Abstract

The LHC-era with $\sqrt{s} = 7$ TeV allows for a new energy-regime to be accessed. Heavy mass-resonances up to $3.5$ TeV/c$^2$ are in reach.

If such heavy particles decay hadronically to much lighter Standard Model particles such as top, Z or W, the jets of the decay products have a sizeable probability to be merged into a single jet. The result is a boosted jet with substructure.

This diploma thesis deals with the phenomena of boosted jets, algorithms to distinguish substructure in these jets from normal hadronization and methods to further improve searches with boosted jets. The impact of such methods is demonstrated in an example analysis of a $Z' \rightarrow t\bar{t}$-scenario on $2$ fb$^{-1}$ of data.


Diese Diplomarbeit beschäftigt sich mit dem Phänomen von konvergierenden Jets, Algorithmen um Unterstrukturen in diesen Jets von normaler Hadronisation zu unterscheiden und Methoden, Suchen mit geboosteten Jets zu verbessern. Der Fortschritt durch solche Methoden wird mit einer Beispielanalyse eines $Z' \rightarrow t\bar{t}$-Szenarios auf 2 fb$^{-1}$ Daten demonstriert.
Erstgutachter: Professor Dr. Peter Schleper
Zweitgutachter: Juniorprofessor Dr. Christian Sander
Contents

1 Introduction 5

2 Theory of elementary particle physics 7
  2.1 History of Particle Discoveries .......................... 7
  2.1.1 Gauge groups of the standard model .................... 9
  2.1.2 Electroweak unification .............................. 9
  2.2 Extensions: e.g. Z' .................................. 11
  2.3 Detector coordinates ................................. 13

3 Kinematics of boosted objects 14
  3.1 Monte Carlo samples ................................ 14
  3.2 Lorentz Boost ....................................... 16
  3.3 Boosted W ........................................... 18
  3.4 Boosted top ......................................... 24
  3.5 Boosted QCD ......................................... 28

4 The LHC-experiment 29
  4.1 Layout ............................................... 29
  4.2 2010 and 2011-runs ................................ 31

5 The CMS detector 33
  5.1 Silicon Tracker ..................................... 34
    5.1.1 Pixel detector .................................. 34
    5.1.2 Stripdetector .................................. 35
  5.2 Calorimeters ....................................... 36
    5.2.1 Electromagnetic calorimeter ....................... 37
    5.2.2 Hadronic calorimeter ............................ 38
  5.3 Muon system ........................................ 38
  5.4 Particle flow ....................................... 40
  5.5 Factorized jet-corrections .......................... 41

6 Clustering and declustering mechanisms 42
  6.1 Clustering of jets .................................. 42
    6.1.1 Cone and sequential recombination algorithms .. 43
    6.1.2 Sequential algorithms: the FastJet package .... 44
    6.1.3 The anti-kt jet clustering algorithm ............ 45
    6.1.4 Cambridge-Aachen ................................ 47
  6.2 Substructure of jets ................................ 48
  6.3 Illustration of Cambridge-Aachen-clustering ........ 49
  6.4 Declustering ....................................... 56
1 Introduction

The Standard Model (SM) of particle physics holds remarkably for half a century by now, but the Higgs-boson as a fundamental prediction of Electroweak Symmetry Breaking (EWSB) is still missing. Many theories beyond the SM propose very heavy particles, decaying into particles with a mass lighter by an order of magnitude. As an example a new gauge boson called $Z'$ with couplings to SM particles quite similar to the Z is studied in this thesis. The main decay channel considered here is $Z' \rightarrow t\bar{t}$. This poses new challenges for detectors and analysis.

Very high energies occurring in $Z'$-decays mean, in context of special relativity, very high boost for the top quarks. The decay products, in the rest system of the top quark, have a sizeable probability to converge in the laboratory system, due to the boost. In this thesis the magnitude of energy needed up to three partons to be merged into one jet are appraised.

The number of jets in the detector is an important distinguishing variable from QCD background that preferentially produces dijet topologies. For merged jets this important variable is lost for searches with conventional methods. But substructure in single jets can be resolved by using sequential recombination jet clustering algorithms in reverse order. Several methods to resolve substructure to “subjets” are discussed. One method utilising the Cambridge-Aachen jet clustering algorithm is chosen and it is graphically demonstrated step for step how the clustering works and in which way this can be used for resolving substructuring with a procedure called “declustering”.

The properties of the so gained subjets are studied on monte carlo, matching, response, quality of mass-reconstruction and behaviour for very highly boosted topologies are thoroughly investigated.

Examples of several CMS analyses already using so called “substructure” are mentioned and finally a search for a $Z'$ with 3 TeV/$c^2$ is performed with conventional and substructure methods in comparison with 2 fb$^{-1}$ of data provided by the CMS experiment. The possible improvement by applying substructure methods is demonstrated and limits on $Z'$ with multiple TeV/$c^2$ mass are set.

An example for a merged jet is shown in Fig. 1 for a highly boosted top in simulation.
Figure 1: A single highly boosted hadronically decaying top out of a Z’-decay with $M_{Z'} = 3$ TeV/c$^2$. The binning in $\eta$-$\phi$-coordinates is arbitrary.
2 Theory of elementary particle physics

This chapter is an overview of the Standard Model (SM) of particle physics, containing a short overview of the hitherto found particles and forces, electroweak unification and symmetry breaking, the hierarchy problem and the Higgs mechanism.

In addition, several theoretical extensions to the SM will be mentioned that produce Z-boson like particles with higher masses, called \( Z' \).

To quantify the likeliness of any interaction between particles to occur, the cross section \( \sigma \) as defined by Fermi’s golden rule is used, stating that it is proportional to the matrix element \( |\mathcal{M}| \) and the phasespace. The differential cross section for an unambiguous two-body scattering into an area of angles in space, \( d\Omega \) is [1]:

\[
\frac{d\sigma}{d\Omega} = \frac{\hbar c}{8\pi} \frac{|\mathcal{M}|^2}{(E_1 + E_2)^2} \frac{p_f}{p_i} \tag{1}
\]

The matrix element describes the fundamental physics of any interaction and can be calculated by application of the Feynman rules, given a Feynman diagram of any interaction.

The factor \( \frac{1}{(E_1 + E_2)^2} \frac{p_f}{p_i} \) contains the available energy for the process \( (E_1 + E_2) \) and the ratio of either ingoing particle momenta \( p_i \) and either outgoing particle momenta \( p_f \). The more energy is available relative to the needed energy for the process, the larger the phasespace will be and the more likely the interaction described becomes.

Experimentally the cross section is determined using the event rate \( \frac{dN}{d\Omega} \) and the luminosity \( \mathcal{L} \):

\[
\frac{d\sigma}{d\Omega} = \frac{dN}{\mathcal{L} \cdot d\Omega} \tag{2}
\]

2.1 History of Particle Discoveries

At the end of the 19th and the beginning of the 20th century, the first elementary particles were found.

In 1897 J.J. Thomson discovered the electron [2]. Two decades later the positive charges in atomic nuclei were discovered by Rutherford.

The neutrons followed closely, discovered by Chadwick in 1932 [3]. Finally even the first antiparticle, the positron, was discovered by Anderson [4] in 1933.

By deep inelastic scattering the first quarks were found 1968 at the Stanford Linear ACcelerator (SLAC). Up, down and strange quark were indirectly
found at this point. Further on, the charm quark was discovered at SLAC and Brookhaven simultaneously [5][6].

With the bottom quark the third generation was seen first at Fermilab in 1977 [7]. The top quark took another 18 years to be found, again by Fermilab [8][9].

An overview of all discovered particles contained in the SM is shown in Fig. 2.

<table>
<thead>
<tr>
<th>Quarks</th>
<th>mass</th>
<th>charge</th>
<th>spin</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2.4 MeV</td>
<td>$\frac{2}{3}$</td>
<td>$\frac{1}{2}$</td>
<td>up</td>
</tr>
<tr>
<td>II</td>
<td>1.27 GeV</td>
<td>$\frac{2}{3}$</td>
<td>$\frac{1}{2}$</td>
<td>charm</td>
</tr>
<tr>
<td>III</td>
<td>171.2 GeV</td>
<td>$\frac{2}{3}$</td>
<td>$\frac{1}{2}$</td>
<td>top</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>photon</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>-1/3</td>
<td>1/2</td>
<td>down</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>-1/3</td>
<td>1/2</td>
<td>strange</td>
</tr>
<tr>
<td></td>
<td>4.2 GeV</td>
<td>-1/3</td>
<td>1/2</td>
<td>bottom</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>gluon</td>
</tr>
<tr>
<td>Leptons</td>
<td>&lt;2.2 eV</td>
<td>0</td>
<td>1/2</td>
<td>$\nu_e$</td>
</tr>
<tr>
<td></td>
<td>&lt;0.17 MeV</td>
<td>0</td>
<td>1/2</td>
<td>$\nu_\mu$</td>
</tr>
<tr>
<td></td>
<td>&lt;15.5 MeV</td>
<td>0</td>
<td>1/2</td>
<td>$\nu_\tau$</td>
</tr>
<tr>
<td></td>
<td>91.2 GeV</td>
<td>0</td>
<td>1</td>
<td>$Z^0$</td>
</tr>
<tr>
<td></td>
<td>0.511 MeV</td>
<td>-1</td>
<td>1/2</td>
<td>$e^-$</td>
</tr>
<tr>
<td></td>
<td>105.7 MeV</td>
<td>-1</td>
<td>1/2</td>
<td>$\mu^-$</td>
</tr>
<tr>
<td></td>
<td>1.777 GeV</td>
<td>-1</td>
<td>1/2</td>
<td>$\tau^-$</td>
</tr>
<tr>
<td></td>
<td>80.4 GeV</td>
<td>+1</td>
<td>1</td>
<td>$W^+$</td>
</tr>
</tbody>
</table>

Figure 2: The particle zoo of the SM [10] excluding the hypothetic Higgs boson
2.1.1 Gauge groups of the standard model

Matter itself is fermionic in nature. Pauli’s exclusion principle applies, which leads to the wealth of chemical phenomena experienced each day.

The three forces contained in the SM, electromagnetic (EM), weak and strong force, are represented by respective bosonic force carriers. This is a clear distinction that leads to completely different phenomena. The number of force carriers or bosons is determined by the underlying mathematic symmetry group in the standard model. In an \( n \)-dimensional symmetry group with \( r \) restrictions, the number of force carriers \( N \) is \( N = n^2 - r \).

EM force is represented by \( U(1)_{EM} \), where “U” means unitary. It is one dimensional without any additional restrictions, therefore carries one force carrier, the photon. The main property of \( U(1)_{EM} \) is the masslessness of the force-carrier, which results in an infinite range for the interaction. \( U(1)_{EM} \) couples to electric charge.

The weak force is formulated in \( SU(2) \), where “S” means that the determinant has to be 1. This is an additional restriction \( r = 1 \). The dimensionality \( n \) is two. The resulting three force carriers are the \( W^+ \), \( W^- \) and \( Z^0 \). The interaction of the weak force is not weak, but very limited in range due to the very high masses of the force carriers. The weak force only couples to the hypercharge of left-handed particles. It is the only force capable of changing the quark and lepton flavour. This can only happen by charged currents.

The last interaction is the strong force of \( SU(3) \) with eight distinct generators and eight fields, the gluons. All gluons carry colour-charge. There are eight different colour-states corresponding to the mathematic generators. For the strong interaction, only bound states of two or three quarks are observed.

If a state bound by the strong force is pulled apart, a pair of particles will be generated as a new colourless state, which needs energy. Pulling a bound state apart equals generating new particles exponentially and therefore requires a lot of energy.

The phenomenon of no free quarks is called asymptotic freedom, effectively limiting the range of the strong interaction, while the process of generating new particles this way is called hadronization.

2.1.2 Elektroweak unification

The EM and weak forces can be unified to a \( U(1) \times SU(2) \)-group that contains a weak isotriplet of vector bosons \( \vec{W} \) containing a negatively, a positively and a neutrally charged vector boson, and a neutral isosinglet \( B \).

The neutral components mix, creating a massless state, the photon, and
an orthogonal massive state, the $Z^0$. The introduction of massive bosons in the electroweak theory comes at the cost of an additional massive scalar, the “Higgs”, named after its inventor, Peter Higgs.

The Higgs particle is the easiest way to introduce mass in the gauge theory, but has not yet been seen by any experiment. Its existence is purely speculative, so far.
2.2 Extensions: e.g. Z’

In several extensions to the SM, very heavy particles associated with a new
gauge group occur. The behaviour of these new bosons is predicted to be
similar to the SM Z bosons. The masses, if these new bosons exist, could be
close to or orders of magnitude heavier than the SM Z boson.

A list of examples for extensions supporting a Z’-particle has been taken
from [11].

- E6-group models [13]
- a new SU(2)R [16]
- several “Little Higgs”-scenarios
- sequential Standard models
- extra dimensions (Kaluza Klein models)
- models with SU(3)c x SU(3)L x U(1)N-gauge symmetries
- models with strong dynamics like topcolour and Breaking Electroweak
  Symmetry Strongly (BESS)

Leptophobic new gauge bosons with otherwise Z0-like couplings can e.g. be
colour singlets in a technicolour scenario. This is called topcolour [12].

The main production of a Z’ in such a scenario will take place by direct
q¯q-annihilation. At the LHC as a pp-collider, this is strongly suppressed by
the Parton Density Function (PDF).

