**Indian Institute of Technology Roorkee**

# Jet Flavor Studies at CMS

Author

Aman Abhishek

Supervisor

Dr. Mikko Voutilainen
Helsinki Institute of Physics
University of Helsinki
Helsinki, Finland

Co-supervisor

Prof. Vir Singh
Department of Physics
Indian Institute of Technology Roorkee
Uttarakhand, India

A thesis submitted in partial fulfillment for the degree of
Master of Science at Indian Institute of Technology Roorkee

April 27, 2015

CERN-THESIS-2018-215

# Declaration of Authorship

April 27, 2015

This is to certify that the master's thesis titled "**Jet Flavor Studies at CMS**", prepared by **Mr. Aman Abhishek**, in partial fulfillment for the award of degree of Master of Science in physics from Indian Institute of Technology Roorkee is a record of his own work carried out under our supervision and guidance.

**Dr. Mikko Voutilainen**
*Supervisor*
Helsinki Institute of Physics
University of Helsinki
Helsinki, Finland

**Prof. Vir Singh**
*Co-supervisor*
Department of Physics
Indian Institute of Technology Roorkee
Uttarakhand, India

*"Understanding physics is like trying to figure out the rules of chess by watching many games. When you thought that you had it all figured out, you would see* ***castling****, and question everything you know. But rather than falling into the abyss of dismay, you would wait more patiently than ever to observe this rare phenomena, and soon it will be a part of your theory, not its exception."*

- analogy given by Richard Feynman, paraphrased and extended here

# *Abstract*

When we collide protons at the LHC, the partons collide and fly out with massive energies. QCD dictates that partons cannot exist alone, and therefore they hadronize instantaneously and lead to a spray of particles, called a jet. Jets are ubiquitous at CMS, and we need to understand them properly to be able to understand the collisions. In this thesis we explore various state of the art algorithms which look at a jet and try to predict which parton it came from. The parent parton's flavor is called the "jet flavor". We implement these algorithms on various samples, compare their results and performance and obtain some benchmark results for future analysis.

# *Acknowledgements*

To state the obvious, this thesis would not have been possible without my supervisor, Mikko. I must thank him for giving me an opportunity to work under him at Helsinki Institute of Physics and exploring the beautiful country. If it were not for his expertise in the field, and his extremely patient and enthusiastic attitude towards my learning, I would not have learnt so much in such a short span of time. His students Hannu and Juska also helped me during my work, and I must acknowledge Hannu specially for his programming file structure that he shared with me, which formed the platform for me to write my code.

I also wish to thank my family and friends, who are always supportive and encouraging through the various ups and downs of my life. And last but not the least I wish to thank Prof. Vir Singh and the Department of Physics, Indian Institute of Technology Roorkee for providing the valuable guidance and allowing me to pursue my master's thesis at Helsinki.

# Contents

# List of Figures

# List of Tables

# Abbreviations and Symbols

| | |
|---|---|
| **LHC** | **L**arge **H**adron **C**ollider |
| **CMS** | **C**ompact **M**uon **S**olenoid |
| **MC** | **M**onte **C**arlo |
| **SM** | **S**tandard **M**odel |
| **PDF** | **P**arton **D**istribution **F**unction |
| **LO** | **L**eading **O**rder |
| **NLO** | **N**ext to **L**eading **O**rder |
| **QGL** | **Q**uark **G**luon **L**ikelihood |
| **QCD** | **Q**uantum **C**hromo **D**ynamics |
| **COM** | **C**enter **O**f **M**ass |

| | |
|---|---|
| $p_T$ | Transverse momentum |
| $\eta$ | Pseudorapidity |
| $\sigma_2$ | Minor axis of the jet cross section in $\eta - \phi$ plane |
| $p_T D$ | Fragmentation function |

*Dedicated to the northern lights, snowstorms, glaciers and huskies; and the wonderful people I met during my stay in Finland and while travelling around.*

# Chapter 1

# Introduction

There was once a time when curious individuals could by themselves cover great lengths to understand the laws of nature. We stood on the shoulders of those giants and saw further, only to discover that nature is as beautiful and full of surprises as it is complex. The scale and complexity of the problems that physicists are trying to solve today necessitate a collective effort. This work is a minuscule contribution to that herculean effort.

CERN (*Conseil Européen pour la Recherche Nucléaire*) stands today as probably the epitome of such an effort and collaboration; transcending all borders of disputes on conflicts in the social arena. Founded in 1954, the CERN is located on the Swiss-French border. It was one of Europe's first joint ventures and now has 21 [2] member states. At CERN physicists are attempting to the fundamental laws of the universe. They have constructed the world's largest, most complex and expensive machine ever built in order to study the basic constituents of matter – the fundamental particles. Inside the LHC, particles smash together close to the speed of light. The process gives the physicists an insight about how the particles interact and thus into the fundamental laws of nature.

Theorists come with their theories and experimentalists design experiments to test them. The instruments used at CERN are particle accelerators and detectors. The accelerators are responsible for boosting the beams of particles to high energies before they are made to collide with each other or with stationary targets. The many detectors, such as CMS observe and record the results of these collisions. The gigantic amount of recorded data is of no utility if one cannot make any sense of out it. Somewhere in the middle of theorists and experimentalists lie data anaylists and phenomenologists - bridging the gap between the two.

The work done in this thesis belongs to that domain. In chapter 2 we explain the relevant theory and how essentially *jets* are our probes in understanding collisions at CMS. In chapter 3 we explain how MC generators have become ubiquitous in the analysis, and its usage in this thesis. In chapter 4 we discuss various state-of-the-art *jet flavor tagging algorithms* that has been proposed by the particle physics community and their implementation. In chapter 5 we compare the results obtained from various jet flavor tagging algorithm and compare their performance. In chapter 6 we discuss the conclusions that can be drawn from the entire analysis.

# Chapter 2

# Theory

## 2.1 The Standard Model

The aim of the Standard Model (SM) of particle physics is to provide the simpler underlying structure to the host of particles found in nature. In some sense SM plays the role that the Periodic Table plays in chemistry. The standard model has stood the test of time and experiments, emerging as a successful theory of particle physics.

