIHandler NULL
```

Appendix B

Implementation of the simulation and analysis code

The most significant bottleneck in the chain of simulations and data analysis is the generation of the full event history. The history consists of all the particles that have been present in the radiation, decay and hadronization chain. Each particle entry contains essential information, such as the momentum and an ID-number indicating the type of the particle. The structure of the event history is conveyed by lists of the mother and daughter particles of each particle.

With a given random number seed the produced event history is always the same. Moreover, the elementary simulation parameters are not changed very frequently. Thus re-generating the particle level data for each analysis is not necessary. It was concluded that a temporary storage to the particle-level data was needed. The greatest challenge with this conclusion is the need for large storage spaces. For good statistics millions of events are needed. Each event can contain thousands of particles.

To minimize the use of storage space, the ROOT software [31] was used. It creates *.root* files that employ a binary data format. Within ROOT there are data structures that make the saving of collision event data ideal. In this work the ROOT version 6 is used, as it brings some convenient fixes to ROOT 5.

To reach a good efficiency, only the particle data needed by the analyses are saved. All the final-state particles are kept, but only some of the temporary particles are saved. This implementation has the benefit that the particle data from each GPMC can be saved in the same format. Each of the generators needs to be handled separately, but after this all the analysis is perfectly generic. The GPMC versions and settings used in this work are presented in Appendix A.

The next level of the analysis is the clustering of final-state particles into jets and determining some jet properties. Given that the particle-level data has been saved, the greatest remaining bottleneck is the jet clustering algorithm. The jet clustering times are considerably shorter than those of the generation of particle-level data. However, with large amounts of data also this phase can take hours.

Thus also the jet data is saved into a similar event structure as the particle data. Since there are only some jets per event, these files are more compact than those of the particle-level data. Along with the jets some preliminary analysis results are saved, such as the jet flavors. This allows the discarding of particle-level data without great losses of information. In this work the jet clustering is done with the FastJet version 3.1.2.

Finally, the processed jet data is turned into graphs. This phase is the one repeated the most, especially concerning plots going into presentations and publications. The necessity of the final polishing of plots should not be underestimated. Also some final parameter cuts can be left to be done at the plotting phase, allowing a fine-tuning of the results. This requires cautiousness, so that it is always known what cuts have been made. The whole analysis process is shown in the flowchart of Fig. B.1.

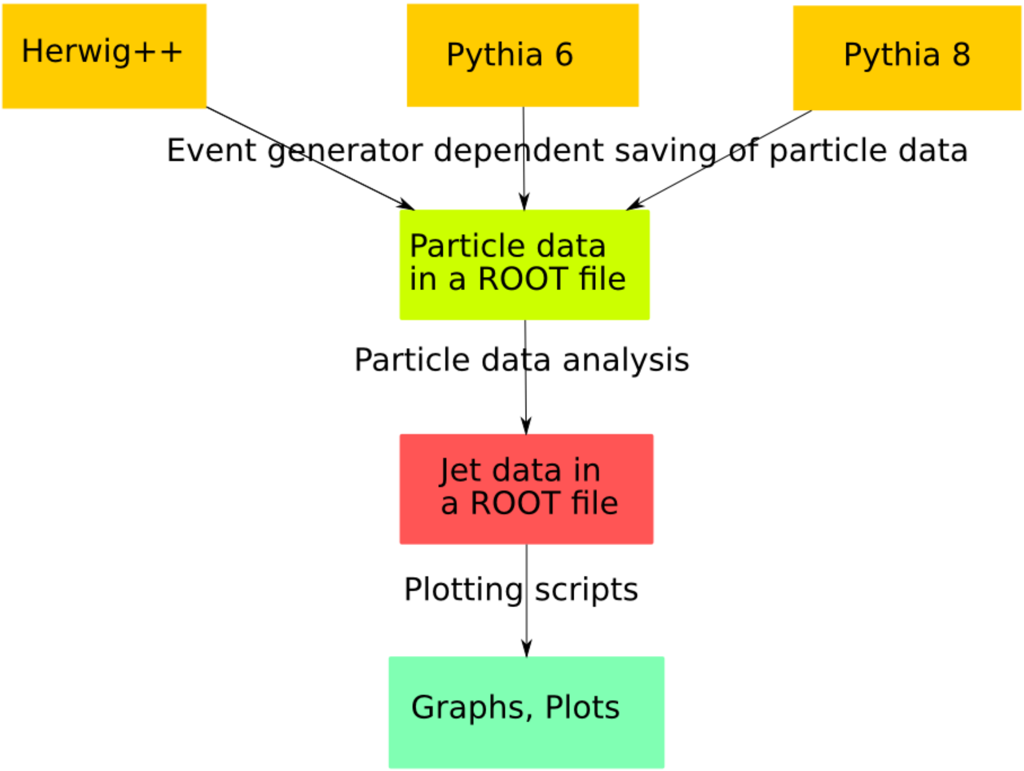


Figure B.1: A flowchart of event generation and data analysis.