The shape of the PDFs and the high mass of the resonance accumulate
to a shift in the measured position of the resonance with respect to its true
mass value to lower masses. The effect gets stronger the higher the mass of
the Z’ is, leading to a dual peak structure as a result of folding a steeply
falling PDF with a peak, where the second peak resembles a mass edge in
close to the true Z’ mass.

As an example for M_{Z’} = 3 TeV/c² decaying to tops, the generated M_{Z’}
(see chapter 3.1 for details regarding the monte carlo samples) of an inclusive
sample is shown in Fig. 3:
Figure 3: Generated $M_{Z'}$ for a $Z'$-resonance at 3 TeV/$c^2$ with a width of 30 GeV/$c^2$ is shown. The sample is inclusive, containing leptonic and hadronic top decays.

The transverse momenta distributions of the tops and W bosons out of the same sample are shown in Fig. 4 and Fig. 5:

Figure 4: Generated $p_{T,t}$ for a $Z'$-resonance at 3 TeV/$c^2$ with a width of 30 GeV/$c^2$ is shown. Only tops in range of $\pm 10\%$ around 172.5 GeV/$c^2$ are considered.
Figure 5: Generated $p_{T,W}$ for a $Z'$-resonance at 3 TeV/$c^2$ with a width of 30 GeV/$c^2$ is shown. Only W bosons in range of $\pm$ 10% around 80.4 GeV/$c^2$ are considered.

### 2.3 Detector coordinates

All information regarding the detector coordinate system is extracted from [25]

The origin is centered in the middle of the detector and beampipe. The y-axis is pointing straight upward and perpendicular to the beampipe. The z-axis is pointing radially inward toward the center of the LHC.

From the z-axis a polar angle $\theta$ is measured. The Lorentz-invariant pseudorapidity $\eta = -\ln\tan\frac{\theta}{2}$ is used together with the azimuthal angle $\phi$ measured from the x-axis as the detector coordinate system. Angular distance $dR$ is defined as:

$$dR = \sqrt{\phi^2 + \eta^2}$$  \hspace{1cm} (3)
3 Kinematics of boosted objects

In case of a particle that decays has high momentum the Lorentz boost into the laboratory system lets all decay-products apparently converge in angle.

The basic kinematic variables we are interested in are therefore the masses of the former particle and the decay products, the distance between the decay products and the energy available in the initial state.

This chapter will first deal with a very basic model of boost, derived from special relativity, then advance to examples of boosted particles and basic phenomenology expected in the context of boosts.

The aim is to quantify the energies needed for the production of substructure by W bosons and tops for the most common jet radius of $dR = 0.5$ used at CMS [23].

3.1 Monte Carlo samples

The nomenclature for the $Z'$-samples is /1_2_3-4/5_6/7:

1. Physical scenario generated, here Zprime for $Z'$ scenario.
2. The mass of the generated resonance in GeV/$c^2$.
3. The width of the generated resonance in GeV/$c^2$.
4. The generator used, here madgraph [22].
5. The season of processing and pileup simulation, “Fall10” for fall 2010 and “Summer11-PU-S4” for summer 2011 with pileup simulated.
6. The CMSSW-tag used for generating the sample, specifying the software version used. Either “START38_V12-v1” for CMSSW 3.8 or “START4_V11-v2” for CMSSW 4.2 were used.
7. The data format of the sample, either “GEN-SIM-RECO” or “AOD-SIM”.

For Chapter 7.1.5 and 9.1.6 a sample generated in fall 2010 (“Fall10”) without simulated pileup was used:

- /Zprime_M3000GeV_W30GeV-madgraph/Fall10-START38_V12-v1/GEN-SIM-RECO

has been used.

For signal more recent samples reproduced in summer 2011 (“Summer11”) with simulated pileup were used:
As background the sample

- QCD_Pt-15to3000_TuneZ2_Flat_7TeV_pythia6/Summer11-PU-S3_START42_V11-v1/AODSIM

has been used. It contains QCD generated in flat $p_T$-bins from 15 to 3000 GeV/c with the pythia6-generator and the tune "Z2", simulated in summer 2011 with pileup in the CMSSW 4.2 environment. As it is generated flat in $p_T$ a proper reweighting is done to arrive at the shape of the theoretical cross section.
3.2 Lorentz Boost

A particle decaying into two particles is considered. Ignoring hadronization the decay particles $A$ and $B$ shall have opposite directions and only transverse momenta:

\[ p_{\perp, A} = -p_{\perp, B} \]
\[ p_{\parallel, A} = p_{\parallel, B} = 0 \]

The boost of a particle of velocity $v$, mass $M$ and energy $E$ is given by the factor $\gamma$:

\[ \gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{E}{M} \]

The Lorentz transformation will be translated into a boost of the parallel ($p_{\parallel}$) and transversal component ($p_{\perp}$):

\[ p'_{\parallel} = \gamma (p_{\parallel} - v \cdot E) \]
\[ r'_{\perp} = r_{\perp} \]

A Lorentz-boost in z-direction transforms the angle of the particle, $\alpha$, to the $z$-axis from 90 degrees to a new value:

\[ \alpha = \arctan \frac{p'_{\perp}}{p'_{\parallel}} = (\gamma - E \sqrt{1 - \gamma^{-2}})^{-1} \]

For $E \gg M$: $E \sqrt{1 - \gamma^{-2}} \rightarrow E$ an approximation for the square root is possible:

\[ \lim_{E \gg M} \alpha \propto \arctan \frac{M}{E - M} \]

The basic effect is a convergence of the observed particles in the laboratory system. This effect is shown in Fig. 6.
Figure 6: The effect of eq. 9 reformulated as a function of $\alpha$ is shown for a W boson with a mass $M_W = 80.4$ GeV/c$^2$.

In the sense of elementary particle physics and jets, this behaviour leads to the merging of the decay products of boosted particles. In the following, the W boson and the top quark each will be important examples for this behaviour.
3.3 Boosted W

The W is a vector boson that can decay into jets. At the LHC this is the particle with the lowest mass that can, if boosted, produce one merged jet with substructure.

For substructure to be produced, two kinematic events need to happen in succession:

1. The decaying particle, e.g. the W boson, has an energy several times its mass in the laboratory frame.

2. The difference in mass between decaying particle and decay products must be large enough to separate the resulting jets.

The W boson is the lowest mass particle that fulfills both conditions with respect to the energy-scales of possible new physics.

This process has to be understood in detail to distinguish fake substructure from true substructure and understand why more massive particles like tops behave the same way.

The basic kinematic behaviour of the W boson and its decay-products $q\bar{q}$ can be approximated by the formula: \[ M_W^2 \approx p_{T,q} \cdot p_{T,\bar{q}} \cdot dR_{q\bar{q}}^2 \] (11)

The pseudorapidity can be rephrased as:

\[ \eta = \frac{1}{2} \ln \frac{|\vec{p}| + p_L}{|\vec{p}| - p_L} \] (12)

In comparison the rapidity is defined as:

\[ y = \frac{1}{2} \ln \frac{E + p_L}{E - p_L} \] (13)

Differences in rapidity are Lorentz invariant.

\[ \Delta y = \Delta y' \] (14)

Rapidity and pseudorapidity become equal in the limit $E \gg M$ and therefore for high energies the angular distance $dR$ is invariant with respect to longitudinal boost.

Transverse momentum components, particle masses and angular distance are the necessary variables for a complete description of the boosted system. According to eq. (11), their correlation has to be:

\[ \frac{[M]}{[dR]} \approx [p_T] \] (15)
Transverse momenta can be expressed by masses and angles between the particle decay products since the angular distance incorporates the longitudinal impulse components. As long as the masses are well defined and the energy of the boosted particle is far greater than its mass, this relation holds. Offshell processes and low energy processes cannot be covered by this approximation.

Several jets merged into one jet will be called a “fatjet”. The angular distance between the jets is the basic variable that decides if two jets or a fatjet can be seen. The dependence of this variable on the transverse boost has to be completely understood to make predictions for physics scenarios if boosted structures will occur.

If one fatjet out of a boosted hadronically decaying W with known transverse impulse is considered, out of the variables of eq. (11) three are unknown: the transverse impulses of the two decay-particles and their angular distance.

Assumptions on correlation to the known variables can be made: transverse impulse $p_T, q^+ + p_T, \bar{q}$ as the fatjet’s transverse impulse, particle mass as the W-mass and the transverse impulses of the decay products shall be considered equal. The equality of the decay products’ transverse impulses is the limit of a decay where the decay products move orthogonal to the boost in the rest system of the W.

In the rest system of the W the decay is isotropic, the mass difference of the W and its decay products lead in every case (besides the decay parallel to the beam axis) to new transverse impulse components that have to be taken into account: $p_{T, mass}$.

Figure 7: Illustration of the transverse separation impulse $p_{T, mass}$ for the W boson. The difference between the boost impulse $p_{T, W}$ and the separation impulse is proportional to the separation distance $dR_{q\bar{q}}$ in the detector.
Figure 7 demonstrates the transverse separation impulse. The difference between transverse boost impulse out of the process that produced the W and the transverse separation impulse that is inherent to the decay of the W is taken to be proportional to the final angular distance separation seen in the detector.

The individual transverse momentum components of the decay products, \( p_{T,q} \) and \( p_{T,\bar{q}} \), can be substituted by generalized variables:

\[
p_{T,q} \propto p_{T,W} - p_{T,mass} \quad (16)
\]

Equation (16) states, that the difference between the transverse impulse boosting the system, \( p_{T,W} \), and the transverse impulse separating the decay products, \( p_{T,mass} \), is proportional to the individual transverse boost of each particle, \( p_{T,q} \) and \( p_{T,\bar{q}} \).

The basic assumption is, that the two conditions named for the production of substructure at the beginning of the chapter can be divided into the two separate impulses, \( p_{T,W} \) and \( p_{T,mass} \).

Basically a necessary amount of energy to boost a system into a certain angular distance is to be computed, but \( p_{T,mass} \) is still unknown. A few definitions at this point, further explanations will follow once the variables are used:

1. \( M_{part} \): the mass of the decaying particle, in this chapter a W boson.
2. \( M_q \) and \( M_{\bar{q}} \): the masses of the decay-products.
3. \( M_{dp} = M_q + M_{\bar{q}} \): the sum of the masses of the decay-products.
4. \( M_{dim} = M_{part} - \frac{M_{dp}}{2} \): the diminished mass.

Using eq. \((15)\) it is possible to express the transverse separation impulse by the difference in invariant mass between W, respective decay product q and the final angular distance \( dR_{q\bar{q}} \):

\[
p_{T,mass} \equiv \frac{M_W - M_q}{dR_{q\bar{q}}} \quad (17)
\]

Substituting \( p_{T,q} \) and \( p_{T,\bar{q}} \) in eq. \((11)\) with eq. \((16)\) and further expressing \( p_{T,mass} \) with eq. \((17)\):

\[
\frac{M_W^2}{dR^2} = (p_{T,W} - \frac{M_W - M_q}{dR_{q\bar{q}}}) \cdot (p_{T,W} - \frac{M_W - M_{\bar{q}}}{dR_{q\bar{q}}}) \quad (18)
\]

Equation (16) can be solved for \( p_{T,W} \) which is the typical value to be computed for fixed angular distances, e.g. jet sizes, to predict if fatjets will occur.
or not. The quadratic components in the solution not depending on $p_{T,W}$ are replaced by $X^2$ as a placeholder.

$$p_{T,W}^2 - p_{T,W} \cdot \frac{2 \cdot M_W - M_{dp}}{dR_{q\bar{q}}} = \frac{X^2}{dR^2}$$  \hspace{1cm} (19)$$

The occurring masses in eq. (19) are substituted by the $M_{dim}$ to tidy up the nomenclature and the quadratic equation is solved:

$$p_{T,W} = \frac{M_{dim} + \sqrt{M_{dim}^2 + X^2}}{dR}$$  \hspace{1cm} (20)$$

The placeholder $X^2$ determines here, whether an upper limit, a lower limit or a function describing the peak-behaviour is considered. This reflects the neglected problem of direction of the momenta $p_{mass}$, completely converting them into transverse momenta. The most likely set of directions is the case for $X = 0$, shown in Fig. 8.

Figure 8: The left plot shows the distance between fully hadronic products of a W decay on parton level, depending on the $p_T$ of the W boson. A MC Z’ with 3 TeV/c$^2$ mass decays to $t\bar{t}$. The W bosons out of hadronic top decays are shown. Only Ws the in range of $\pm$ 10% around 80.4 GeV/c$^2$ are shown. The right plot shows the deviation of the bin with most entries in each row from the prediction.

Figure 8 shows how well eq. (20) works.

All placeholders $X$ have been found empirically, but there are some considerations involved: For upper and lower limits, the placeholder must have
the dimension of a mass due to unit considerations. Only two masses are known and where originally included in the quadratic part of the solution: $-M_{dp}$ and $M_{W}$. Functions of those masses are candidates for a solution.

For the W boson the decay product masses are typically negligible and even if considered, light flavour jet masses are ill defined due to offshell-effects.