Four forces can describe all observed interactions between particles in nature[7]. These fundamental forces are the gravitational force, the electromagnetic force, the weak force and the strong force. Most of the forces that we encounter in our “daily life” can be described by the former two, while the latter two have a very short interaction length and are therefore only observed in nuclear and particle physics. Standard model includes all forces except gravitation, but it still manages to explain collisions at LHC with high accuracy because of the relative weakness of the strength of gravity at the small scale of protons.

| **Force** | **Strength** | **Mediator** | **Theory** |
|---|---|---|---|
| Gravitational | $10^{-42}$ | graviton | Geometrodynamics |
| Weak | $10^{-13}$ | $Z^0$, $W^-$, $W^+$ | Flavordynamics |
| Electromagnetic | $10^{-2}$ | $\gamma$ | Electrodynamics |
| Strong | 10 | gluons | Chromodynamics |

Table 2.1: Comparison of the strength of fundamental forces and their mediators

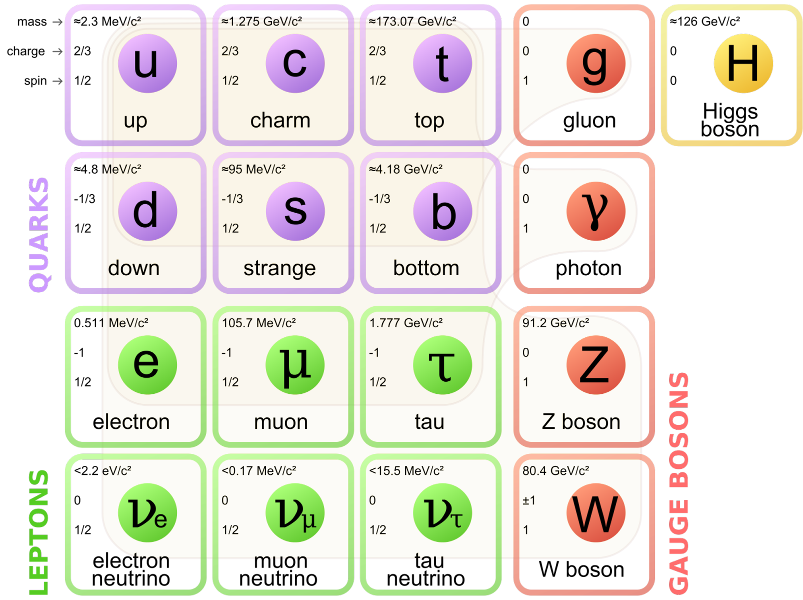



FIGURE 2.1: Standard Model[1]

**Quarks** - Fundamental particles that experience strong interaction are called quarks. They are grouped into three families with two quarks each and carry a fraction of either $2/3$ or $-1/3$ of the elementary charge. The six quarks are distinguished by their flavor, which is a set of quantum numbers that are conserved in all interactions except the weak interaction, where flavor-changing is possible. Quarks have an additional quantum number called *color*. It can have 6 values in total - red, green and blue and their “anticolors”. A combination of three particles containing all three colors(or anticolors), or a color-anticolor combination is color-neutral or “colorless”. All naturally occurring particles in nature are colorless.

**Leptons** - Leptons are fundamental particles that do not experience strong interaction. The electron $e$ and its heavier siblings, the muon $\mu$ and the tau $\tau$ , have neutral partners, the neutrinos $\nu_e$, $\nu_\mu$ and $\nu_\tau$ .

**Gauge bosons** - Interactions in the SM are mediated by gauge bosons. Gauge bosons are the quanta of gauge fields described by the quantum field theory. The electromagnetism is mediated by a single vector boson, the photon. Spontaneous symmetry breaking leads to the three bosons of the weak interaction the $Z^0$, $W^+$ and $W^-$ boson. Eight gluons mediate the strong interaction. The spin-statistics theorem states that all bosons have integer spin whereas all fermions carry a half-integer spin. In contrast to the photon which has no electric charge, gluons carry the charge of the strong interaction called color. Therefore, gluons participate in the strong interaction themselves. Like all fundamental particles with an odd-integer spin, the listed vector bosons mediate forces that are both attractive and repulsive. On the other side, particles with an even-integer spin mediate forces that are either attractive or repulsive. A Higgs boson with spin 0 is responsible for attributing mass to all the other particles. The hypothetical graviton is assumed to be a spin 2 particle.

## 2.2 Hadronization

According to SM, colored particles cannot exist in isolation but always in a combination with other colored particles to form colorless hadrons. This is called "color confinement" which means that the strong force increases on increasing the separation. Another concept called "asymptotic freedom", dictates that quarks and gluons can move in freedom as long as they stay close (the order of the diameter of a nucleon $\sim 10^{-15}$) to the other colored particles that make the configuration colorless.

This implies that the quarks and gluons, also called partons, inside protons are free to collide with partons of the other proton when we do *pp* collisions at LHC. When two partons collide with such massive energies as 14 TeV, they can not fly out freely because of confinement. When such a parton gets high momentum away from the particles that make the package colorless, a new particle - antiparticle pair is created solely from the energy of the strong confining bond, and a new color singlet is formed and is free to exit the system. In high-energy collisions the initial color singlets are often in a virtual state and decay further to hadrons until energies of single hadrons are low enough for forming a real particle. In this process numerous hadrons are created from one initial parton. This process is known as *hadronization.*

Hadronization is a phenomenologically studied concept, and is pictured in the Lund string model. According to this model one may imagine stretching a rubber band, and when the energy becomes large enough it breaks into two rubber bands (figure 2.2). Most of the hadrons formed after hadronization are unstable, and they decay to lighter particles - electrons, protons and photons. At CMS it takes $< 10$ ns for these hadrons to traverse the CMS tracker and calorimeters. Some hadrons like charged pions and kaons live long enough to be measured, but still many of the particles born in the hadronization process undergo decay before even hitting the first layers of the tracker just a few centimeters away from the interaction point.

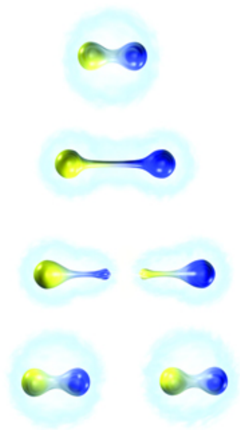


FIGURE 2.2: Lund string model [9]

## 2.3 Some Terminologies

We must introduce some basic terminologies and convention before beginning a discussion on jets.

### 2.3.1 Event

An event encompasses all the phenomena initiated by one proton bunch crossing near the origin of the detector's coordinate system. Normally physicists are only interested in hard scattering, where new physics is found. Duration of an event is in the scale of nanoseconds.

### 2.3.2 Coordinate System

The coordinate system used for the CMS detector is defined as follows -

1. Origin - center of the detector where the collisions take place, coincides with COM frame for easier analysis
2. X-axis - parallel to the radial direction of LHC
3. Y-axis - pointing up, perpendicular to he plan of the LHC
4. Z-axis - parallel to the proton beam

Normally, cylindrical coordinate system variables are used during analysis.

### 2.3.3 Transverse Momentum

The momentum along the X-Y plane is conserved at a collision, and the transverse momentum $p_T$ $(=\sqrt{x^2+y^2})$ is conserved. Note that the momentum of the COM along Z-axis is zero.

### 2.3.4 Pseudorapidity

Pseudorapidity is defined as -

$$\eta = -\ln(\tan(\theta/2)) \tag{2.1}$$

where is $\theta$ the angle from the positive z-axis. It is a measure the scattering angles of particles with respect to the beam axis, and also when speaking of coverage of detectors. The number N of particles flying to one unit in $\eta$ is nearly constant throughout the detector, i.e. $\frac{dN}{d\eta} \approx constant$. Moreover $\eta$ approaches rapidity, $y^{(1)}$, when the particle's speed approaches $c$ and its mass energy is much smaller than its total energy.

## 2.4 Jets and their Clustering

A shower of particles hit the different layers of the detector after a parton hadronizes. A jet is the shower of particles coming from a single parton. What we observe at say, a small part of the detector, is an overlap of particles coming from various jets, or even particles like a photon which might not be part of any jet (if it comes from the hardest subprocess). This means that jets are our handles through which we see the collision process and observe any new physics whatsoever.