In the shorter runs it is sufficient to use a good modern computer. However, this becomes challenging when the amount of GPMCs, GPMC tunes and collision energies is increased. Most importantly, the storage requirements are difficult to handle. The CERN server LXPLUS was used to deal with these issues. The most significant bottleneck, event generation, was simple to parallellize. Separate processes need to be given separate seed values and the final results need to be combined. Using ROOT this is straightforward.

The jet clustering phase is considerably more difficult to parallellize. A parallellization is not worth implementing, since the run times are generally short enough (a few hours). It is most convenient to copy the analyzed jet data to a local computer for plotting and final analyses purposes. Thus, the CERN server provided both computation and storage resources for this work. This was indispensable for analyses to be made in the present scope.

The whole software implementation can be found in the open github repository of Ref. [32]. The repository is roughly divided into the section of event generation & jet clustering and the section of analysis & plotting. The github source can be used to get the whole software running. However, the external packages such as PYTHIA 8 have to be installed separately. Basic installation instructions are given in the repository. As a word of warning, PYTHIA 6 and HERWIG++ may be challenging to set up. For the HERWIG implementation some changes have been made to THEPEG source code. Without these alterations the code will not work. The changes were necessary because of the stiffness of the HERWIG++ interface. PYTHIA 6 is utilized with the help of ROOT, so using this is not simple, either.

Appendix C

Listing of example events

The subject of the quality of jet flavor definitions is theoretically challenging. The best method for studying the possible shortcomings of the definitions is to have a look at single events. In this scrutinous study we inspect a sample of 600 dijet events. The events were required to have $\alpha < 0.3$, back-to-back angle $\phi > 2.8$ and $p_T > 30\,\mathrm{GeV}$ for the two leading jets.

A listing of events in which any of the flavor definitions studied has failed is provided in Table C.1. The reasons for failing were obtained using the visualization format presented in Appendix D. The performance of the LO, CLO, LO & CLO and LO & FS definitions is shown. For each definition the found flavor is indicated. A zero flavor means that no match has been found.

We consider the LO & FS definition to *fail* when it gives a different result than the LO & CLO flavor (if this is non-zero). This is due to the fact that the FS definition finds a flavor practically always. Thus, when the LO & CLO definition finds a flavor, the result is in some ways more robust than the LO & FS flavor. For the FS partons, it is possible that there are multiple FS partons within a jet. Thus e.g. gluon splitting jet can lead to bad results. That is, gluon splitting where both the quarks end up in one jet.

The color coding of the table is given to facilitate reading. Red and green indicate a failed or a successful flavor tagging, correspondingly. A disagreeing algorithmic flavor is marked with blue and a sole algorithmic flavor is marked with yellow. In the descriptions, physically uninteresting reasons are marked with gray. This means events in which flavor tagging has been unsuccessful only because of the underlying MC technicalities. In the few physically interesting events it is questionable whether the LO partons should be used for flavor tagging.

It is observed that both the hybrid definitions handle the physically uninteresting cases well. Sometimes there is a disagreement, which leaves an uncertainty in the flavor. Both the hybrid definitions have a weakness in the gluon splitting. The LO & CLO definition fails to understand gluon splitting, which produces separate jets. Conveniently, the LO & FS definition always assigns some flavor to the jets. Thus even the non-LO jets are given a flavor.