If $X^{2} = M_{W}^{2}$, an upper limit can be found as shown in Fig. 9

![Figure 9](image)

Figure 9: The left plot shows the distance between fully hadronic products of a W decay on parton level, depending on the $p_{T}$ of the W boson. A MC $Z'$ with 3 TeV/$c^2$ mass decays to $t\bar{t}$. The W bosons out of hadronic top decays are shown. Only Ws in the range of $\pm$ 10% around 80.4 GeV/$c^2$ are shown. The right plot shows the fraction of events covered by the upper limit prediction.

The upper limit covers about 50% of the events, which is not a very good prediction at all.

If $X^{2} = -M_{dp}$ a lower limit can be set. Since the masses of the decay products were previously neglected, a crude approximation is using the combined bottom- and charmed quark masses as $M_{dp} = 5.5$ GeV/$c^2$. The result is shown in Fig. 10.
Figure 10: The left plot shows the distance between fully hadronic products of a W decay on parton level, depending on the $p_T$ of the W boson. A MC $Z'$ with 3 TeV/$c^2$ mass decays to $t\bar{t}$. The W bosons out of hadronic top decays are shown. Only Ws in the range of $\pm 10\%$ around 80.4 GeV/$c^2$ are shown. The right plot shows the fraction of events covered by the lower limit prediction.

A binning effect is clearly visible in the coverage plot due to the different scales of the $p_T$-bins and angular bins. 85% of events are covered by the lower limit for small angular distances. For a normal jet size of $dR = 0.5$ an estimate of $\approx 300$ GeV for $\approx 80\%$ of the events can be made this way.

The goal of the next chapter is to generalize the two-body-prediction to the three-body-case.
3.4 Boosted top

The top quark decays in the fully hadronic case into three quarks in the final state. Without a fourth generation, the top quark as the most massive quark might be preferred in many decays of very heavy mass-resonances of several TeV/c².

In this case a more complex structure has to be considered, where the top first decays to a b quark and a W boson and then the latter decays into two quarks. The additional challenge is that contributions to the final momenta occur at each junction in the decay-chain. This happens in random directions, which was negligible for the W decay, since no matter the direction, the angular distance of both decay-products is narrowed.

This and not negligible decay product masses become an issue, since from a triangle in η-φ-space with about equal length of the three sides, if each cascade-decay happens at a right angle to the boost in its respective rest-system, to a linear decay-shape, if each cascade-decay happens parallel or antiparallel to the beam axis in its respective rest-system, can happen. Therefore two transverse separation impulses \( p_{T,mass} \) are of concern for the final state interpretation of the top decay: from the top decay and from the W decay thereafter.

The triangular case features the closest angular distances, the linear decay the farthest between the two momenta lying the farthest apart.

Taking eq. (15) and considering the W as a single jet, the first stage of the decay can be computed similar to Chapter 6.2 with a peak-function, an upper and a lower limit:
Figure 11: The left plot shows the distance between fully hadronic products of the first stage of a top decay on parton level, depending on the $p_T$ of the top. A MC Z' with 3 TeV/$c^2$ mass decays to $t\bar{t}$. The W bosons out of hadronic top decays are shown. Only tops in the range of ± 10% around 172.5 GeV/$c^2$ are shown here. The right plot shows the deviation of the bin with most entries in each row from the prediction.

Figure 12: The left plot shows the distance between fully hadronic products of the first stage of a top decay on parton level, depending on the $p_T$ of the top. A MC Z' with 3 TeV/$c^2$ mass decays to $t\bar{t}$. The W bosons out of hadronic top decays are shown. Only tops in the range of ± 10% around 172.5 GeV/$c^2$ are shown here. The right plot shows the fraction of events covered by the upper limit prediction.
Figure 13: The left plot shows the distance between fully hadronic products of the first stage of a top decay on parton level, depending on the $p_T$ of the top. A MC $Z'$ with 3 TeV/c² mass decays to $t\bar{t}$. The W bosons out of hadronic top decays are shown. Only tops in the range of $\pm$ 10% around $172.5 \text{ GeV/c}^2$ are shown here. The right plot shows the fraction of events covered by the lower limit prediction.

Fig.: 11 shows that the peak-distribution is perfectly described, similar to the W boson-case. The biggest difference is in shape. The distribution of the events is symmetric with respect to the peak-function (Fig. 11), because $M_{dp} \approx 0.5 \cdot M_t$.

The upper limit in Fig. 12 and the lower limit in Fig. 13 both cover $\approx 80\%$ of the events. This is much better than for the W, which is to be expected, since the decay-products’ masses are better defined.

For a standard jet size of $dR = 0.5$ a lower limit of $p_{T,t} \approx 420 \text{ GeV}$ can be read off the function in Fig. 13 with a coverage of $\approx 80\%$. The added complication of a second decay after this stage is a problem for estimating whether two or three of the jets will be merged or not in a given jet size. To get another handle, the dependence of the W boson’s decay product’s angular distance from $p_{T,t}$ can be studied.

The most basic assumption is $\frac{p_{T,t}}{2} = p_{T,W}$ for a lower limit, if the transverse impulse of the top greatly exceeds the topmass. The light flavour quark-jet masses are again neglected.
Figure 14: The angular distance between the quark-jets out of the W boson of a top decay on parton level, depending on the top $p_T$. A MC $Z'$ with 3 TeV/$c^2$ mass decays to $t\bar{t}$. The W bosons out of hadronic top decays are shown. Only tops in range of $\pm$ 10% around 172.5 GeV/$c^2$ are shown here.

Fig. 14 demonstrates that the lower limit-assumption is well-founded, but no more information can be derived this way. The general $\frac{1}{x}$-shape of the distributions stays the same, no matter what decay. It is safe to assume generality of the kinematic equations derived for any decay of any particle, known or unknown.

For the standard jet size of $dR = 0.5$ a lower limit on $p_T$, of $\approx$ 320 GeV for a boosted $q\bar{q}$-structure can be read off the function in Fig. 14. All resulting limits and expectations for the most common jet size are summarized in Tab. 3.4:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$dR_{q\bar{q}}(p_{T,W})$</td>
<td>310</td>
<td>320</td>
<td>390</td>
</tr>
<tr>
<td>$dR_{W}(p_{T,t})$</td>
<td>420</td>
<td>560</td>
<td>690</td>
</tr>
<tr>
<td>$dR_{q\bar{q}}(p_{T,t})$</td>
<td>320</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Table for upper, lower limits and peak expectations of the angular distances between top decay products depending on transverse momenta of either top or W boson for $dR = 0.5$.

The difficulties and properties of decay-chains in context of boost have been shown at the example of the top decay. The next goal is to find distinctions from QCD background.
3.5 Boosted QCD

QCD is the main background for all studies of substructure, because it can produce jets of very large invariant masses with substructure due to final state gluon emission.

According to [27] the “leading-order (fixed-coupling) differential QCD jet mass distribution” goes:

\[
\frac{1}{n} \frac{d \sigma}{d M^2} = \frac{1}{M^2} \, \frac{\alpha_s C_i}{\pi} \left( \ln \frac{R^2 p_T^2}{M^2} + O(1) \right)
\]  

(21)

For large \( p_T \) the running of the strong coupling constant will be largely compensated by the logarithm. High energy QCD produces jets of large invariant masses.

This becomes an issue for studies with heavy boosted particles like tops or W bosons, since QCD will always have higher cross sections and produce jets with jetmasses equal to the sought particle masses.

An obvious distinction of QCD from conventional searches for heavy particles is the jet-multiplicity that is suppressed by orders of \( \alpha_s \) for each additional jet. This distinction is lost in the boosted case. Boosted particle searches need a kinematic distinction to not misidentify QCD and hadronization as substructure.

This can be found by following a calculation by Salam on a two-pronged decay into particles i and j and defining a new parameter \( z \) for \( p_{T,i} < p_{T,j} \):

\[
z = \frac{p_{T,i}}{p_{T,i} + p_{T,j}} = \frac{p_{T,i}}{p_T}
\]  

(22)

Equation (22) can be redefined by the use of \( z \) to:

\[
M^2 \approx z(1 - z)p_T^2 dR^2_{ij}
\]  

(23)

Out of equation (23) two things can be learned: QCD has a soft divergence and a collinear divergence. Hadronization is a QCD-problem. In each step of hadronization, a soft and collinear splitting to occur is most likely.

In Chapters 6.2 and 6.3 it was shown that massive particles decay at wide angles.

The difference in angular properties can be used to distinguish hadronization from substructure and unmerge fatjets into their composite parts, called “subjets”. The methods to do this will be specified in the next chapter based on the foundations of this chapter.
4 The LHC-experiment

The Large Hadron-Collider (LHC) is the most powerful accelerator built so far. It is made of superconducting niob-titan cavities cooled down to 1.9 K by superfluid helium [14].

The LHC is situated in Geneva, inheriting the former Large Electron Positron-collider’s (LEP) tunnel with a circumference of 26.7 km.

According to the technical data report [25], the center-of-mass energy at each interaction point ($\sqrt{s}$) is 14 TeV and is currently at 7 TeV. It can accelerate up to about 2800 bunches of 115 billion protons.

The injection of protons or heavy ions is done by a linear preaccelerator and three smaller synchrotrons, the “Proton Synchrotron Booster”, “Proton Synchrotron” and “Super Proton Synchrotron” [15].

4.1 Layout

The bunch crossings will take place at four different interaction points, each containing an experiment.

Figure 15: Layout of the underground installations of the LHC [18].

29
Figure 15 shows the layout of the vaults containing the experiments:

The Large Hadron Collider beauty-experiment (LHCb) has been designed to measure CP-violation and search for indications of new physics with bottom quarks [17].

ALICE is an experiment solely dedicated to heavy ion-physics. Its main component is a time projection chamber (TPC) [19], containing no calorimeter. It has only a weak magnetic field of 0.5 Tesla, since heavy ion-collisions produce very many particles at low energies.

There are two multipurpose detectors, A Toroidal Lhc ApparatuS (ATLAS) and the Compact Muon Solenoid (CMS).
4.2 2010 and 2011-runs

Since 2010 the LHC is operating with an energy of 3.5 TeV per beam. Nearly 50 pb$^{-1}$ were delivered in 2010 (Fig. 16).

![LHC 2010 RUN (3.5 TeV/beam)](image)

**Figure 16:** Official luminosity report 2010 [24].

In 2011 the bunchspacings were continuously reduced, increasing the luminosity. By 09/2011 the delivery of roughly 1 fb$^{-1}$ per month was possible (Fig. 17).
Figure 17: Official luminosity report 10/2011 [24].

In this thesis, 2 fb$^{-1}$ of data recorded by CMS will be used.
5 The CMS detector

CMS is a multipurpose detector. The information in this section is according to the first technical data report of the CMS collaboration [25].

CMS consists of several subdetectors with specifically tailored abilities to reconstruct all electromagnetically and strongly interacting particles traversing it.

In Fig. 18 a profile of the detector is shown. It is built symmetrically around the beampipe.

The innermost layer around the beampipe is the silicon tracker (see Chapter 4.1), depicted as light gray.

The calorimeter (green and orange) is built around the tracker (see Chapter 4.2), but still inside the superconducting coil (dark grey) that produces a magnetic field of 4 T.

The outermost subdetector consists of iron yokes and drift chambers which form the muon system (see Chapter 4.3).

A forward and a barrel section of the detector can be defined. The barrel is the main part of the detector, designed to reconstruct particles with high transverse momenta. The barrel section’s detectors are made of layers par-
allel to the beam line. In detector coordinates, the barrel is in $|\eta| \leq 3$

The forward part of the detector consists of endcaps on each subdetector, perpendicular to the beamline, to extend the reach of detection to jets with high longitudinal momenta and are mainly used for reconstructions of the complete energy deposition in the detector. In detector coordinates, the forward calorimeter reaches to $3 < |\eta| < 5$.

5.1 Silicon Tracker

The tracking system is divided into two parts: the pixel and the strip detectors.

The working principle of a silicon detector is to apply a high voltage in reverse bias on a doted flat piece of silicon that basically resembles a diode. If a charged particle passes through the silicon, ionization can occur. Electron-hole pairs are produced that are separated by the high voltage, creating not only one current, but two currents:

One at the positive and one at the negative junction. In analogue mode, the shape and timing of both currents can be used to identify the position of the charged particle’s passthrough and distinguish it from noise. The amplitude can be used to calculate the ionization loss of energy $E$ per distance $x$: $\frac{\delta E}{\delta x}$. The reconstruction software uses the entries of incidents within the tracker modules to reconstruct the most likely charged particle trajectories for each event.

Due to the magnetic field, trajectories of charged particles will be bent, making an accurate measurement of their energy possible by only using the track informations for low energy particles.

5.1.1 Pixel detector

Each pixel of the pixel detector has a size of $100 \cdot 150 \, \mu\text{m}$. The pixel detector consists of several layers of modules out of which each contains several pixelchips.

In the barrel, 3 layers of pixel detectors are placed. Two wheels of turbine-like design as in Fig. 19 on each side cap off the forward part of the detector.
Figure 19: Several pixel detector-chips on a wheel for the pixel forward detector [20].

Pixels are placed closest to the beampipe ($\approx 10$ cm) for the best possible vertex reconstruction, especially for vertices with small separation.

5.1.2 Stripdetector

At distances of $20 - 55$ cm to the beampipe, the strips of the stripmodules have sizes of $10$ cm $\cdot 80$ $\mu$m. The inner barrel tracker in this region is made of four layers, the inner disc tracker has four discs on each side.