The problem is then how does one cluster particles into jets. One must note that there is no unique definition of a jet. The state-of-the-art algorithm used by both CMS and ATLAS collaboration is anti-$k_t$ algorithm [8], and it produces regular shaped conical jets. We use this algorithm in this thesis and we shall define its working briefly. We define $d_{ij}$ as the distance between particles $i$ and $j$, and $d_{iB}$ as the distance between particle $i$ and the beam.

$$d_{ij} = min(k_{ti}^{-2}, k_{tj}^{-2}) \frac{((y_i - y_j) + (\phi_i - \phi_j))^2}{R^2} \tag{2.2}$$

$$d_{iB}^2 = k_{ti}^{-2} \tag{2.3}$$

(1) $y = \frac{1}{2}\ln\frac{E+p_z c}{E-p_z c}$

Here $k_{ti}$, $y_i$, $\phi_i$ are respectively the transverse momentum, rapidity and azimuthal angle of $i^{th}$ particle and R is proportional to the jet size. The clustering process starts by finding the smallest of the distances $d_{ij}$ and $d_{iB}$, and then merging particles $i$ and $j$ if $d_{ij} < d_{iB}$. Iteration proceeds until no particle j is found that fulfills $d_{ij} < d_{iB}$. Then $i$ is considered a complete jet and the algorithm continues to process other particles of the event.

## 2.5 Samples

In the actually observed data as well as the simulated event, we can separate the $pp$ scattering into three categories. The categorization is based on the type of force - strong, weak or electromagnetic. This provides a useful way to do the analysis and make conclusions. These categories are called "samples" in the analysis, and they are generated separately for convenience. The Feynman diagrams for these samples are given in figure 2.3.

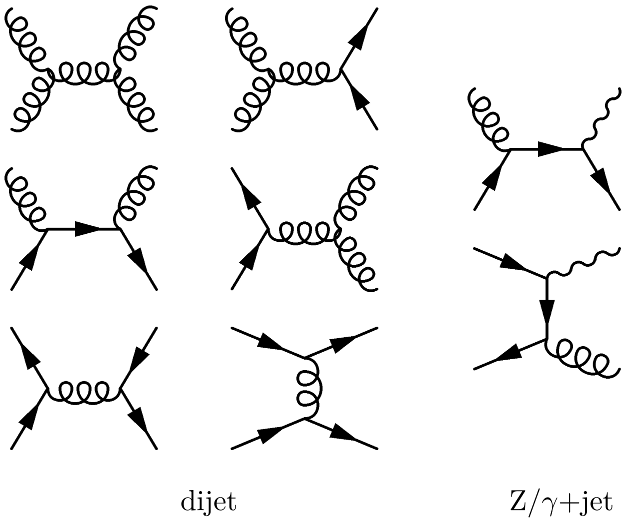



FIGURE 2.3: LO diagrams for samples

The Feynman diagrams for Z+jet and $\gamma$+jet sample are the same, the only difference being (not shown here) is that the $Z^0$ boson decays further, while $\gamma$ does not. In our analysis we close all the decay channels of the $Z^0$ boson except the one leading to decay into a muon pair, $\mu^+\mu^-$. We do this to simplify the analysis, and in the real event too it is easier to pick out $Z^0$ only when it decays via the muonic channel, and ignore the other events. $\mu^+\mu^-$ is an easily identifiable signature of $Z^0$ as compared to other decay products which will get mixed with the same product coming from other sources.

## 2.6 Quark Gluon Likelihood Variables

As known both from first principles and a large collection of experimental measurements, hadronic jets initiated by gluons exhibit differences with respect to jets from light-flavor quarks (u, d, s). In general, the following observations have been made[5] -

- the particle multiplicity is higher in gluon jets than in light-quark jets
- the fragmentation function of gluon jets is considerably softer than that of a quark jet
- gluon jets are less collimated than quark jets

Quark Gluon Likelihood (QGL) variables are defined so that one can differentiate a quark jet from the gluon jet. Three such variables are defined as explained in the following sections.

### 2.6.1 Multiplicity

The simplest and most studied variable that can be used is the multiplicity, i.e. the total number of particles reconstructed within the jet. For a given pT scale, quark jets are expected to have (on average) lower values of multiplicities with respect to gluon jets.

### 2.6.2 $\sigma_2$

Jets have a conical structure that can be projected in the $\eta - \phi$ plane. Gluon hadronization is expected to produce jets which are "wider" than jets induced by quark hadronization. The jet width may be quantified by the two principal components of the second moment of the constituent distribution in the $\eta - \phi$ plane. The shape of a jet can be approximated by an ellipse which is characterized by its two principal axes, the major and the minor axis, and the orientation of the major axis in the plane. A $2 \times 2$ symmetric matrix, $M$, is constructed with the following elements -

$$M_{11} = \sum_i p_{T,i}^2 \Delta\eta_i^2 \tag{2.4}$$

$$M_{12} = \sum_i p_{T,i}^2 \Delta\phi_i^2 \tag{2.5}$$

$$M_{12} = M_{21} = -\sum_i p_{T,i}^2 \Delta\phi_i \Delta\eta_i \tag{2.6}$$

The major ($\sigma_1$) and minor ($\sigma_2$) axes of the jet can be computed from the eigenvalues $\lambda_{1,2}$ of the matrix $M$ by -

$$\sigma_1 = \sqrt{\frac{\lambda_1}{\sum_i p_{T,i}^2}} \tag{2.7}$$

$$\sigma_2 = \sqrt{\frac{\lambda_2}{\sum_i p_{T,i}^2}} \tag{2.8}$$

It has been found observationally that the minor axis $\sigma_2$ is much more sensitive to the differences between quark and gluon jets, and hence it is used as a QGL variable.

### 2.6.3 Fragmentation function

Quarks have a harder fragmentation function compared to gluons and are therefore more likely to produce jets with hard constituents that carry a significant fraction of the jet energy. This can be expressed with the $p_T D$ variable, defined as -

$$p_T D = \frac{\sqrt{\sum_i p_{T,i}^2}}{\sum_i p_{T,i}^2} \tag{2.9}$$

where the sum runs over the jet constituents. From its definition, it stems that $p_T D \rightarrow 1$ for jets made of only one particle that carries all of its momentum, and $p_T D \rightarrow 0$ for a jet made of an infinite number of particles.

# Chapter 3

# Monte Carlo Generators

## 3.1 Overview

Complex final states are formed when elementary particles are collided with very high energies at LHC. Such collisions produce large multiplicities of hadrons, leptons, photons and neutrinos. However, establishing the relation between these final states and the underlying physics description is an extremely nontrivial task, and that is the following reasons -

1. We do not possess a complete understanding of the underlying physics.
2. Any analytical approach is made intractable by the large multiplicities.