Table C.1: Events with unsuccessful flavor tagging for four flavor definitions: LO, CLO, LO & CLO and LO & FS (LO = leading order partons, CLO = momentum-corrected partons, FS = final-state partons). The flavor numbers follow the particle data group MC numbering scheme [33]. 0=failure, 1=u, 2=d, 3=s, 4=c, 5=b, 21=g.

#event	LO	CLO	LO & CLO	LO & FS	Reason of failing.
16	0	0	0	2	An initial parton is split into 2 distinct jets.
23	1	0	1	1	Corrected parton rapidity slightly off.
31	21	0	21	21	A very wide jet/a jet with small side jets.
36	21	0	21	21	An initial parton is split into 3 distinct jets.
44	0	21	21	1	Original parton rapidity slightly off.
50	0	21	21	21	Original parton rapidity slightly off.
65	0	21	21	21	Original parton rapidity slightly off.
110	0	21	21	21	An initial parton is split into 2 distinct jets.
113	0	0	0	1	Original parton rapidity off, strong radiation.
126	0	0	0	1	An initial parton is split into 2 distinct jets.
141	0	21	21	21	Original parton rapidity slightly off.
142	21	0	21	21	Corrected parton rapidity off.
151	21	0	21	21	Corrected parton rapidity slightly off.
154	0	2	2	4	Original parton rapidity slightly off.
168	21	0	21	21	Corrected parton ϕ off.
169	21	0	21	21	Corrected parton rapidity off.
174	0	21	21	3	Original parton rapidity slightly off.
177.1	0	3	3	3	Original parton rapidity slightly off.
177.2	0	0	0	21	An initial parton is split into 2 low-energy jets.
191	0	2	2	2	Original parton rapidity slightly off.
207	0	0	0	21	Both parton rapidities slightly off.
226	1	0	1	1	Corrected parton rapidity slightly off.
243	21	0	21	21	Corrected parton rapidity slightly off.
295	0	0	0	21	Second jet is not collimated.
296	0	1	1	21	Original parton rapidity slightly off.
306	21	0	21	21	Corrected parton rapidity slightly off.
311	0	0	0	21	An initial parton is split into numerous jets.
340	0	0	0	2	An initial parton is split into two distinct jets.
359	1	0	1	1	Corrected parton rapidity slightly off.
365	0	0	0	21	An initial parton is split into numerous jets.
373	21	0	21	21	Corrected parton rapidity slightly off.
378	21	0	21	21	Corrected parton rapidity off.
383	21	0	21	21	Corrected parton rapidity slightly off.
390	0	21	21	21	Original parton rapidity slightly off.
406	0	0	0	1	Original parton rapidity off, strong radiation.
418	0	2	2	2	Original parton rapidity slightly off.
446	0	21	21	2	Original parton rapidity slightly off.
517	0	21	21	1	Original parton rapidity off.
527	1	0	1	1	Corrected parton rapidity slightly off.
531	0	21	21	2	A wide jet.
532	21	0	21	21	Corrected parton rapidity off.
590	0	1	1	1	Original parton rapidity off.
595	0	21	21	1	Original parton rapidity off.
599	0	3	3	21	Original parton rapidity slightly off.

Appendix D

Event visualizations

It is useful to look at the (y, ϕ)-structures of the events. We take into a special study the events with failing flavor definitions, given in Appendix C. This visualization is the way in which the descriptions for reasons of failing were obtained.

In the following figures there is a lot of information, so the notation is somewhat subtle. It is beneficial to present as much information as possible in these figures to get a full view of the situation. As a basis, there is a heatmap that indicates a Gaussian filtering of the p_T values of the final-state particles. Therefore it is important to note that the heatmap is not showing any strict physical quantity. High-energy jets produce wide profiles even if the jets are very well collimated. However, this provides an excellent method for understanding the jet structure. On the heatmap circles with $R = 0.5$ are drawn to show the jet positions. A color coding is given to indicate the jet p_T ordering: in a descending order the four leading jets are dark red, light red, magenta and cyan. All the possible remaining jets are dark blue.

Onto the plot described above, some partonic-level information is drawn. All the FS partons are plotted. The marker size of each parton is proportional to its p_T on a log scale. Gluons are red squares, quarks blue circles and anti-quarks magenta circles.