For the outer ranges of above $55$ cm to the beampipe, the strips are of $25$ cm $\cdot 180$ $\mu$m size. The outer barrel tracker consists of six layers, while the endcaps contain nine discs.

The overall material budget of the silicon tracker, its cooling and support structure is a problem, because it introduces a fraction of a radiation length before the actual calorimeters in $\approx 1.1 < |\eta| < 2$ [21]. The readout channel support structure for the tracker is shown in Fig. 20.
5.2 Calorimeters

Calorimeters are used to measure an energy deposition of particles, stopping and absorbing them completely in the process.

For charged particles and photons there are two basic processes in the calorimeter:

1. Radiation: The charged particle enters the electromagnetic field of a dense matter structure. This results in a photon emission.

2. Pair production: A photon initially emitted by the charged particle will be able to produce electron-positron pairs in an electromagnetic field.

On average, any of those processes happens each so called “radiation length”. This leads to an exponential decrease of energy per particle in this cascadic decay, called electromagnetic shower, until no further pair production is possible.

For strongly interacting particles hadronic interactions are the equivalent.
Energy can be measured in the calorimeters by measuring the spread, intensity and intrusion depth of the cascade, using e.g. Avalanche Photo Diodes (APD) to measure activity of electromagnetic showers directly or by using a material that emits photons due to excited states returning to ground states after excitation by ionization due to particles. This is called a “scintillator”

5.2.1 Electromagnetic calorimeter

The Electromagnetic CALorimeter (ECAL) consists of lead tungstate (PbWO$_4$) crystals with the density of lead and the opacity of glass. Combining these properties, charged particles’ energy depositions can be measured with an excellent precision by avalanche photodiodes on each crystal bar, building a module.

The ECAL is made out of towers comprised of a 5 · 5 modules each that are bound to a common trigger and readout structure. A view of the ECAL in the barrel is shown in Fig. 21.

![Figure 21: Inner view of the ECAL [20].](image)

The ECAL was designed to best reconstruct a possible diphoton decay of a Higgs and has a very high density. This limits the possible efficiency of the

37
hadronic calorimeter, because there is more than one hadronic interaction
length already due to ECAL and tracker.

5.2.2 Hadronic calorimeter

The Hadronic CALorimeter (HCAL) is a sample calorimeter. Scintillators
are built in turn with absorbers out of stainless steel and brass.

In $(|\eta| < 1.740$, detector coordinates are defined in Chapter 5.1) each
tower covers an angle of $\Delta \eta \times \Delta \phi = 0.087 \times 0.087$. The installation of an
HCAL-part is shown in Fig. 22.

![Installation of a HCAL-part into the solenoid](image)

Figure 22: Installation of a HCAL-part into the solenoid [20].

The performance of the HCAL of CMS suffers from more noise experi-
enced than by design and an excessive material budget in front of the HCAL.

5.3 Muon system

Minimally Ionizing Particles (MIP) traverse the complete detector. The
muon is the typical MIP. A system of iron yokes is designed to return the
magnetic field to the coil by gaseous detectors.

The gaseous detectors are comprised of drift tubes, gas-filled tubes with
a wire in their center. Between tube and wire a high voltage is applied. An ionizing particle will detach an electron from a gas-atom in the tube. Ion and electron are separated and accelerated by the high voltage, leading to further ionization, causing an avalanche.

By measuring the time of detection, cosmic muons can be identified as first activating the muon system. Muons produced in hard interactions can be identified by entries in muon system and tracker. The iron yokes of the muon system are shown in Fig. 23.

Figure 23: Outer view of the Muon system’s drifttubes, interspersed with layers of iron [20].
5.4 Particle flow

Colored particles quickly hadronize due to asymptotic freedom by gluon emission. Charged and neutral components are produced. The goal of the Particle Flow (PF) algorithm applied by CMS is to identify all charged particles in an event by combining the information of all subdetectors [26].

This leads to three possible improvements:

1. Charged particles with low momentum can be more accurately measured in terms of energy by tracker entries than by calorimeter information. The error of the calorimeters $S_{\text{CAL}}$ is influenced by poisson statistics, leading to a $S_{\text{CAL}} \propto \frac{1}{\sqrt{E}}$ behaviour. For high momenta, tracks equal straight lines, but the relative calorimeter error decreases drastically in this case. The assessment with less estimated error is used by the algorithm.

2. The identification of particle flavour allows for specific response corrections.

3. The ratio of charged particles and neutral particles ($\approx 66:33$) inside jets is known. By correctly identifying the charged components, an extrapolation of the neutral component is attempted.

The algorithm itself works by matching tracks reconstructed in the tracker and muon system with calorimeter entries. First ECAL-entries are matched with tracks and assigned a likelihood of electron or photon-incidents. Up to a certain likelihood cut, tracks and ECAL-entries will be identified as either photons (ECAL hit and no track match) or electrons (ECAL hit and track match, no muon system match). Identified particles and all entries leading to these particles are removed from consideration for further likelihood assessments.

For all supported particles, the algorithm continues identification assessment. Ideally, all tracks are accounted for and the error of the measurement is reduced by extrapolation of the neutral component deposited in the calorimeters, especially the HCAL.
5.5 Factorized jet-corrections

In CMS, a factorized jet correction approach is used. The theoretical idea is, that correction factors due to detector effects, pileup, underlying event, flavour dependent response and Monte Carlo (MC) differences can be separately assessed and applied as correction factors. The corrections available so far are:

L1 Offset-correction due to pileup, based on jet-area or the number of vertices in the event.

L2 Relative correction smoothing the jet-response in $\eta$, computed with MC.

L3 Absolute correction smoothing the jet-response in $p_T$, computed with MC.

L2L3residual Correction of differences between MC and data, smoothing MC differences in jet response in $\eta$ and $p_T$. Only applied on data, not MC.
6 Clustering and declustering mechanisms

This chapter deals firstly with the “clustering” of jets, followed by the “declustering” in order to find substructures. Constraints on the declustering will be derived, culminating with the introduction of three distinct algorithmic improvements for declustering.

6.1 Clustering of jets

To cluster a jet means to reconstruct the jet momentum out of pure calorimeter entries or (for Particle Flow [28]) calorimeter entries corrected by information out of the tracker and the muon system, called clusters.

For this purpose, two theoretical problems have to be taken care of. Infrared (IR) safety and collinear safety.

IR safety means: if soft objects are randomly added to the list of clusters, the output must not be affected. The most basic example would be to have two adjacent hard partons in the generator that will be separate jets after clustering. Adding soft radiation between the two partons can merge both hard partons into one jet for an IR-unsafe algorithm.

Collinear safety deals with collinear splittings. A collinear splitting is e.g. a gluon emission by a quark at a small angle. A collinear unsafe algorithm will cluster two jets in this case, if two adjacent calorimeter entries happen to fit into one jet size with the unsplit entry. A collinear safe algorithm will cluster only a single jet, no matter if there is a collinear splitting or not.

IR and collinear (IRC) safety is also an issue in perturbative calculations. A loop of infinitely low mass in the matrix element of the process corresponds to a soft cluster in the detector and to IR safety. The result in perturbative calculations is a negative logarithmic divergence [27]. On the other hand, collinear splittings in the matrix element correspond to a positive logarithmic divergence. Both divergences cancel if the algorithm is IRC safe.

A comparison of theory to experimental data has to include the use of IRC-safe algorithms or is only valid to a certain order after which the used algorithm becomes IRC-unsafe.
6.1.1 Cone and sequential recombination algorithms

Two kinds of algorithms are commonly used for the recombination of all objects to four-momenta. Either cone algorithms or sequential recombination algorithms.

Cone algorithms work using a fixed radius $R$ as an angular distance in detector coordinates (see Chapter 5.1) with respect to a seed, recombining all objects within that radius to a single four-momentum using a recombination scheme. The recombination scheme, no matter if objects are recombined by sequential recombination or cones, defines how the momenta are combined. The “E-scheme” constructs the sum of the four-momenta of all clusters designated to the jet and is the one used in the following.

The basic concept of sequential recombination is to choose a mathematical metric defining a distance-parameter $d$ between two objects $i$ and $j$, a jet size $R$ and a scheme for recombining the clusters.

All clusters in the detector are objects and their distance among each other is named $d_{ij}$ while the distance to the beamline is called $d_{iB}$ [29]. During every step of the sequential recombination the objects closest together in $d_{ij}$ are recombined until $d_{ij} > d_{iB}$. In this case the four vector at that time is promoted to a jet.

Sequential recombination is collinear safe by definition, since collinear splittings and hadronization in general will coincide with close distances in $d_{ij}$. IR-safety though can only be achieved by imposing a requirement of minimal energy on the final four-momenta.

For the standard metrics used at the LHC, the distances can be parametrised as:

$$d_{ij} = \min(p_{T,i}^{2^n}, p_{T,j}^{2^n}) \frac{dR_{ij}^2}{d_{\text{max}}^2}$$  \hspace{1cm} (24)

$$d_{iB} = p_{T,i}^{2^n}$$  \hspace{1cm} (25)

The exponent $n$ can be positive, negative or zero.

For $n < 0$, explicitly $-1$ here, the standard algorithm in CMS: anti-kt (anti-kt). Every negative exponent yields the same type of behaviour.

For $n = 0$ the most simple algorithm Cambridge-Aachen (Cambridge-Aachen).

For $n > 0$, explicitly +1 here, an algorithm mostly used at lepton colliders, kt, is the result. Like for the anti-kt-type, all positive exponents yield the same behaviour.

For all algorithms mentioned the general behaviour will be explained in the following chapters. In the next section, the actual process of clustering using the metrics of eq. (24) and eq. (25) will be explained in detail.
6.1.2 Sequential algorithms: the FastJet package

FastJet [30] is the software package used to cluster jets, that is called from the standard CMS software (CMSSW). It is open source and the main reason for its implementation are two algorithmical improvements on the basic process of sequential recombination.

The normal process includes (taken from [31]):

1. Calculate $d_{ij}$ for each pair of objects and $d_{iB}$ for every object in the detector, once. This yields an $O(N^2)$ of calculations.

2. Find the minimal distance. This yields another $O(N^2)$ and is done $N$ times.

3. Recombine the pair or promote the cluster to jet. This is of $O(1)$, done $N$ times.

Step two is dominant, resulting in $O(N^3)$ calculations.

FastJet improves this by replacing step one. Instead, two arrays are constructed. One for the distance of the object to the corresponding nearest neighbour and another for the distance to the beamline. This yields $O(N^2)$ calculations.

Further improvement is reached by implementing so called “Voronoi diagrams” instead of arrays that contain the distances to all nearest neighbours surrounding a cluster in the respective metric. The final result of FastJet yields $O(N \ln N)$ calculations.

The format that clusters and jets in FastJet are saved in are “pseudojets”. A pseudojet contains not only the four-momentum, but also the hierarchical history of every recombination of clusters leading to that object and all four-momenta of its constituent clusters. This inherited history is an important issue for declustering.
6.1.3 The anti-kt jet clustering algorithm

Anti-kt is the commonly used algorithm. The main virtue of using a negative exponent to $p_T$ is that the resulting jets are conelike.

Anti-kt jets are round in apparel and have a perfectly circular area, meaning that all clusters within radius $R$ of centre of energy of the jet constitute the jet with an area of $\pi R^2$.

The measure of area for a jet is done by introducing “ghost particles” (ghosts). Ghosts are infinitely soft clusters that are added to the collection of clusters from the detector. For any IR-safe algorithm, the outcome of the clustering does not change by the introduction of ghosts.

The usage of ghosts permits two meaningful area-definitions [32]:

1. The “passive area”, comprising the distance where a solitary ghost particle will be added to a jet. This is a measure for intake of diffuse radiation.

2. The “active area”. A random number of ghost particles is randomly distributed. The active area is defined as the number of ghosts clustered into the jet, divided by the ghosts per unit area and averaged over many sets of ghosts. In this case, jets comprised out of only ghosts will be clustered, called “ghost jets”, which compete with other jets. The active area is a measure for the impact of radiation like pileup.

These definitions for area are used for jet corrections, since contributions of PU and UE are dependant on $\eta$ and are regarded as constant per unit area, averaged over time. The area is a measure for any offset-correction that needs to be applied.

For the special needs of boosted topologies, the active area defines the angular distance, where two jets will be merged.

Sapeta [33] shows for anti-kt, that active and passive area are equal and $\pi R^2$. The basic behaviour of the anti-kt algorithm is resilience to PU and UE, as well as simple perturbative calculations and jet corrections.

Defining the parameter $z$:

$$z = \frac{\min(p_{T,i}, p_{T,j})}{p_{T,i} + p_{T,j}}$$  \hspace{1cm} (26)

the distance requirement for two hard partons, $i$ and $j$, to be clustered into a single jet or not, can be derived [33].

Three scenarios have to be taken into account:

1. For $\frac{d_{ij}}{p_T} < 1$ : one single and circular jet is formed.
2. For $1 < \frac{d_{ij}}{R} < \frac{1}{1-z}$: two distinct jets that are adjacent are built and the jet with lower energy produces a crescent-shape indentation in the harder jet.