Under such a scenario, numerical methods, namely, MC methods, come into play. The complexity of the problem is broken by the divide and conquer strategy - split the problem into smaller, numerically solvable pieces. All main aspects of the events are simulated separately, such as hard-process selection, initial and final-state radiation, beam remnants, fragmentation, decays etc, and put together in one piece, giving a MC simulation of the event. We expect that these simulated events match up to the experimentally observed ones at CERN. Such a setup of MC generator becomes an extremely valuable tool in understanding the laws of physics via the following pathways -

1. We can compare results from the simulation and with results obtained from the experiment and therefore test the theory.
2. We can study the results of the experiments in the future.

For the analysis carried out in this thesis, we have used PYTHIA 8[6].

## 3.2 PYTHIA 8 Setup

PYTHIA is a widely used software package across the whole experimental community of the LHC collaboration, and a benchmark for comparing results. The current PYTHIA generator is the product of more than 35 years of development. In the beginning, all code was written in Fortran 77. PYTHIA 8.100 was the first full release of a complete rewrite to C++.

For our analysis, we need need to simulate the dijet, Z+jet and $\gamma$+jet samples (as described in section 2.5), and PYTHIA provides convenient switches to do so. But first, we list in table 3.1 the switches which remain unchanged across all samples-

| **Switch** | **Value** |
|---|---|
| Beams:eCM | 8000 |
| PhaseSpace:bias2Selection | on |
| PhaseSpace:bias2SelectionPow | 4.5 |

Table 3.1: PYTHIA settings common across samples

The first switch sets the energy of each beam at 8 TeV. The other two are responsible for *biased sampling*. The probability of generating events as a function of $p_T$ falls as $\sim e^{-p_T}$. This means computational time is not getting distributed evenly- oversampling occurs for low $p_T$ and under sampling for high $p_T$. Biased sampling tries to counter this effect by introducing a factor of $p_T^\alpha$, so that

$$P_{event}(p_T) \sim p_T^\alpha e^{-p_T} \tag{3.1}$$

$\alpha = 4.5$ is taken as the standard for analysis done at CMS. In the following sections we discuss sample specific switches.

### 3.2.1 Dijet Sample

We use all hard QCD scattering processes for dijet sample, as shown in table 3.2. Dijet sample needs no other settings.

| **Switch** | **Value** |
|---|---|
| HardQCD:all | on |

Table 3.2: PYTHIA settings for dijet sample

### 3.2.2 Z+jet Sample

We turn on the two processes as shown in figure 2.3. Also, as described in section 2.5, the Z+jet sample needs special care because all non-muonic decay channels of $Z^0$ need need to be turned off (third and fourth line in table 3.3). Furthermore, the production of $W^+$ and $W^-$ bosons also needs to be turned off (fifth line in table 3.3) to not waste computational power and time on events that are not of interest to us.

| **Switch** | **Value** |
|---|---|
| WeakBosonAndParton:qqbar2gmZg | on |
| WeakBosonAndParton:qg2gmZq | on |
| 23:onMode | off |
| 23:7:onMode | on |
| 24:onMode | off |

TABLE 3.3: PYTHIA settings for Z+jet sample

### 3.2.3 $\gamma$+jet Sample

We turn on the processes shown in figure 2.3, there is an extra process $gg \to g\gamma$, which is a box graph.

| **Switch** | **Value** |
|---|---|
| PromptPhoton:qg2qgamma | on |
| PromptPhoton:qbar2ggamma | on |
| PromptPhoton:gg2ggamma | on |

TABLE 3.4: PYTHIA settings for $\gamma$+jet sample

# Chapter 4

# Jet Flavor Tagging Algorithms

Three jet flavor tagging algorithms were implemented and analyzed as a part of this thesis work. Here we describe their working.

## 4.1 Physics Definition

### 4.1.1 Overview

The physics definition matches the outgoing partons from the hardest subprocesses with the leading jets[3]. The matching is based on whether the initial parton is going in the same direction as the jet or not. This algorithm generally gives a larger fraction of unmatched jets as compared to other algorithms. Even though this algorithm was implemented only in PYTHIA results can be expected to be fairly MC generator independent, because MC generators differ in the way they treat hadronization but not the scattering matrix. This definition is used by CMS for flavor tagging. The straightforward approach implies that this algorithm is quite robust as long as the NLO terms in the event do not contribute significantly, or if we can manage to filter them out.

### 4.1.2 Algorithm

1. Cluster particles to obtain $i$ jets for a given event. The particles to be clustered depends on the sample under consideration.

   (a) dijet sample- cluster all final state particles.

   (b) Z+jet sample- cluster all final state particles except the $\mu^+, \mu^-$ pair coming from the hardest subprocess.

   (c) $\gamma$+jet sample- cluster all final state particles except the $\gamma$ coming from the hardest subprocess.

2. Single out the outgoing parton(s) from the hardest subprocess. For a dijet sample, there are two of these whereas for Z+jet and $\gamma$+jet samples there is only one.

3. Check if the event is "good" according to the criterion mentioned in Appendix A. If not, skip event.

4. Loop over $i$ jets. Do flavor tagging-

   (a) Calculate $\Delta R^{(1)}$ between the jet and the parton(s) obtained from step 2.

   (b) If $\Delta R(jet, parton) < R$, assign the flavor of the parton to the jet. In case of dijet sample, it is possible that the jet was already assigned with a flavor by the other parton. In which case, mark the jet as *unmatched*.

   (c) If the jet did not get assigned a flavor from the previous step, mark it as *unmatched*.

## 4.2 QCDaware Definition

### 4.2.1 Overview

The QCDaware definition is currently being used by ATLAS for jet flavor tagging[4]. It has a "top-down" approach - it takes the partons right before hadronization and uses QCD rules to work its way down the "decay tree" to locate the outgoing parton from the hardest subprocess. The way it traverses the tree is implemented and made publicly accessible, and we use that as a plug-in in our algorithm.

### 4.2.2 Algorithm

1. Make a list, $l_1$ of partons that are in preparation of the hadronization process.

2. Cluster particles in $l_1$ using the *QCDaware* algorithm.

3. We obtain jets $j_1, j_2..j_k$, with their flavors as assigned by the QCDaware algorithm. Rescale the four-momentum of these jets by a very small number (the default rescale factor is $10^{-18}$), turning them into *ghost-jets*.

4. Make a list, $l_2$ of final state particles. For Z+jet and $\gamma$+jet sample exclude from this list the $\mu^+, \mu^-$ pair and $\gamma$ coming from the hardest subprocess, respectively.

(1) $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$, where $\Delta\phi \in [0, 2\pi)$

5. Cluster particles from $l_2$ and the $k$ jets from step (3). Use general clustering specifications. The new set of jets obtained $J_1, J_2..J_p$ will be used for the analysis in the subsequent steps.

6. Check if the event is "good" according to the criterion mentioned in Appendix A. If not, skip event.

7. Loop over $p$ jets $J$-
   (a) The flavor of the jet is the flavor of the hardest ghost-jet it contains.
   (b) The jet is tagged *unmatched* if there is no ghost-jet inside it.

## 4.3 Hadronic Definition

### 4.3.1 Overview

The Hadronic definition is also called *Algorithmic definition* in the CMS analysis[3]. This definition is used by default for all b-tagging purposes. Being one of the prominent jet flavor tagging algorithms at CMS, in this thesis we test its efficiency beyond its immediate scope of b-tagging.