Also the LO partons are indicated. For these the marker size is given an additional multiplier of 2. Gluons are given as green squares, quarks as dark blue circles and anti-quarks as cyan circles. The corrected LO parton momenta are marked with a magenta cross. The ϕ-axis is drawn from $-\pi$ to $3\pi/2$ in order to get a good visualization also for the border zone. Jet circles are drawn only once, but the other markers are drawn in their every manifestation.

The short descriptions of the drawn events can be checked in the Table C.1. The zeroth event (Fig. D.1) is an example of a normal event in which jets are clustered successfully. Most of the events are like this. The unsuccessful events are simply more interesting in a debugging-sense.

In Figs. D.2 and D.3 such events are given in which the LO or the CLO definition fails. This manifests usually as a small deviation in the rapidity-direction. In

Fig. D.4 a first example of the failure of both the LO and CLO definitions is given. One of the LO partons has been split into two or three distinct jets. Fig. D.5 gives a generic example of a wide jet and the trouble it causes.

Fig. D.6 shows trouble caused by low-energetic jets. The second jet is not collimated at all, contradicting the basic assumptions concerning jets. Finally, in Fig D.7 a very energetic event is given; a gluon splits into numerous jets. With this selection a brief but thorough listing an understanding of the various problems with flavor definitions is obtained.

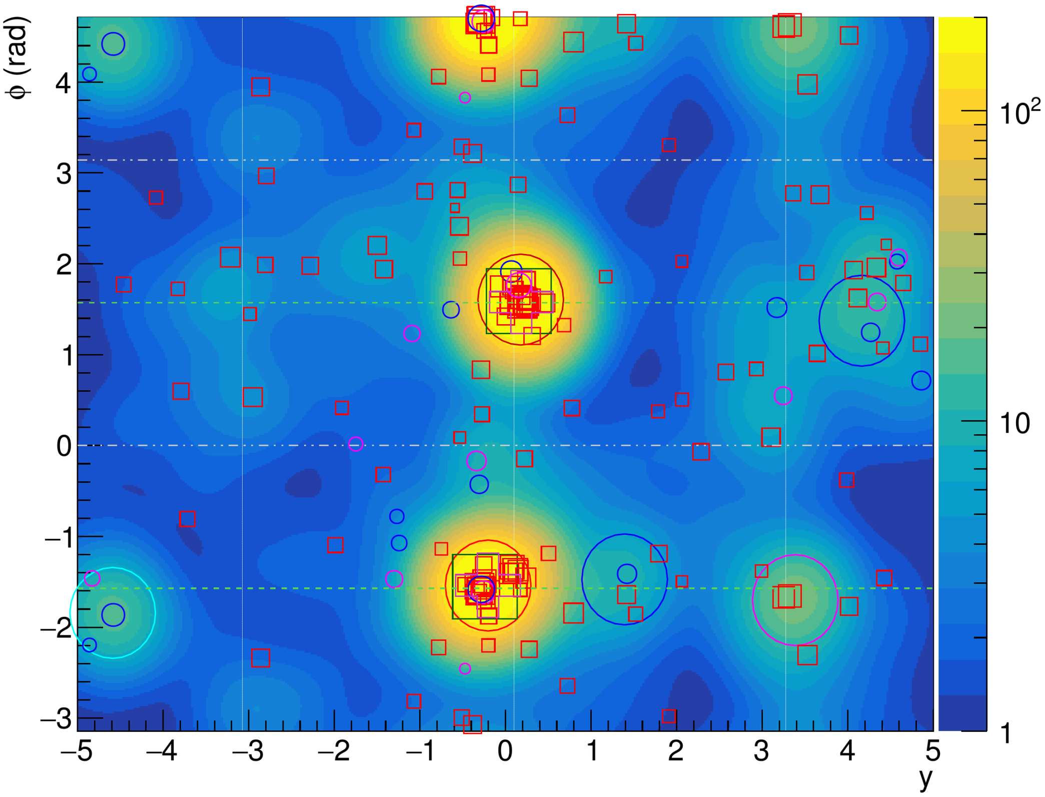


Figure D.1: Event 0.