3. For $\frac{1}{1-z} < \frac{d_{ij}}{R}$: two circular jets are formed.

Case two is the rare exception from circular jets produced by anti-kt. Anti-kt will cluster two hard partons together into a single jet if they are close enough together. The harder parton then accumulates clusters from the softer parton. The result is one jet where the original two hard partons cannot be distinguished anymore. No substructure is resolvable.
6.1.4 Cambridge-Aachen

For Cambridge-Aachen, active and passive jet-areas are not circular. It is not commonly used in conventional analyses at the LHC, but is the most important algorithm for searches with substructure techniques.

For the passive area in a two-particle event, three regions can be defined:

1. For $\frac{d_{ij}}{R} < 0.5$ : the jet is formed circular and comprised of both particles.

2. For $0.5 < \frac{d_{ij}}{R} < 1$ : the passive area of the jet increases. The second particle acquires a passive area of its own before it is clustered together with the rest of the hard parton. The overall area is extended by the latter area, producing noncircular jets.

3. For $1 < \frac{d_{ij}}{R} < 2$ : the passive area is decreased. The second particle will build a jet of its own that takes passive area from the hard jet.

For the active area, three different regions are definable:

1. For $\frac{d_{ij}}{R} < 0.75$ : the active area is about 0.8. Ghost jets will fill the detector and take active area from each other.

2. For $0.75 < \frac{d_{ij}}{R} < 1$ : there is an increase in active area. The two hard jets will claim all adjacent clusters and ghost particles before being clustered together, increasing the active area.

3. For $1 < \frac{d_{ij}}{R} < 2$ : the active area is decreased to about 0.8. Two distinct jets will be clustered and take active area from each other.

In Cambridge-Aachen the concept of area is fragile and can only be redeemed if the jet size is small enough to cluster every hard parton into a single jet by itself.

If two partons are clustered together, two independently clustered jets are clustered and then ideally merged in the final step of recombination. The basic structure of the event is still accessible by declustering methods (see Chapter 7.2).
6.2 Substructure of jets

Substructure of jets can be one jet containing more than one group of gaussian-distributed clusters. Substructure can also be a non gaussian component, equaling an offset. It can as well consist of another gaussian group of clusters, literally a second hard jet.

Three different types can therefore be defined:

I: Subjet from uncorrelated sources, overlapping the hard jet considered or clustered together with it. It is typically soft, originating from proton-leftovers, initial state radiation, beam-rests and/or scatterings, e.g. pileup (PU) and underlying event (UE).

II: Subjet from correlated sources, clustered together with the hard jet considered, originating from the same primary vertex, but another branch of the Feynman diagram.

III: Subjet from correlated sources, originating from the decay of a single boosted particle, clustered together into a single jet, e.g. Fig. 24.

This thesis refers to type III if not explicitly mentioned otherwise.

Figure 24: Hadronic boosted top decay into a single fatjet with three subjets.
6.3 Illustration of Cambridge-Aachen-clustering

In this chapter, a clustering of a jet will be shown, step by step. This jet will be declustered in the end. It is possible to look for substructure during the clustering by eye.

These pictures were produced by asking the pseudojet for its parents until the constituent-level was reached, making a picture at each level. The order of recombination is not the same as during the clustering in a temporal sense, but in a hierarchical and spatial one.

Pseudojets with many parents will therefore start to merge first.

Figure 25: Arbitrary binning of clusters, constituting a generated fully hadronic top decay clustered into one jet at the end.

Figure 25 starts with 92 Clusters. The binning is four times as fine as the cells of the calorimeter, but still many constituents can hide in any of the bins. Identified leptons for example make their own constituents, since PF-clusters are used.
Figure 26: Arbitrary binning of clusters, constituting a generated fully hadronic top decay clustered into one jet at the end, second step of clustering.

The second step of the clustering is Fig. 26. One picture was omitted, since only one single merging occurred. The particles closest in angle in the areas with the highest particle density were recombined, so that no effects outside of the binning are visible, yet.

Figure 27: The first bins are emptied and already three distinct collimated structures can be seen. The first at (0.55, 0.7), the second at (0.63, 1) and the last at (0.68, 0.7).
Figure 27 demonstrates the first visible recombinations.

Figure 28: The collimated structures further merge, while the outer soft clusters have a lot less constituents. Most likely they consist of PU or UE.

In Fig. 28 the arrows show the first mergings of clusters.

Figure 29: The first outer clusters are merged, while the free space around the hottest clusters grows. Between 0.6 to 0.7 in $\eta$ and 0.7 to 1 in $\phi$ we still have some clusters that cannot be conclusively allocated by eye.

Figure 29 demonstrates the accumulation of energy in the area with highest particle density and energy at $\eta = 0.55$ and $\phi = 0.7$. 
Figure 30: For the clusters between 0.6 and 0.7 in $\eta$, the working of the $E_{\text{scheme}}$ can be seen with respect to the next figure. By four-momentum combination the difference to the next figure is striking, since hard clusters will be combined.

Three distinct jets can be discerned by eye in Fig. 30. At this point the main structures have been recombined and will only proceed to accumulate energy from nearby clusters.

Figure 31: The soft outer clusters are being recombined. It is seven hierarchical steps further into the clustering. The number of constituents for those clusters is up to $2^7$ less than the number of clusters at the start of the procedure.
Figure 32: A three-pronged structure of the top decay being clustered can be seen. Although 92 close and indistinguishable clusters were present at the start of clustering, the jet algorithm has found the right recombinations. By going backwards in the sequence of clustering, later on, it is possible to find this view, revealing three jets instead of one, again.

Figure 33: The soft outer clusters will merge with each other at this point. Still they constitute almost no $p_T$, unlike the hot clusters. Searches for boosted substructures allow to ignore almost everything soft, unlike in normal applications for jets. This will be used during the declustering.
Figure 34: The hard clusters finally recombine. The most probable candidates for the quark-jets out of the W recombine first. As shown in Chapter 6 this is a direct result of the kinematic behaviour of the $q\bar{q}$-angular distance regarding $p_{T,t}$.

Figure 35: The last recombinations follow. The outliers are orders of magnitude less energetic than the subjets and are recombined at wide angle. During the whole clustering, the angular ordering is retained. If looked at from a hierarchical perspective, every step further on monotonically enlarges distance for recombinations.
This chapter shows the working not only of Cambridge-Aachen clustering by angular distance, but also of combining this procedure with the hierarchical pseudojets’ structure that will be used in Chapter 7.2.

The important main facts are:

- All constituents are ordered monotonically by angular distance.
- Going down a hierarchical step of the pseudojet’s parentage decreases angular distance and transverse momentum.
- Outlying soft structures are easily distinguishable by eye from the hard structures of the jets.
6.4 Declustering

The basic concept for finding substructure is to use an appropriate distance-
metric, that resolves the type of substructure that is searched for.

Here Cambridge-Aachen will be used, since it prefers collimated struc-
tures. As long as high transverse momentum is correlated with a small
angular distance between multiple clusters, Cambridge-Aachen will cluster
the corresponding parts together, first. Each hard parton will be clustered
into a “subjet” of its own until the distance of separation between the two
closest hard partons is reached. Then both will be clustered together, two
“subjets” inside one “fatjet”.

This idea gives rise to use the working structure of FastJet pseudojets,
which contain the history of recombinations of all constituent-clusters of the
jet, to access the “subjets”.

While going backwards in the history of recombination shown in Fig. 37,
a sorting algorithm has to be used, that allows to find all subjets without
declustering down to constituents or producing soft outlying subjets out of
pure UE. How this can be achieved will be explained in detail in Chapter
7.3.
Figure 37: Overview of pseudojets hierarchical parentage-history. Each step down from the green triangle, the final jet, equals one call to a “has parents”-routine or one step back in the history of clustering. Every block is a pseudojet, the red blocks are constituent level.
6.5 Constraints on declustering

Declustering is the application of a sorting algorithm on top of the jet clustering, defining new “subjets” within the original jet, called “fatjet”. The sorting algorithm that works backwards through the history of clustering of a pseudojet, addresses three issues at each junction:

1. Are both parents considered substructure?
2. If not, which parent is more likely to contain substructure and shall be probed further?
3. After each step, is there possibly more substructure to be found or shall the algorithm make a full stop here?

For the first question, boosted kinematics have to be considered (see Chapter 6), which will be done in detail in the sections two and three of this chapter.

The second question can be dealt with by the simple idea that more energy means more boost, so substructure is more likely in a hard parent than in a soft one.

The third question is implemented in different versions for different analyses. This will be explained in detail in Chapter 8. General answers can be derived to some extent, which will be explained in sections 7.3.1 and 7.3.4.

6.5.1 Calorimeter granularity

The granularity of the calorimeter is finite. This has to be considered when looking at the minimal distance two parents can have and still be considered distinct substructure.

Particle Flow allows for tracker corrected clusters, which equals a much finer granularity than by calorimeter towers alone, but the neutral component of the jet can only be extrapolated.

If two jets hit the same or adjacent calorimeter towers, perhaps even overlapping, it cannot be decided anymore which jet originally contained which energy, no matter if the charged components can still be distinguished.

The tower distance of the hadronic calorimeter, being \( dR = 0.087 \) in \((|\eta| < 1.74)\), is taken as an absolute minimal distance, shown in Fig. 38.
Figure 38: Using the kinematic space of the W as an example, this figure shows on the one hand the theoretical curve of the angular distance between decay products of the W depending on $p_{T,W}$ and on the other hand shows the effect of the calorimeter-granularity. As can be seen, only extremely high energies would not be resolvable in the detector. For the Z'-MC of 3 TeV shown here, this is not an issue.

Using the approximation formula eq. (18) (see Chapter 6.2) for the boosted W boson, shown in Fig. 38, this equals a $p_T$-cut of $\approx 1850$ GeV/c, exceeding most theoretical scenarios in reach. This cut therefore effects overlapping jets, not boosted jets.

6.5.2 Massdrop

A possibility to distinct substructure from parts belonging to a jet or soft UE or PU is:

Any splitting of a fatjet $A$ into subjets $B$ and $C$ of interest means to reduce the mass of the more massive subjet, $M_B > M_C$ below a threshold relative to the mass of the fatjet $M_A$. This is arbitrarily taken to be 66%.

$$M_B < 0.66 \cdot M_A$$

This is a phenomenological approach. It was first developed for Higgs studies, where the process $H \to b\bar{b}$ in associated Higgs production was considered. For a light Higgs of 120 GeV/c$^2$ the mass $\frac{M_B}{M_A}$ is expected to be of order 10% and thus smaller than the cut chosen.

The massdrop is identical to a shift perpendicular to the theoretical curve.
in Fig. 38 to lower energies and angular distances in kinematic space. Any massive particle boost is characterised by its mass and the decay product masses. Any decay chain will be resolved into a minimum of the decay product masses until actual hadron masses.

The method will correctly identify possible elements of the decay chain, but also incorrectly identify hadronization as such elements. An upper limit on the number of subjets to be found has to be included from start when this method of identification is used. It is only applicable if the expected number of subjets for a search channel is known.
6.5.3 W-method

Another method to find substructure is to define two conditions:

1. QCD has a soft and a collinear singularity. Hadronization is found in the kinematic space at low $p_T$ and low angular distances $dR$.

2. The W boson is the least massive interesting particle for substructure.

Considering these two conditions, it is possible to divide kinematic space into two areas. A QCD-exclusion-area and a “W boson or more massive”-particle-area, as shown in Fig. 39.

Figure 39: The W boson is shown as well as the typical linear exclusion-function for QCD-splittings. Everything to the lower left of the red line is excluded as hadronization or collinear QCD-splitting, while all structures above the line are considered to be substructure.

A Cambridge-Aachen clustering forces the declustering to undergo a monotone decrease in angular distance and transverse impulse at each step.

The declustering performs a movement in kinematic space towards the excluded region. Once it is reached, the process can be stopped, because nothing of interest (as defined as axiom 2) can be found anymore.

Using this method, the number of subjets found can vary freely. As long as the basic conditions hold true, good results can be expected.
### 6.5.4 Jetsize

Jet sizes are chosen for each analysis taking into account the angular size of jets due to hadronization and the UE and PU that is expected to be accumulated per unit area of the jet.

For analyses with substructure another aspect can be added. For the W-method, a minimal transverse impulse of a fatjet directly derived from the chosen jet size, that is necessary for substructure to occur, suppresses undesirable declustering of diffuse inhomogenous jets. Such a cut is demonstrated in Fig. 40. Later in the analysis a cut for a jetsize of 0.8 is used.

---

**Figure 40:** The kinematic space of the W boson is shown with the peak function. In red an exemplary jetsize is marked and in green the correlated transverse momentum required for a boosted structure to occur.
6.6 Improvements on algorithms for declustering

Three general types of algorithm for improvement are available for searches for substructure.

Their position in the process of clustering and declustering is fixed and depending on the method, at least one may be mandatory. Here this will be trimming.

The Fig. 41 gives an overview where which improvement can be run:

Figure 41: Flowchart for the use of improvement-algorithms. Red boxes are mandatory elements, grey boxes are eligible. The subjects are the end-product.
6.6.1 Pruning

First introduced by [34], pruning is a “bottom→up”-technique, that is applied during the clustering of the jets.

The properties of QCD as explained in Chapter 2.4 allow for using kinematic variables to discern between correlated hard recombinations of the hard scattering process and uncorrelated soft contributions of PU or UE.