### 4.3.2 Algorithm

1. Make a list, $l_1$ of partons that are in preparation of the hadronization process.

2. Add to $l_1$ *b-hadrons*, *c-hadrons* and *s-hadrons* based on the following definition-
   (a) *b-hadron* is defined as a hadron containing at least one $b/\bar{b}$ quark and other lighter quarks.
   (b) *c-hadron* is defined as a hadron containing at least one $c/\bar{c}$ quark and other lighter quarks
   (c) *s-hadron* is defined as a hadron containing at least one $s/\bar{s}$ quark and other lighter quarks

3. Rescale the four-momentum of the particles from $l_1$ by a very small number (the default rescale factor is $10^{-18}$) turning them into so-called *ghost particles*.

4. Make a list, $l_2$ of final state particles. For Z+jet and $\gamma$+jet sample exclude from this list the $\mu^+, \mu^-$ pair and $\gamma$ coming from the hardest subprocess, respectively.

5. Cluster particles to obtain $i$ jets for a given event. The particles that are clustered are taken from $l_1$ and $l_2$.

6. Check if the event is "good" according to the criterion mentioned in Appendix A. If not, skip event.

7. Loop over $i$ jets.

   (a) Jets are assigned flavor based on the (ghost) hadrons-

      i. A jet is considered a $b$ jet if there is at least one $b$ ghost hadron clustered inside it.

      ii. A jet is considered a $c$ jet if there is at least one $c$ and no $b$ ghost hadron clustered inside it.

      iii. A jet is considered a *unmatched* jet if there is no $b$ or $c$ ghost hadron clustered inside it.

   (b) Jets are assigned flavor based on the (ghost) partons-

      i. A jet is given the flavor of the hardest parton inside it.

      ii. A jet is considered *unmatched* otherwise.

   (c) The *unmatched* jets from (a) are assigned flavor obtained from (b).

# Chapter 5

# Results and Discussion

## 5.1 Flavor Fractions

One useful information that a jet flavor tagging algorithm can provide us with is what fraction of jets come from a particular parton as a function of $p_T$. Such information can be plotted as a stack for convenient visualization. The stack plots of flavor fractions obtained from applying the physics definition on the three samples is given in figure 5.1, accompanied with the dominant Feynman diagrams.

### 5.1.1 Insight into Parton Density Distribution

A jet with high $p_T$ is formed from a parton with high $p_T$. A high $p_T$ parton in turn corresponds to a large, $x$[1] in the PDF. Therefore, to find out which are the possible incoming partons in the Feynman diagrams at high $p_T$, one must look at the large $x$ PDF. Similarly, for low $p_T$, PDF corresponding to small $x$ must be referred, and hence there is a linear correspondence between $x$ and $p_T$. Of course, the outgoing partons in the Feynman diagram can be drawn in any way and are not influenced by the PDF. In this sense, the Feynman diagrams that are "unlocked" can be ordered as a function of $p_T$. Note that the aforementioned facts imply that for such an ordering of Feynman diagrams, since the PDF remains the same for various samples, the sequence in which the incoming partons appear must also remain the same. During this discussion, it will be really helpful for the reader to also refer to figures 5.2 and 5.3, which shows the dominant Feynman diagrams explicitly as a function of $p_T$ and the PDF of a proton at 8 TeV collision energy, respectively.

[1] Bjorken x: fraction of hadron's (proton) momentum carried by a parton

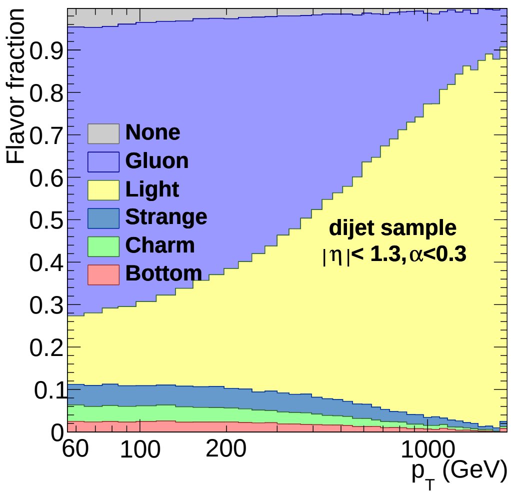
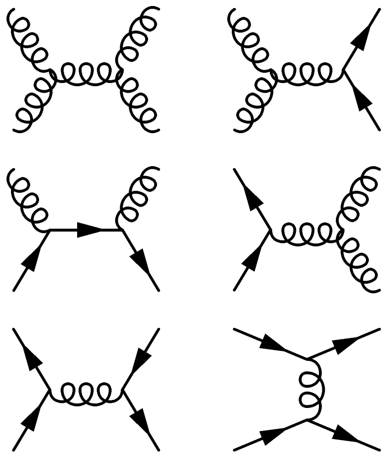
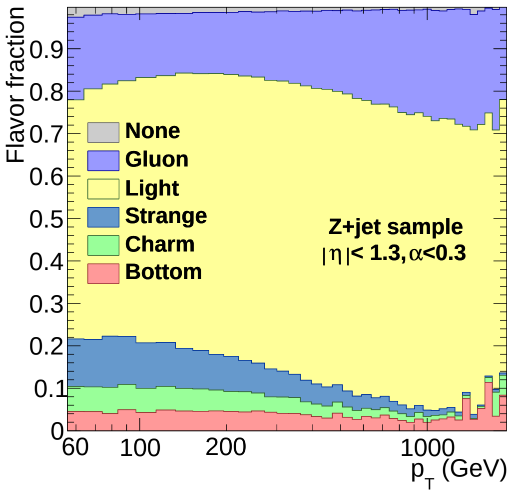
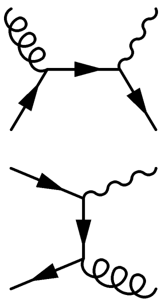
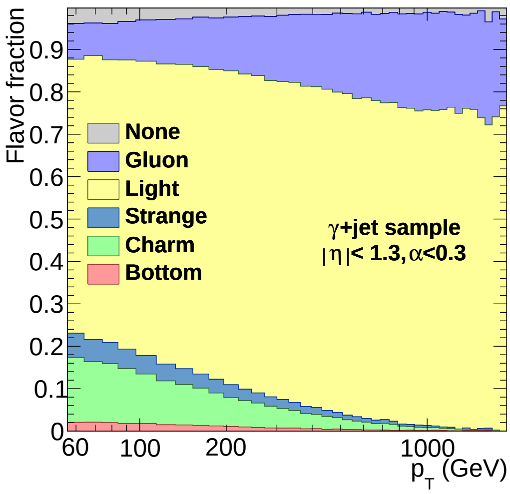
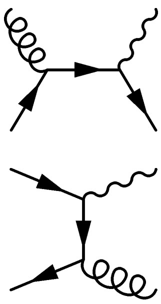
<table>
<tr><th></th><th>Stack Plot</th><th>Feynman Diagrams</th></tr>
<tr><td>dijet</td><td>
</td><td></td></tr>
<tr><td>Z+jet</td><td>
</td><td></td></tr>
<tr><td>γ+jet</td><td>
</td><td></td></tr>
</table>