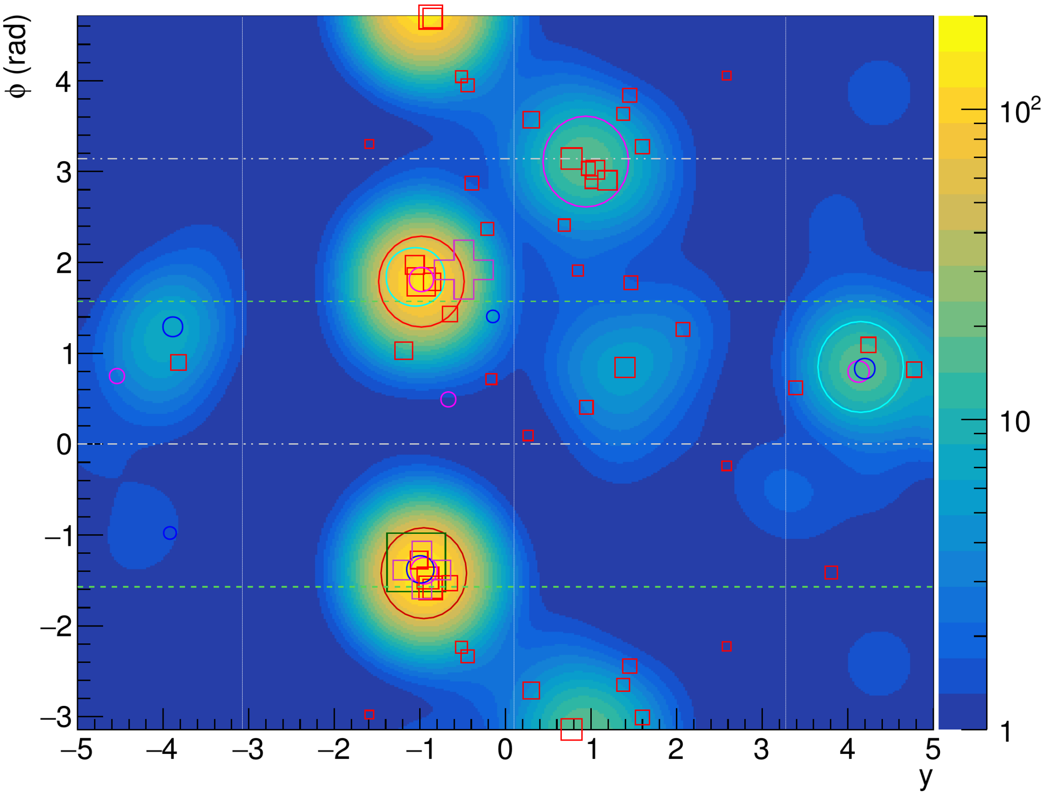


Figure D.2: Event 23.

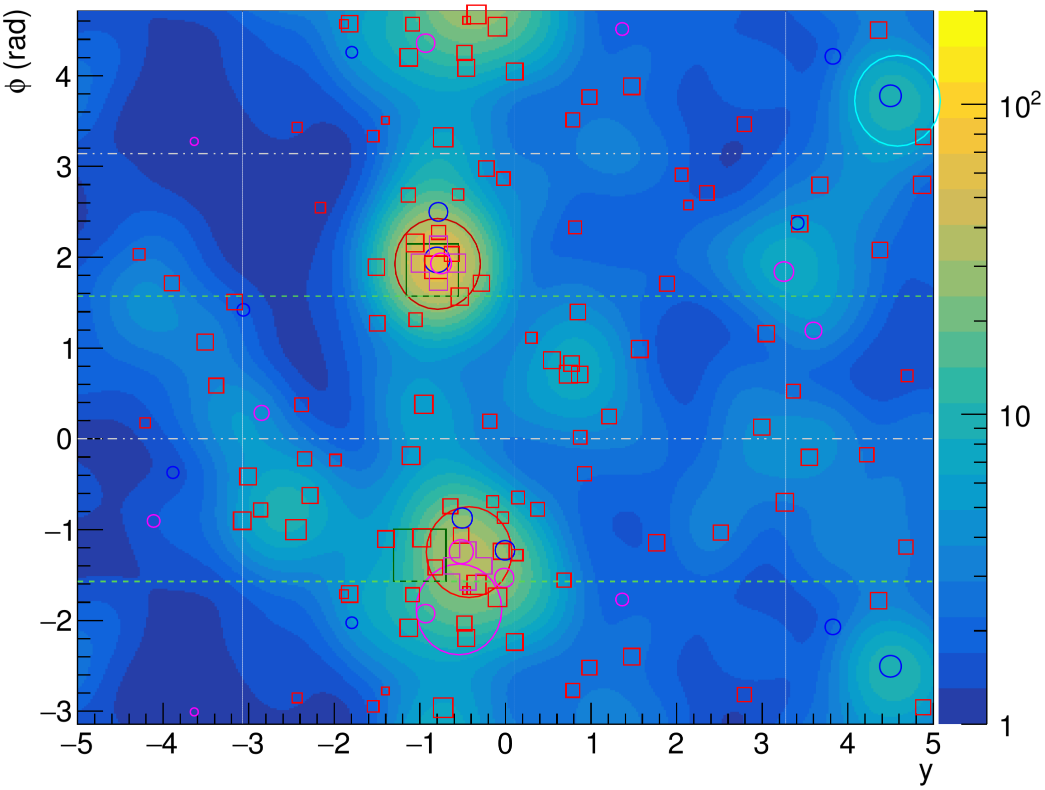