By defining the softness variable:

\[ z = \frac{\min(p_{T,i}, p_{T,j})}{p_{T,i} + p_{T,j}} \]  \hspace{1cm} (28)

and taking the angular distance \( dR_{ij} \) between two clusters \( i \) and \( j \) that are to be recombined by a sequential recombination algorithm, two cuts are applied:

1. \( z < z_{\text{cut}} \)
2. \( dR_{ij} < D_{\text{cut}} \)

This vetoes all wide-angle-recombinations that are soft. The cuts can be used with all sequential recombination algorithms. It has been shown to have an impact mostly on contributions of UE, but corrects QCD jets to lower masses.

For searches for heavy particles this side effect is considered to be beneficial, since jets with high mass out of QCD that are considered background are shifted to lower masses, migrating out of the signal-regions.

A second kind of implementation of the same principle is to involve the same cuts as a “top→down”-technique while declustering to veto QCD-jets. An example for the use of such a technique is the Boosted Higgs search mentioned in Chapter 8.1.

In this thesis only the first type of pruning will be used.
6.6.2 Trimming

PU, UE and ISR are soft contributions compared to hard scatterings. This is especially the case when considering boosted jets with transverse momenta of several hundreds of GeV/c.

Soft clusters can be removed by a procedure called “trimming” that cuts on the ratio $X$ of the $p_{T,i}$ of the object and the $p_{T,jet}$. Trimming was first implemented by [35].

This technique is a “top$\rightarrow$down”-technique. After the inclusive jetclustering with a given clustering sequence, the jets are examined by a method of declustering like the W-method.

If the output is a division of the former fatjet $A$ into $N$ subjets, any cluster $i$ not passing a cut on a ratio $X = \frac{p_{T,i}}{p_{T,A}}$ is removed, for example: $X > 0.1$. Thus, soft contributions are removed. Only the hardest parts of the scattering remain.

The second version is applied if the method of declustering sequentially unravels the fatjet. While going backwards in the clustering history, the parents $i, j$ are found by declustering the pseudojet $A$. If $p_{T,i} < p_{T,j}$ the softer parent $i$ will be removed if a cut on the ratio $X = \frac{p_{T,i}}{p_{T,A}} > 0.1$ is not passed.

Whichever version of trimming is applied, the result is a drastic reduction of the active jet-area for all recombination algorithms, increasing the sturdyness against contaminations. The improvement is therefore antiproportional to the area taken by the hard scattering.

As shown in [35], trimming is the more effective for true mass reconstruction, the more boosted the structure is and the less subjets in total are clustered into one fatjet.

In this thesis only the second of type of trimming will be used.
6.6.3 Filtering

The concept of filtering was proposed in [36] for Higgs searches in the channel $H \rightarrow b\bar{b}$ in the context of boosted Higgs out of associated production. The distance of separation $d_{R_{ij}}$ of the first substructure already identified by some declustering procedure is used to run an exclusive clustering on the former jet with a cutoff parameter $d_{cut} = \min(0.3, \frac{d_{R_{ij}}}{2})$.

The $n$ filtered jets with the new smaller radius are ordered in $p_T$ and only the $N$ jets expected in the process searched for are retained. The rest is discarded.

A variant of this technique that will be used in this thesis is to “recluster” each subjet with a new size and/or algorithm. The same separation distance as for the classical filtering is used.

Filtering, however achieved, is done at the end of the procedure of declustering and has been shown to reduce soft contaminations at the edges of hard subjets, further refining the result.
7 Previous CMS analyses using substructure techniques

This chapter is a short overview of several analyses using substructure techniques in the CMS collaboration.

7.1 Boosted Higgs search

Searches for associated Higgs production with a boosted bottom quark pair, $VH \rightarrow b\bar{b}$, are inspired by the paper “Jet substructure as a new Higgs search channel at the LHC” [38].

The basic idea is to recover the $b\bar{b}$-channel for low mass Higgs, which has the highest total cross section, but the most dominant SM background, by selecting events out of the tail in the transverse momentum distribution of associated Higgs production, triggered by the leptonic decay of the associated vector boson or a missing transverse energy variable.

So far, no PAS has been published by CMS, but an AN [39], covering the search channels $W(\mu\nu)H$, $W(e\nu)H$, $Z(\mu\nu)H$, $Z(ee)H$ and $Z(\nu\nu)H$ in a mass range of $110 \text{ GeV}/c^2$ to $135 \text{ GeV}/c^2$ for not boosted jets, where the possible extension to the boosted case is explicitly mentioned. The use of substructure techniques in these channels has so far been discarded, because the cross section for boosted associated Higgs production is very low and not enough integrated luminosity has yet been recorded for this kind of analysis.

The tools for a boosted Higgs search are available since 2010 [40]. The $b\bar{b}$-pair is planned to be resolved by a massdrop (see Chapter 7.3.2) criterion applied on declustering of a Cambridge-Aachen jet of $R = 1.2$, then taking half the separation distance of an interesting splitting and use this distance as a new jetsize in a filtering (see Chapter 7.4.3) method of the first kind. The leading three subjets in transverse momentum, found by this method, are taken to be $b\bar{b}g$ and the two bottom quarks are tagged by secondary vertex methods.
7.2 Search for boosted W and boosted top

A search for heavy mass resonances decaying into top quarks is [41]. It is tried to resolve fully hadronic top decays for the highly boosted scenario with three merged jets, called type 1, and an intermediately boosted scenario, where only the quarks out of the W are merged (see Chapter 6.3 for a more detailed explanation of high and intermediate boost), called type 2.

The analysis strategy is to tag the W in type 2 decays or the top in type 1 decays with a hemisphere algorithm. Substructure in decays of type 1 is resolved by the W-method (see Chapter 7.3.3) with trimming and a transverse momentum cut of 350 GeV/c on a CA-fatjet with $R = 0.8$. The W-jet and the b-jet out of decays of type 2 are reconstructed with a jet pruning algorithm that uses a mass drop (see Chapter 7.3.2) criterion of $M_{\text{subject}} < 0.4 \cdot M_{\text{fatjet}}$ after a pruning (see Chapter 7.4.1) of the fatjet.

Two channels are used, the “type 1+1”-channel using two highly boosted tagged tops for heavy mass resonances and the “type 1+2”-channel using a highly boosted and an intermediately boosted top for mass ranges of $1 - 2$ TeV/$c^2$. The limits derived this way are shown in Fig. 42:

![Figure 42: Limits on a heavy mass resonance decaying to $t\bar{t}$](image)

Figure 42: Limits on a heavy mass resonance decaying to $t\bar{t}$. A combination of search channels for two fully boosted tops, “type 1+1”, and a fully boosted top with an intermediately boosted top (W jet and b jet), “type 1+2”, is shown [41].
8 Analysis using substructure

To demonstrate the power of substructure techniques, an example analysis on a $Z' \rightarrow t\bar{t}$ scenario will be performed, where the $Z'$ is very heavy, so most tops are boosted. First MC studies will be shown for the properties of the methods used to resolve the boosted topologies in this scenario.

Second a complete example analysis on a $Z'$ with 3 TeV/c$^2$ mass and 30 GeV/c$^2$ width will be shown. The analysis will be divided into a part using no substructure methods, to show how well conventional methods perform in boosted scenarios, and a second part, using substructure methods, to quantify the possible improvement.

8.1 Used substructure definition for clustering and jet corrections

For the declustering, the basic Cambridge-AachenTopJetAlgorithm program was modified. It was improved by declustering into up to eight subjets (a subjet can be declustered up to three times), filtering and optional pruning.

- A jet size of $dR = 0.8$ was chosen for the initial clustering of fatjets with Cambridge-Aachen and a $p_T,\text{min} > 50$ GeV/c was required. The fatjets must be in the barrel: $|\eta| < 2.5$.
- While clustering, no pruning is applied.
- During each step of the declustering, it is first checked, whether the pseudojet to be declustered has more than 180 GeV/c transverse momentum. This value is chosen according to the jet size (see chapter 7.3.4 where $p_T = 200$ GeV/c where derived for $R = 0.8$) and an appraised uncorrected pseudojet energy scale with PF of 0.9 that was multiplied with the initial value.
- If either the angular distance of the parents $dR_{ij}$ in relation to the transverse momentum of the initial pseudojet that is to be declustered $p_{T,A}$ is below $dR_{ij} < 0.4 - 0.0004*p_{T,A}$ (W-method, see chapter 7.3.3) or in adjacent hadronic calorimeter cells according to the “modified adjacency” method of the TopJetAlgorithm, declustering of that pseudojet is stopped.
- A pseudojet considered for declustering will undergo a splitting by the “has parents”-routine. If the parent with less momentum exceeds 10%
of the initial pseudojet momentum, it will be considered a subjet. If it is soft, the soft parent will be discarded (“Trimming”, see chapter 7.4.2) and the hard parent will undergo another declustering. This implies an effective requirement of $p_{T,min} > 18 \text{ GeV/c}$ for subjets.

- After all further declustering attempts have failed, filtering takes place by reclustering each subjet $i$, that was declustered out of pseudojet $A$ together with subjet $j$, with the new jetsize $R = \min(0.8, dR_{ij}/2)$ and the anti-kt algorithm. In rare cases subjets with very few constituents can be reclustered into single constituents by this procedure, which is prevented by a safeguard: If a filtered subjet has less than 75% of the former subjet transverse momentum, the filtering will be omitted. In these cases, the subjets are collimated by construction, so the impact of this procedure seems to be negligible.

- If no substructure is found in a fatjet, filtering still takes place and the filtered fatjet is redefined as the single subjet of the fatjet.

- After filtering, L2L3 jet corrections are applied on all subjets. L2L3 residual jet corrections are additionally applied on all subjets on data.

Subjets typically have smaller and variable sizes. They feature a lot higher energies than normal jets, reducing the relative error of the calorimeter significantly (see Chapter 4.2). There are no special jet corrections for subjets yet. The corrections for anti-kt with a jetsize of $R = 0.5$ are applied. It is assumed that the response variations depending on $\eta$ and $p_T$ are due to the detector and therefore the same for any energy deposition. The dependence on jet algorithm is a dependence of shape. The average number of calorimeter cells with a relative error dependant on the $p_T$ of the jet in a given $\eta$-bin is anti-proportional to the relative error of the energy measurement of the whole jet. The less calorimeter cells are used and the higher the energy deposited is, the less relative error is expected for the measurement. By this logic, any jet correction derived on any algorithm will be an improvement, but smaller jet sizes would need more bins in $\eta$ for an appropriate correction. The jet corrections applied are expected to shift the subjet energy scale close to unity.

Matching in the following chapters is done by:

- Only generated particles with particle status “3” are considered in the matching.

- All generated lighter than top quarks are assigned to either the top quark or antitop quark of the Z'-decay by minimal angular distance.
• The generated W bosons are assigned to the top quarks by charge.

• If only one bottom quark is assigned to a top quark, the leftover up-type and down-type quarks are assigned to the W boson corresponding to the top the quarks were previously assigned to.

• If two bottom quarks are assigned to a top quark, the one closer to the W boson of the same top quark in angular distance is assigned to that W boson. The left bottom quark is assigned to the top as b-jet.

• Once all generated partons are assigned, from each parton the closest subjet in angular distance is sought and then matched to that generated parton.

• A “hemisphere” in any process in this thesis is defined by either being matched to a generated parton assigned to the top quark or the antitop quark.
8.1.1 Combinations of filtering, trimming and pruning

Each of the improvements on declustering presented in this chapter is not exclusive and can be combined with another procedure. The use of pruning, trimming and filtering in combination has been studied by [37] and shown to be beneficial.

The procedure of declustering used in this thesis requires trimming. The decision if a parent of a fatjet is substructure or not, by appraising if it is hard or soft, is trimming of the second type and will be used with a ratio of $X > 10\%$. All graphs shown will therefore be labeled with trimming (“T”).

The effects of additionally using filtering (“F”), pruning (“P”) or even both will also be shown and discussed.

![Multiplicities of different jet improvements per hemisphere](image)

Figure 43: The mean multiplicity of subjets per generated top-$p_t$ on parton level for the two leading fatjets per $Z'$ event with $M_{Z'} = 3$ TeV/$c^2$ and $\Gamma_{Z'} = 30$ GeV/$c^2$ for fully hadronic top decays is shown. Different algorithmic improvements are displayed. “T” labels trimming, “F” labels filtering and “P” labels pruning.

The multiplicity in Fig. 43 shows the expected behaviour. At low transverse momenta of the top, jets are lost due to requirements of a minimal transverse momentum and a decay in the barrel. The ideal multiplicity should be three for hadronic top decays and is reached at $p_{T,t}^{gen} \approx 500$, but for higher energies additional subjets due to FSR become increasingly likely.

Filtering has no impact on the multiplicity by design. The only difference is in the use of pruning or not. Pruning was originally introduced for searches for electroweak bosons. In a three-body decay the b-jet is the most likely jet
to be pruned away. This is visible in the mass-distributions in Fig. 44, where less top quarks and more W bosons are be found per peak, when pruning is introduced.

\[ \text{Mass of } \Sigma_{\text{Subjets}} \text{ for different algorithmic improvements} \]

Figure 44: The masses of the summed four-momenta of the subjets \( M_\Sigma \) in the two leading fatjets per event is shown for fully hadronic top decays of \( Z' \) with \( M_{Z'} = 3 \text{ TeV/c}^2 \) and \( \Gamma_{Z'} = 30 \text{ GeV/c}^2 \). The ideal peak positions are shown by vertical lines at \( M_W \) and \( M_t \). “T” labels trimming, “F” labels filtering and “P” labels pruning.