FIGURE 5.1: Flavor fractions from Physics Definition

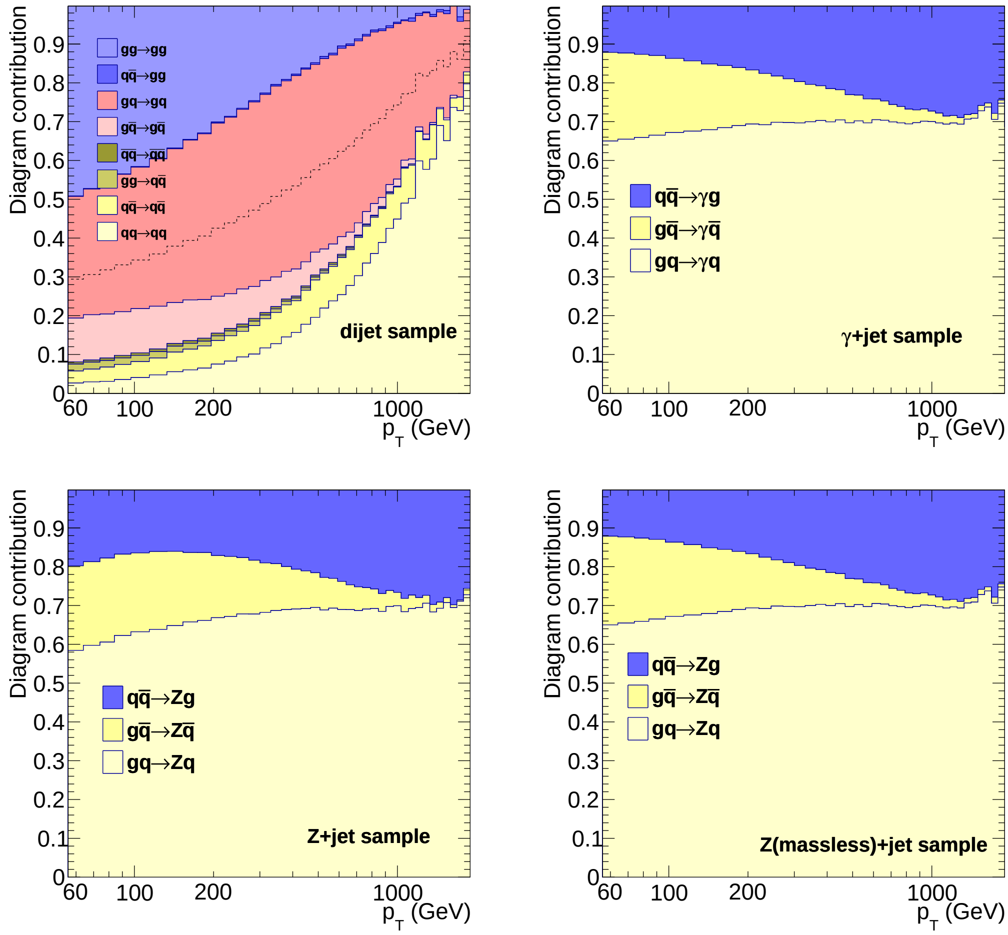



FIGURE 5.2: Variation of the dominance of all the LO Feynman diagrams as a function of $p_T$

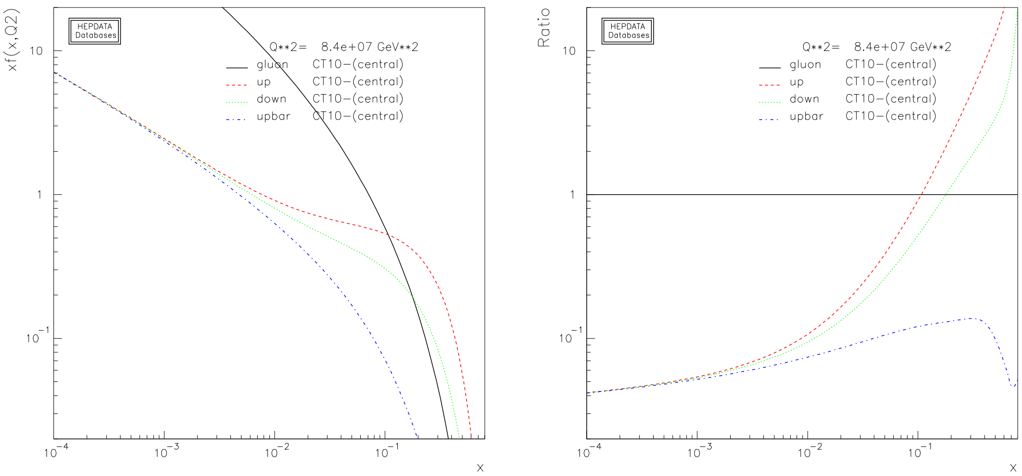



FIGURE 5.3: PDFs of some of the dominant species inside a proton, as a ratio of gluon (right) and as their normal values (left).

**Dijet** - The Feynman diagrams with the highest contribution as a function of $p_T$ are listed from top left to bottom right.

- **Low $p_T$**: Gluons are in relative abundance at $x \sim 10^{-2}$, and the top two diagrams are both dominant, leading to $\sim 29\%$ quark, $\sim 69\%$ gluon and $\sim 2\%$ unmatched jets. The extra 40% gluon jets can be attributed to the stronger coupling of a gluon with gluon than a gluon with quark.
- **Intermediate $p_T$**: The relative abundance of quarks and anti quarks increases at intermediate $x \sim 10^{-1}$, and the two middle and bottom left diagrams offer a significant contribution.
- **High $p_T$**: At $x > 10^{-1}$, abundance of gluons and antiquarks falls, while that of up quarks rises. This makes the bottom right diagram most dominant, suppressing all others. In the asymptotic limit $x \to 1$, only the bottom right diagram remains, in principle giving 100% quark jets on extrapolation beyond $p_T = 2000GeV$.

**Z+jet and $\gamma$+jet** - The Feynman diagrams with the highest contribution as a function of $p_T$ are listed from top to bottom, and they are the same for both the samples, leading to pretty similar stack plots. The upper diagram contributes significantly through all $p_T$ because of lack of other stronger candidates. The bottom diagram needs an antiquark as an incoming parton, which becomes relatively more likely only at high $x$. However, then too it is always a factor of $\sim 5$ less than the gluon abundance. We believe that the fluctuations in the flavor fraction of heavy quarks is due to low statistics. These samples are pure in quarks, with $> 70\%$ jets being quark jets over all $p_T$.

### 5.1.2 Insight into Electroweak Couplings

It is of interest to note that the bands of heavier quarks (b, c and s) in the Z+jet and $\gamma$+jet samples hint at the nature of the coupling constants. While a $\gamma$ couples $\propto e^2$, Z boson does not see electric charge, $e$. This implies that for the $\gamma$+jet sample, the thicker band of c quark jets can be attributed to its higher electric charge 2/3 e as compared to b or s quark $-1/3$ e. From the figure it can be seen that at low $p_T = 60$ GeV, amongst b, s and c jets are around 11%, 22% and 66% respectively, and that distribution is more or less maintained for all $p_T$. For the Z+jet sample the distribution is comparatively more even, and the weight of the quark plays the dominant role in its production, not its electric charge.