Figure D.3: Event 44.

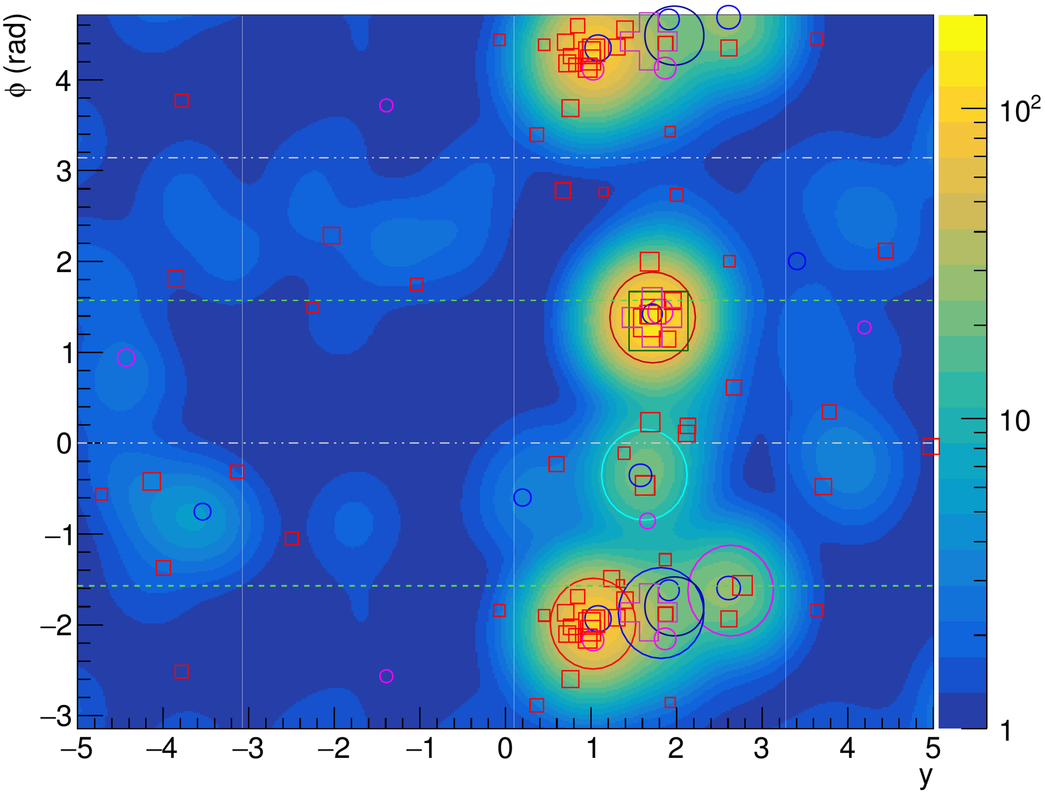


Figure D.4: Event 16.