For every combination of algorithmic improvements a peak around \( M_W = 80.4 \text{ GeV/c}^2 \) and \( M_t = 172.5 \text{ GeV/c}^2 \) is clearly visible. The invariant mass distributions of the recombinations of all subjets in a given fatjet \( M_\Sigma \) show the predicted behaviours. For the application of filtering shifts the peaks to lower masses but decreases the peak width. Pruning decreases subjet multiplicity. Less statistics for top quarks is the direct consequence as is the improvement on the W-peak where former top events migrate. This may still be outbalanced by background reduction, but for this thesis, pruning will not be used for the analysis part.

The best peak position, leanest gaussian and lowest tail is produced by using all three methods in combination. This is done at the cost of losing signal efficiency. These results confirm the findings in [37].

8.1.2 Correlation of transverse momentum of fatjets and subjets

Trimming and filtering serve as counter-measures for PU and UE. Their purpose is to remove energy from these sources to lay bare the substructure.
Figure 45: Correlation between transverse momentum of fatjet and the fourvector-sum of their constituent subjets for the two leading fatjets in fully hadronic decays of top quarks out of a $Z'$-sample with $M_{Z'} \approx 3$ TeV/$c^2$ and $\Gamma_{Z'}$.

The effect of trimming plus filtering is visible in Fig. 45. In many cases, fatjets are corrected to lower values of $p_T$ in the fourvector-sum of their subjets. In rare cases, a correction to higher values takes place by the jet correction, but this effect is smaller than the migration to lower energies.

This algorithmic correction can remove up to a half of the fatjet transverse momentum.

8.1.3 Correction of multiplicity

Declustering by the W-method (see Chapter 7.3.3) is done to obtain a number of subjets that corresponds well to the expected number of partons. If a highly boosted hadronic top decay produces a b-jet and two quarkjets, it should be possible to see these represented by reconstructed subjets, unless they are out of the acceptance of the detector.

This is an issue for all analyses with decay-chains. For QCD as a common background, every single emission of a gluon adds another factor of $\alpha_S$ to the matrix element, making processes the more unlikely the more jets are expected. Multijet events are suppressed in QCD. Boosted multijet events are additionally suppressed due to the high energies needed for such processes.

The subjet multiplicity is therefore a valuable distinction between a possible signal producing boosted jets and QCD.
Figure 46: Mean jet multiplicity of different jet types, binned in transverse momentum of a top generator parton for fully hadronic top decays out of a $Z'$ with $M_{Z'} = 3 \text{ TeV}/c^2$ and $\Gamma_{Z'} = 30 \text{ GeV}/c^2$. A cut on transverse momentum of 50 GeV/c and an $\eta$-cut of 2.5 was applied to the generated top quarks. The blue area shows the variance of the generator expectation. Subjets correspond in this plot to the number of subjets inside fatjets with substructure.

Figure 46 shows that at low transverse momentum for the generator top quark for all algorithms and the generator expectation, the multiplicity rises with $p_{T,\text{top}}^{\text{gen}}$ since less jets will be lost in the beamline or due to the applied $\eta$-cut. At 200 GeV the curve for fatjets flattens, indicating that collimated partons are merging into single fatjets. The same effect is seen for the ak5-jets at 320 GeV. This fits with the calculations performed in Chapter 6.

Once the top is sufficiently boosted for the respective jet size, the seen multiplicity for conventional jets drops steadily.

The number of subjets is not affected and surpasses the three expected subjets at 1.1 TeV. This can be explained by the higher likelihood of additional boosted splittings at very high energies, since the phasespace is very...
large. This FSR-effect would also compensate some of the multiplicity-loss of the other algorithms, if it really is responsible.

8.1.4 Transverse momentum response

Note that no cuts on $|\eta|$ and $p_T,\text{min}$ were applied to the generated partons for the response plots. For the matching all subjets with a $p_T,\text{min}$ of 50 GeV/c out of all fatjets were taken into a consideration. Only if the event is decaying hadronically and three different matches are found per generated top parton, the event will be counted in the following figures.

This is done to show the accuracy of the representation of subjets at the generatorlevel without mismatching errors or double counting. A Hadron level with dedicated generated subjet collections are a possible improvement for this kind of study in the future.

All plots in this chapter were normalized to an integral of one. The response is defined as $\frac{p_T,\text{measured}}{p_T,\text{generated}}$ and measured to investigate the subjet energy resolution. The matching quality is appraised by the angular distance $dR_{\text{matching}}$ between matched generator parton and found subjet.

Figure 47: Normalized correlation of $p_T$-response and angular distance between generated parton and matched subjet for the b-jet of fully hadronic top decays.
Figure 48: Normalized response of the generated bottom parton to the matched subjet of fully hadronic top decays.

Figure 49: Normalized angular distance between the generated bottom parton and the closest subjet of fully hadronic top decays.

A bad matching leads to a bad response, as shown in Fig. 47 for the bottom quark. Most events are nicely centered at a value a bit below unity at a very close matching-distance.

Collimated jets at very high energies facilitate a lower relative calorimeter error and less corrections, but still the response is below unity, as seen in figure 48. The matching works well for a large part of the events, but a substantial tail can also be seen for the matching of the b-jet in Fig. 49. These bottom quarks were either out of acceptance or not resolvable.
The Figures 50, 51 and 51 show the same quantities already shown for the b-jet, but for the up-type quarks out of the W decay.

Figure 50: Normalized correlation of $p_T$-response and angular distance between up-type (u- or c-flavour) genparticle and matched subjet of fully hadronic top decays out of the W decay.

The difference in Fig. 50 to the b-jet-correlation is a better shaped response. Most events lie at close distances and a bit too low responses. Some events have a very high response.

Figure 51: Normalized response of the up-type (u- or c-flavour) generated parton out of the W decay to the matched subjet of fully hadronic top decays.

The response of this light flavour quark is more gaussian-shaped than the
response of the b quark. The tendency towards a bit too low $p_T$ is an effect of trimming and filtering.

Figure 52: Normalized angular distance of the up-type generated parton out of the W decay to the closest subjet of fully hadronic top decays.

The Figures 50, 51 and 51 show the same quantities already shown for the b-jet and up-type quarks, but for the down-type quarks out of the W decay.

Figure 53: Normalized correlation of $p_T$-response and angular distance between down-type (d- or s-flavour) generated parton and matched subjet for the second q-jet of fully hadronic top decays out of the W decay.

The correlation in Fig. 53 of the down-type quark-jet is almost identical
to the up-type.

![Figure 54: Normalized response of the down-type generated parton out of the W decay to the matched subjet of fully hadronic top decays.](image)

The response for the down-type quarks’ jets in Fig. 54 is almost at unity. The different responses of the heavier flavours reflect in the different responses of up- and down-type quarks in Fig. 54 and Fig. 51 due to the differently shaped response of the heavier charmed quark.

![Figure 55: Normalized angular distance of the down-type generated parton out of the W decay to the matched subjet of fully hadronic top decays.](image)

Fig. 55 shows that matching works fine for the down-type-jets.
Figure 56: Responses for different flavours of the jets out of a hadronically decaying W. PdgId 1 is down-, 2 is up-, 3 is strange- and 4 is charmed-flavour. Negative values label antiquarks.

Fig. 56 demonstrates the differences in response for heavier flavours compared to light flavours. Filtered and trimmed subjets, found by the W-method, are reliable jet-representations. Occasional mismatchings, not found jets or overcorrected jets will occur, which is normal for any kind of jet reconstruction method.
8.1.5 Mass reconstruction

The anti-kt algorithm was introduced for LHC for its accurate representation of jet energy and mass and resistance to contaminations due to high activity in the detector, typical for hadron colliders.

Fatjets are not able to compete in this, since they will accumulate more PU and UE. Any subjet-algorithm needs to correct for that fact.

Figure 57: Dotted lines show $M_W$ and $M_t$ in the simulation. Top decays from a $Z'$ with $M_{Z'} = 3$ TeV/c$^2$ and $\Gamma_{Z'} = 30$ GeV/c$^2$. Masses of the two leading jets reconstructed with anti-kt with $dR = 0.5$, Cambridge-Aachen fatjets with $dR = 0.8$ as well as the sum of subjet-fourvectors inside those fatjets, labeled as Subjet, are shown for $\eta < 2.5$ with $p_{T,min} > 50$ GeV/c.

Figure 57 shows three peaks for each algorithm. The first peak consists of single jets, the second peak is close to $M_W$ and the third peak to $M_t$. The peak-heights and -positions differ slightly for the algorithms. There is a pronounced tail to high masses of the fatjet is significantly reduced by the subjet algorithm. The width of the top-peak is about similar for all applied algorithms.

The main improvement of using substructure is a correlation of subjet multiplicity and reconstructed mass that is explained in the next chapter.
In Fig. 58, fatjets of up to 300 GeV/c^2 migrate close to the expected values of $M_W$ and $M_t$ using the subjet algorithm. The effects of trimming and filtering in the jets are thus substantial.
8.1.6 Correlation of subjet multiplicity and reconstructed mass

The main improvement by using subjets is the property of getting the expected multiplicity out of boosted topologies inscrutable for conventional jet algorithms. This can be shown by plotting the masses of the sum of the subjet fourvectors for different numbers of subjets: e.g. if three subjets are found within a fatjet, it would be the three subjet mass. For a top-event a mass of 172.5 GeV/c\(^2\) would be a sensible expectation.

Figure 59: Mass reconstructed from the sum of the fourvectors of subjets out of two leading fatjets in fully hadronic top decay, sorted by subjet multiplicity per fatjet. Plotted for Z' decays with \(M_{Z'} = 3\) TeV/c\(^2\) and \(\Gamma_{Z'} = 30\) GeV/c\(^2\).

Figure 59 demonstrates the correlation between subjet multiplicity and mass. Two subjets often result in the W-peak plus a tail towards the top-peak. The tail is the product of incomplete declustering, where one subjet was not reconstructed. This can occur naturally, when the b-jet and one of the quarkjets are close to each other.

The three subjets mass distribution leads to a nice top-peak, as expected.

Four subjets produce on the one hand a top-peak, but on the other hand a pronounced tail to high masses. Five Subjets very rare, as far as the simulation suggests.
8.1.7 Graphical declustering

To show the effects of the declustering and the improvement-algorithms directly, a full declustering of the jet clustered in Chapter 3.1.5 will be shown in this chapter.

An arbitrary $\eta$-$\phi$-binning is used and for every picture all constituents’ $p_T$ of the pseudo-jet are stacked. The $p_T$-scale is chosen logarithmically to show soft outlying clusters, which should be removed by either trimming or filtering.

Figure 60: Display of one fatjet reconstructed with Cambridge-Aachen and a jetsize of $R = 0.8$. No jet corrections were applied. Shown is a highly boosted top with $p_T = 945$ GeV/c.

Figure 60 shows a top-decay merged into a single fatjet reconstructed with Cambridge-Aachen. The fatjet consists out of a few very hard clusters and lots of soft outlyers. Some structure can be seen, at least two jets are distinguishable by eye.
Figure 61: First declustering of the fatjet of Fig. 60. Two subjets were found.

In the first step of the declustering in Fig. 61, the clusters with the largest angular distances to the jet center have been removed. It was further zoomed in for better visibility. No noticeable effect in mass or transverse impulse can be accounted for.

Figure 62: Second declustering of the red subjet in Fig. 61. A total of three subjets was found.
Soft contamination at $\phi \approx 0.7$ has been removed by trimming in Fig. 62. Three hard clusters remain with only marginal remaining soft contamination of the green subjet.

The mass has been corrected in this case close to the top quark mass and 40 GeV/c of transverse impulse were removed by trimming.

No further declustering happens, since the W-method (see Chapter 7.3.3) becomes effective and vetoes any further attempt of declustering.

![Figure 63: The subjets of Fig. 62 after application of filtering (see Chapter 7.4.3) and jet corrections (see Chapter 9.1).](image)

As a final step, the filtering takes place. Half the distance of separation has been taken as new jet size for an clustering of the subjets with anti-kt. The last soft contaminations have been removed in Fig. 63.

Afterwards ak5-jet corrections were used on the subjets, also in this step. Filtering and jet corrections balance themselves out, while only the most collimated structures remain after the declustering-process.

The final result is the resolution of a simulated boosted hadronical top decay. In this case with a mass very close to the toppeak.
9 Data analysis

An exemplary analysis searching for a possible Z' of 3 TeV/c^2 mass and 30 GeV/c^2 width with an arbitrary cross section of 0.35 pb (constant for all shown Z' masses in this thesis) is shown to quantify the possible improvements by the use of subjets in comparison to conventional methods. For a better cross section estimate see section 8.2. As a background, only QCD has been studied. Top decays from SM sources are expected to have vanishing cross sections for generating boosted structures at the LHC. For details regarding the MC samples see Chapter 5.2.

For the analysis data taken by the CMS experiment was used. Specifically samples and runs reconstructed and reconstucted from the start of CMS data taking 2010 up to beginning of August 2011. The “HLT_DijetAve*” triggers were used and the integrated luminosity in this thesis amounts to \( L^{-1} = 1.96 \text{ fb}^{-1} \). The MC samples are reweighted to the expectation for 1.96 fb\(^{-1}\) and compared to data corresponding to the same integrated luminosity.