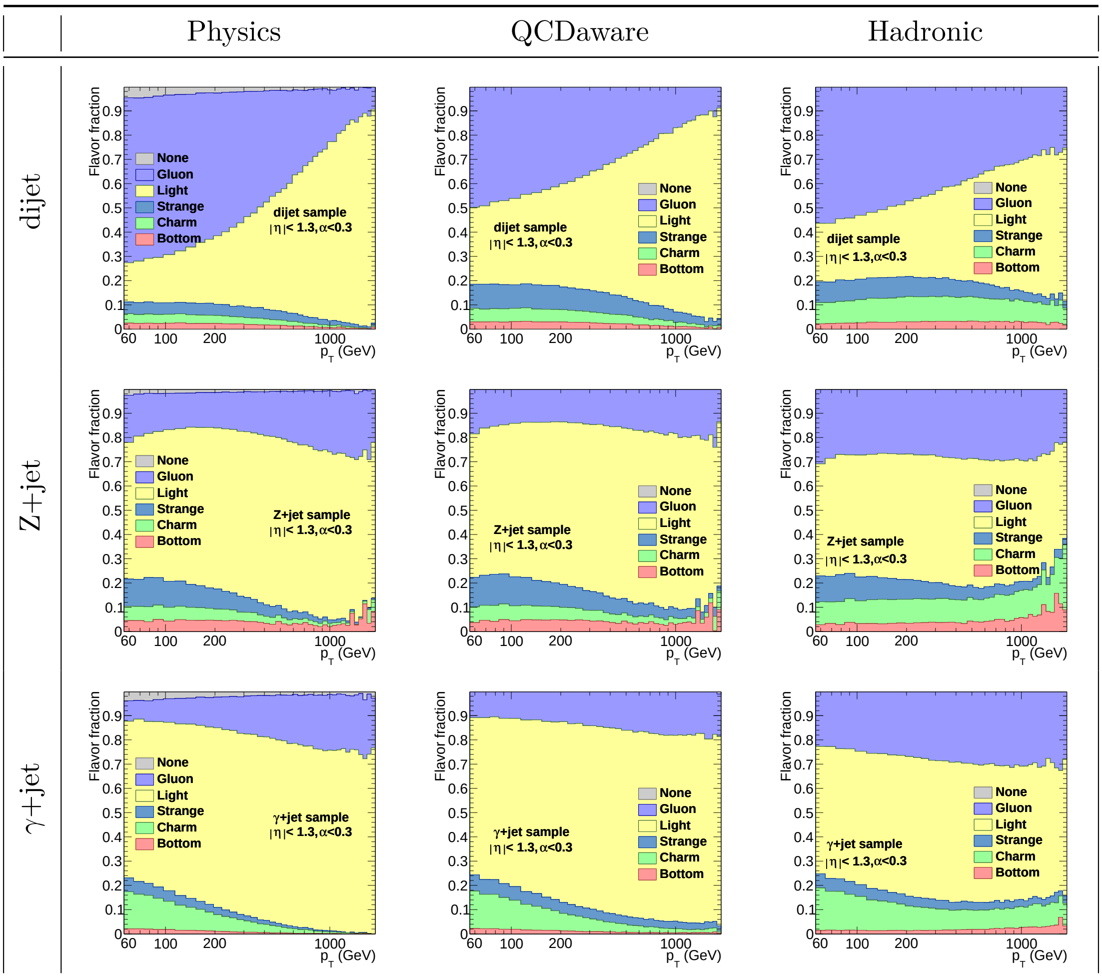



FIGURE 5.4: Flavor fractions from all algorithms

### 5.1.3 Analysis of the Performance of Algorithms

The performance of QCDaware and Hadronic Definition can be measured again in terms of the similarity of their stack plots to the benchmark set by the Physics Definition (Table 5.4). The differences make the following features apparent:

- **Hadronic definition**: The manner in which this definition proceeds can be visualized as stacking the b and c fractions first. The algorithm is made to tag any jet with b(c) hadron as b(c) flavored, making it quite biased towards these heavy flavors in the light of gluon splitting, $g \rightarrow b\bar{b}$ and $g \rightarrow c\bar{c}$ during parton showering. With increasing $p_T$, gluon splitting only increases, leading to an even more abundant production of b and c jets at high $p_T$. By construction, physics definition is blind to gluon splitting.

- **QCDaware definition**: If one quark from a $q\bar{q}$ pair falls within the jet radius of a gluon quark (to which it did not belong), QCDaware will think of this as a hard gluon emission $q \to gq$ and incorrectly label the gluon jet as a quark jet. Such "stray" quarks contaminate the gluon jets. The same can happen if an antiquark falls in a quark jet. QCDaware will consider this as gluon splitting $g \to q\bar{q}$ and label the jet as gluon. However, the probability of this happening than the former, because the antiquark falling in the quark jet in the latter case must also be of the same flavor as the quark to be considered as gluon spitting.

## 5.2 QGL Variables and Robustness

The whole premise and utility of defining quark-gluon likelihood variables is to be able to distinguish a quark jet from a gluon jet. The manner in which say, a quark with a given $p_T$ hadronizes to produce a (quark) jet should not depend on the process which produced it. This fact gives a strong basis to test the robustness of a jet flavor algorithm. At a fixed collision energy if we look at the QGL variables of the quark jet in $[p_T, p_T + \delta p_T]$, then it should look the *same* across samples (processes).

In figure 5.5 we compare the three QGL variables obtained from each of the algorithms for quark and gluon jets, across samples. The full and hollow markers correspond to the quark and gluon jets respectively, while different colors denote different samples. A robust flavor definition will be able to overlay all the hollow markers separately from all the full markers, while maintaining a clear structure for each. It is evident from the from the figures that the physics definition gives the best performance, followed by QCDaware definition and Hadronic definition.

It seems that the leading order (LO) parton is the most decisive in terms of the final jet properties. Hadron definition mixes light flavors due to gluon splitting, $g \to q\bar{q}$ and hard gluon emission, $q \to gq$ when it looks at the hardest parton right before hadronization. QCDaware tries to improve upon this by reconstructing the LO order parton from the final parton shower partons, using the allowed QCD vertices.