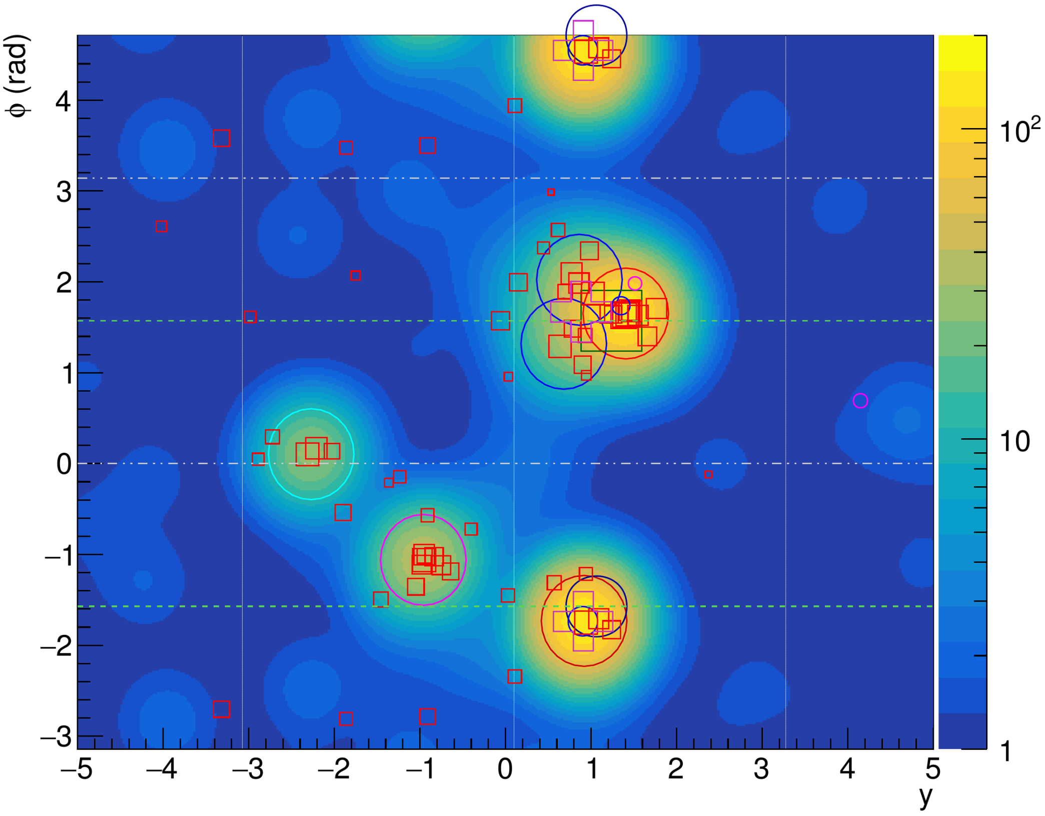


Figure D.5: Event 31.

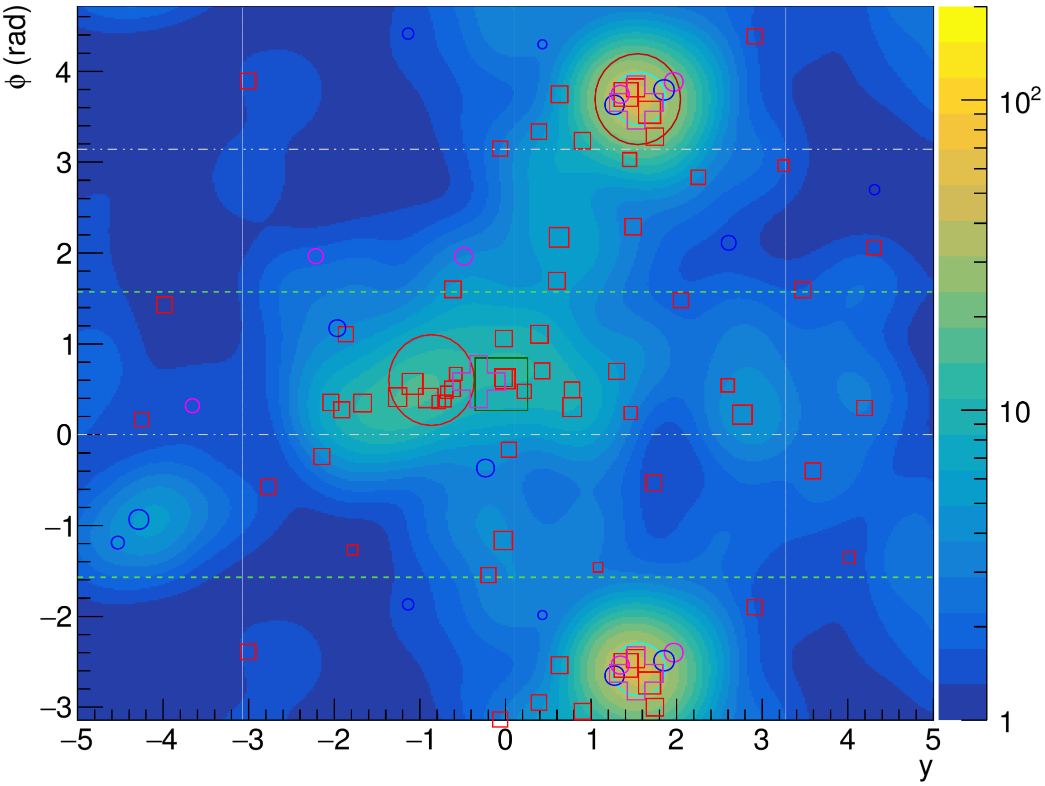


Figure D.6: Event 295.

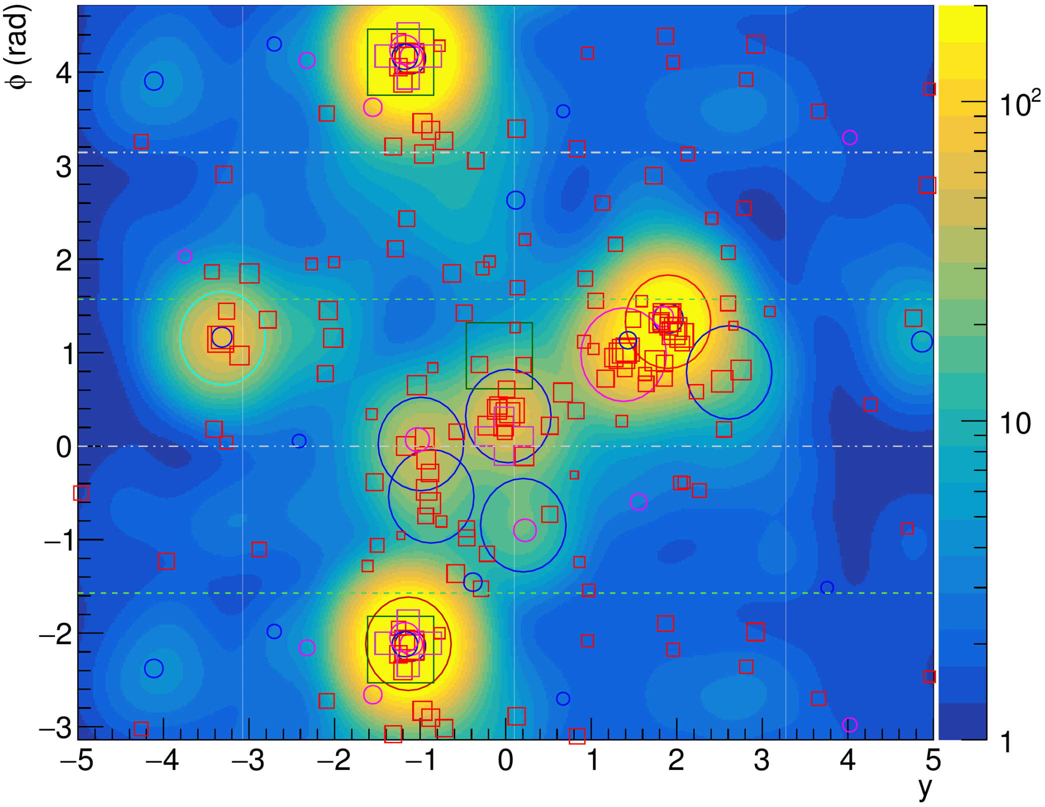


Figure D.7: Event 311.