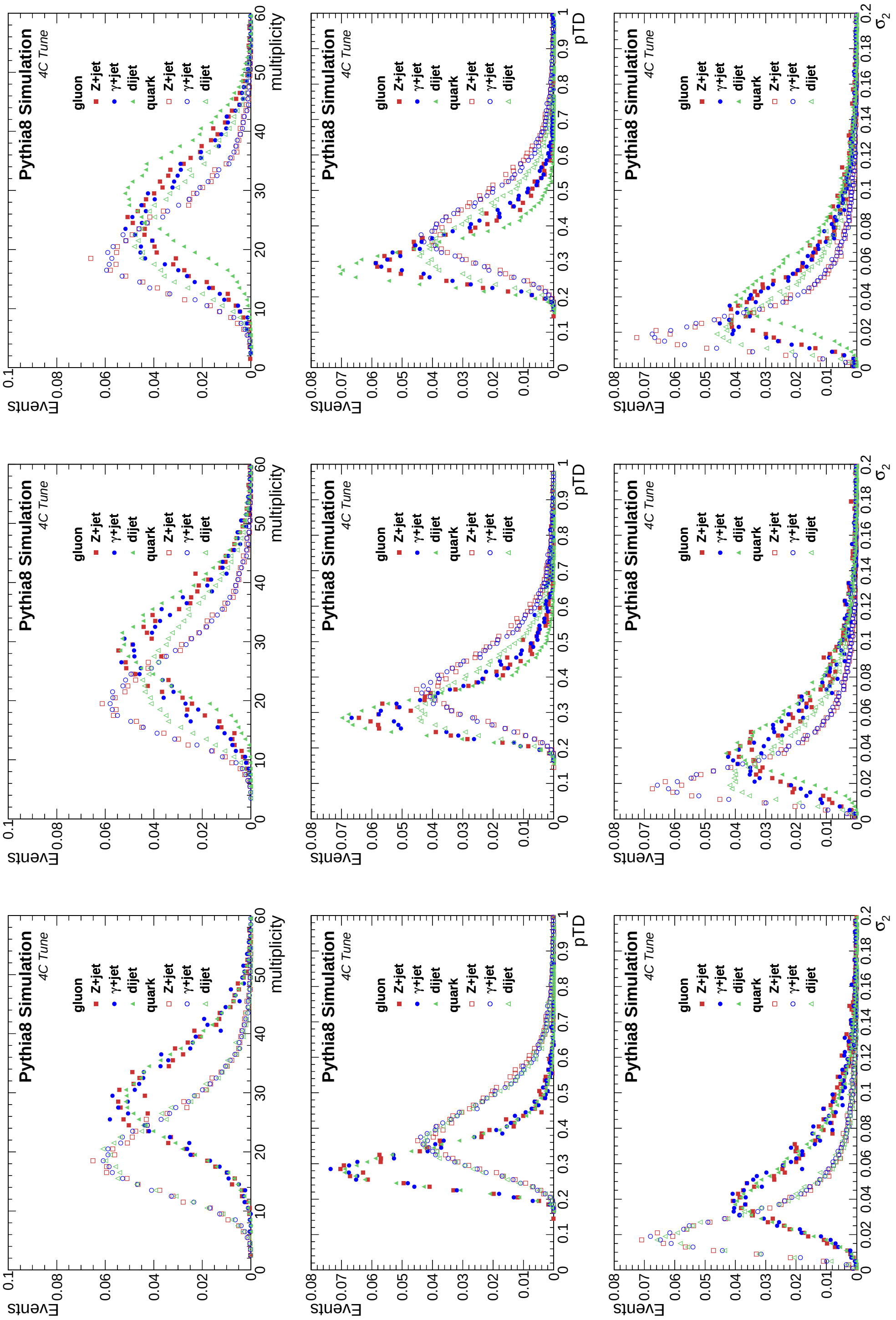



FIGURE 5.5: QGL variables

# Chapter 6

# Conclusion

It is evident from the results that the physics definition outperforms all the other flavor tagging algorithms. The logical basis of physics algorithm is quite straightforward, and hence is robust performance. The question which arises then is why are physicists still trying to come up with more algorithms?

PYTHIA is a LO event generator, but nature of course, is not only restricted to LO. The physics definition too relies only on the LO terms in *pp* scattering. In principle it is possible to use NLO MC generators, and extend the physics definition by matching one more outgoing parton with one more jet. The problem with this approach is obvious - it cannot be extended to the infinite number of Feynman diagrams that exist in reality. An ideal algorithm cannot rely on the "bottom-up" approach used by the physics definition. It should have the theoretical backing to it to be able to work "top-down", for easy implementation, and accommodate as many diagrams as possible.

QCDaware follows a very promising idea, and perhaps it needs to be investigated further why it does not give results same as physics definition. To rigorously test this definition it will be useful to follow the decay chain produced by the MC generator being used and and compare to see where and how QCDaware fails to locate the parton from the hardest subprocess. If QCDaware is tweaked and perfected, it should reproduce the results of the physics definition when LO MC generators like PYTHIA are used, and outperform the LO physics definition and match the NLO physics definition when NLO MC generator is used. Physics definition therefore provides a very useful benchmark that an algorithm needs to attain before extending its scope to NLO and beyond.

It must also be stated that the Hadronic definition is particularly bad at flavor tagging because it is not really meant for the purpose. The focus of Hadronic definition is to look for b quarks, as they indicate the presence of new physics. After b quarks, it looks for c quarks indiscriminately (no feature to accommodate gluon splitting). It is not surprising then that it overestimates the number of heavy quark jets by a huge margin.

The prime result of this thesis work is that the physics definition is the most robust definition when jet flavor tagging is concerned. The benchmarks set by this definition would be useful for future analysis.

# Appendix A

# Event Selection

Depending on the sample, the following criterion is imposed on the event. The purpose of doing this is to select only those events which are necessarily fulfilling LO approximation. It must be noted that such a selection can be applied to experimental data as well because they concern the observables - jets, and leptons and $\gamma$ from the hardest subprocess.

- dijet
  1. $\delta\phi(jet_1, jet_2) > 2.8$
  2. $\alpha = \frac{p_{T_{jet_3}}}{average(p_{T_{jet_1}}, p_{T_{jet_2}})} < 0.3$
- Z+jet
  1. $\delta R(lepton_1, jet_1) > R$, $\delta R(lepton_2, jet_1) > R$
  2. $p_{T_{lepton_1}} > 20\ GeV$, $p_{T_{lepton_2}} > 10\ GeV$
  3. $p_{T_{jet_2}} < 0.3(p_{T_{lepton_1}} + p_{T_{lepton_2}})$
  4. *70 GeV < dimuon invariant mass < 110 GeV*
- $\gamma$+jet
  1. $\delta R(\gamma, jet_1) > R$
  2. $p_{T_{jet_2}} < 0.3 \times p_{T_\gamma}$

# Bibliography


[1] http://en.wikipedia.org/wiki/Standard_Model#/media/File:Standard_Model_of_Elementary_Particles.svg.

[2] http://home.web.cern.ch/.

[3] https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagMCTools.

[4] https://fastjet.hepforge.org/trac/browser/contrib/contribs/QCDAware/trunk/README.

[5] The CMS Collaboration. Cms physics analysis summary. *CERN Document Server*, November 2013.

[6] T. Sjöstrand et al. An introduction to pythia 8.2. October 2014. arXiv:1410.3012 [hep-ph].

[7] Alan D. Martin Francis Halzen. Quarks and leptons: an introductory course in modern particle physics.

[8] Gregory Soyez Matteo Cacciari, Gavin P. Salam. The anti-$k_t$ jet clustering algorithm. *Journal of High Energy Physics*, Issue 04, April 2008.

[9] University of Groeningen. Hadron structure with antiprotons. http://www.rug.nl/kvi/research/hnp/research/panda. Accessed 01-April-2015.