The following variables will be used in the analysis:

1. \( H_{T,\text{fat}} = p_{T,\text{fat}1} + p_{T,\text{fat}2} \) : the scalar sum of transverse momenta of the two leading fatjets.
2. \( M_{\text{fat}} \) : the invariant mass of a fatjet.
3. \( |\Delta \eta| = |\eta_1 - \eta_2| \) : the absolute difference of pseudorapidity of the two leading fatjets.
4. \( |\Sigma \eta| = |\eta_1 + \eta_2| \) : the absolute sum of pseudorapidity of the two leading fatjets.
5. \( |\Delta \phi| = |\phi_1 - \phi_2| \) : the absolute difference of azimuthal angle of the two leading fatjets.
6. \( N_{\text{Sub}} \) : the number of subjets inside one fatjet.
7. \( M_{2,\text{Sub}} \) : the minimal invariant mass out of all combinations of two subjets possible in one fatjet.
8. \( M_\Sigma \) : the invariant of the combination of all subjets in one fatjet.
9. \( M^{\Sigma} \) : the invariant mass of the combination of all subjets in both leading fatjets.
9.1 Basic selection cuts with conventional methods

No substructure analysis will be used in this chapter. Jets are defined by the Cambridge-Aachen algorithm with a jetsize of $R = 0.8$, representing an analysis with conventional methods.

The fatjets were calculated in the range $p_{T,min} > 50$ GeV/c in $\eta < 2.5$. On data, a “DijetAverage*-trigger is used. Also vertex- and scraping filters were applied with standard values. Due to technical reasons, the subjet collections replace the constituent collections of the fatjets in the current program setup, no jet identification could be applied and typical noise cut variables are inaccessible. The standard noise cuts are being replaced by quality cuts on $\eta$- and $\phi$-variables as well as a very high transverse momentum cut. Each plot in chapter 10 is shown after all cuts on the variables shown before were applied.

Figure 64 shows the distributions of signal, background and data for $H_{T,fat}$, which is the scalar sum of transverse momentum of the leading two fatjets in the detector.

![Figure 64: Distribution of the scalar sum of transverse momenta of the leading two fatjets for 1.96 fb$^{-1}$. The vertical line marks the later cut position.](image)

No noise cuts were applied yet, therefore the expected distribution in Fig. 64 differs from the data at very high masses. The low $H_{T,fat}$ region is affected by the “Dijet370U”-trigger and is removed by a high cut value of $H_{T,fat} > 1500$ GeV/c. In an intermediate region of $H_{T,fat} \approx 1.5$ to 2.5 TeV/c data and QCD simulation agree reasonably considering the large uncertain-
ties of the QCD calculation.

The tail of the Z' signal to low $H_{T,fat}$ is produced by two different circumstances:

1. The very high mass Z' is almost exclusively produced by quark-antiquark-annihilation. The parton distribution function for this process is suppressed, leading to a second peak in the mass distribution before the real peak (see Fig. 3) and therefore to a tail in the $H_{T,fat}$-distribution of the signal.

2. The sample is inclusive, containing hadronic and leptonic top decays. Leptonic top decays contain missing energy due to the neutrino in the decay, producing a tail in the $H_{T,fat}$-distribution of the signal.

For every step of the cutflow, $S/B$, $S/\sqrt{S+B}$, signal efficiency and background efficiency are shown to motivate each cut, where S is the signal from the Z' simulation and B the background consisting of QCD simulation.

Figure 65: Shown are $S/B$ and $S/\sqrt{S+B}$ on the left axis and signal and background efficiencies on the right axis. Distributions as function of a cut value for the scalar sum of transverse momenta for the leading two fatjets for 1.96 fb$^{-1}$ is shown. Signal S denotes a Z' with 3 TeV/c$^2$ mass and 30 GeV/c width as signal. Background B denotes the QCD simulation.

Figure 65 shows the effect of cutting on $H_{T,fat}$. A value of $H_{T,fat} > 1.5$ TeV/c$^2$ was chosen to still be sensitive on Z' masses of 2 and 4 TeV/c$^2$. 

90
The invariant mass of the fatjet is expected to be close to the topmass for a possible signal. Cuts on the fatjetmass are the next possibility to improve the analysis.

![Figure 66](image.png)

Figure 66: Distribution of the invariant jet mass of the leading two fatjets for 1.96 fb$^{-1}$. The vertical line marks the later cut position.

Still without noise cuts, the invariant mass of the fatjets in Fig. 66 is ill modeled for low and high masses.

The signal distributions shows a peak around the top quark mass and a high mass tail.
Figure 67: Shown are $\frac{S}{B}$ and $\frac{S}{\sqrt{S+B}}$ on the left axis and signal and background efficiencies on the right axis. Distributions as function of a lower cut value for the invariant jet mass of the leading two fatjets for 1.96 fb$^{-1}$ is shown. Signal S denotes a Z’ with 3 TeV/c$^2$ mass and 30 GeV/c width as signal. Background B denotes the QCD simulation.

Figure 67 demonstrates the impact of a lower mass cut. $M_{fat} > 120$ GeV/c$^2$ was chosen to not cut away much of the signal while still significantly reducing the acceptance.
Figure 68: Shown are $\frac{S}{B}$ and $\frac{S}{\sqrt{S+B}}$ on the left axis and signal and background efficiencies on the right axis. Distributions as function of an upper cut value for the invariant jet mass of the leading two fatjets for 1.96 fb$^{-1}$ is shown. Signal S denotes a $Z'$ with 3 TeV/c$^2$ mass and 30 GeV/c width as signal. Background B denotes the QCD simulation.

Figure 68 shows the effect of an upper mass cut. The signal has a high mass-tail. Signal efficiency drops more steeply than background efficiency on any possible upper mass cut. No upper mass cut was used.

Cuts on the eta- and phi-distributions are used to remove noise events on data.
Figure 69: Distribution of the absolute difference in $\eta$ of the leading two fatjets for 1.96 fb$^{-1}$. The vertical line marks the later cut position.

Figure 69 shows that a few possible noise events are removed by a cut on $|\Delta \eta|$. This is the first quality cut. The modeling of the $\eta$-distribution is the first, where data points and expectation roughly agree.

For the quality cuts the cut motivations are ommitted, since there is virtually no effect of these cuts on signal or background efficiencies. A $|\Delta \eta| < 2.5$-cut removes noise. Noise events are expected to be randomly distributed in $\Delta \eta$ and to be effected by this cut.
Figure 70: Distribution of the absolute of the leading two fatjets’ summation in $\eta$ for 1.96 fb$^{-1}$. The vertical line marks the later cut position.

In Fig. 70 it is demonstrated that some possible noise events can be affected by cutting on $|\Sigma \eta|$. A $|\Sigma \eta| < 1.5$-cut is a quality cut. Noise events are expected to be randomly distributed in $\Sigma \eta$ and to be effected by this cut.

Figure 71: Distribution of the absolute of the leading two fatjets’ difference in $\phi$ for 1.96 fb$^{-1}$. The vertical line marks the later cut position.
In Fig. 71 it is shown that a cut on $|\Delta \phi|$ is very effective on possible noise events. A $|\Delta \phi| > 2$-cut removes noise. Noise events are expected to be randomly distributed in $\Delta \phi$ and to be effected by this cut.

Table 1 shows the cutflow so far.

<table>
<thead>
<tr>
<th>Cut</th>
<th>QCD</th>
<th>$Z'$_2000_20</th>
<th>$Z'$_3000_30</th>
<th>$Z'$_4000_40</th>
<th>data</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC reweighted</td>
<td>3.84E+10</td>
<td>534</td>
<td>678</td>
<td>692</td>
<td>/</td>
</tr>
<tr>
<td>2 jets, $p_T &gt; 50$ GeV/c, $</td>
<td>\eta</td>
<td>&lt; 2.5$</td>
<td>3.84E+010</td>
<td>530</td>
<td>672</td>
</tr>
<tr>
<td>$H_{T,fat} &gt; 1500$ GeV/c</td>
<td>12020</td>
<td>118</td>
<td>274</td>
<td>62</td>
<td>7889</td>
</tr>
<tr>
<td>$M_{fat} &gt; 120$ GeV/c$^2$</td>
<td>1848</td>
<td>98</td>
<td>210</td>
<td>50</td>
<td>1626</td>
</tr>
<tr>
<td>$</td>
<td>\Delta \eta</td>
<td>&lt; 2.5$</td>
<td>1828</td>
<td>98</td>
<td>210</td>
</tr>
<tr>
<td>$</td>
<td>\Sigma \eta</td>
<td>&lt; 1.5$</td>
<td>1696</td>
<td>92</td>
<td>210</td>
</tr>
<tr>
<td>$</td>
<td>\Delta \phi</td>
<td>&gt; 2$</td>
<td>1696</td>
<td>92</td>
<td>210</td>
</tr>
</tbody>
</table>

Table 2: Cutflow for fatjets for 2 fb$^{-1}$. For the $Z'$ notation the first number is the mass of the $Z'$, the second number is the width the resonance.

The final S/B-ratio after an analysis with conservative methods on boosted tops is $\approx 12$% and a $S/\sqrt{S+B}$-ratio of $\approx 3.4$. Such an analysis is applicable with a good substraction of background or can be improved by using substructure-methods. The signal to background ratios for the different masspoints after all cuts are shown in Tab. 9.1.

<table>
<thead>
<tr>
<th>S/B $Z'$_2000_20</th>
<th>S/B $Z'$_3000_30</th>
<th>S/B $Z'$_4000_40</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.12</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 3: Signal to background ratios after basic selection with fatjets for three different $Z'$ masses. For the $Z'$ notation the first number is the mass of the $Z'$, the second number is the width the resonance.

A signal to background ratio of 12% with an unsound modeling of the QCD background is not sufficient.
9.2 Advanced selection cuts with substructure methods

This chapter will introduce cuts on the substructure of the fatjets on top of the cutflow of Chapter 9.2.1.

The subjet multiplicity is a possible discriminator between QCD and signal. Before it is used, it has to be shown that it is accurately modeled. True substructure in QCD is a higher order process with hard gluon emissions in the matrix element and not modeled in leading-order Pythia6-MC.

![Distribution of subjet multiplicity inside the leading fatjet for 1.96 fb⁻¹. The vertical line marks the later cut position.](image)

Figure 72: Distribution of subjet multiplicity inside the leading fatjet for 1.96 fb⁻¹. The vertical line marks the later cut position.

The signal distribution in Fig. 72 shows that multiplicities of two to four subjets is the main signal region. Two subjets correspond to leptonic decays, three subjets to the highly boosted hadronic decays and four to highly boosted hadronic decays with final state radiation.
Figures 72 and 73 show that noise events still persist in higher subjet multiplicity bins. The one and two subjets-bins are not well modeled. The three, four and five subjets-bins appear to be well modeled.

An excess is seen at very high subjet multiplicities. This could be traced to monster events with tens of TeV/c transverse momentum, filling the whole detector. These events will be discarded by the cuts on masses of subjets.
Figure 74: Shown are $S/B$ in the upper left plot and $S/\sqrt{S+B}$ in the upper right plot. Distributions as correlation of the number of subjets in the leading two fatjets for 1.96 fb$^{-1}$ are shown. Signal S denotes a Z' with 3 TeV/c$^2$ mass and 30 GeV/c width as signal. Background B denotes the QCD simulation.

Figure 74 shows the effect of cuts on subjet multiplicity of first and second fatjet. A cut of $N_{Sub} \geq 3$ was chosen. The background efficiency does not drop to zero, but by two orders of magnitude.

The subjet multiplicity cut also rejects semileptonic und dileptonic decays.

The minimal mass out of all possible recombinations of two subjets inside of a fatjet is called as $M_{2,Sub}$. This quantitiy should be close to $M_W$ for signal events.
Figure 75: Distribution of $M_{2,\text{sub}}$ in the two leading fatjets for 1.96 fb$^{-1}$. The vertical lines marks the later cut positions.

In Fig. 75 it is shown that the signal now exceeds the background expectation in the region close to $M_W$. No such excess has been seen in data and the modeling works reasonably well for the variable $M_{2,\text{Sub}}$. 
Figure 76: Shown are $\frac{S}{B}$ and $\frac{S}{\sqrt{S+B}}$ on the left axis and signal and background efficiencies on the right axis. Distributions as function of a lower cut value for $M_{2,\text{sub}}$ for 1.96 fb$^{-1}$ is shown. Signal S denotes a $Z'$ with 3 TeV/c$^2$ mass and 30 GeV/c width as signal. Background B denotes the QCD simulation.

Figure 76 demonstrates the effect of a lower cut on the minimal recombined mass of two subjets. The goal is to make a counting-experiment with almost no background.

A cut on $M_{2,\text{sub}} > 50$ GeV/c$^2$ has been used.
Figure 77: Shown are $\frac{S}{B}$ and $\frac{S}{\sqrt{S+B}}$ on the left axis and signal and background efficiencies on the right axis. Distributions as function of an upper cut value for $M_{2,\text{sub}}$ for 1.96 fb$^{-1}$ is shown. Signal S denotes a $Z'$ with 3 TeV/c$^2$ mass and 30 GeV/c width as signal. Background B denotes the QCD simulation.

Figure 77 shows the effect of an upper cut on $M_{2,\text{sub}}$. This is used as a further cleaning cut against the high events of high subjet multiplicity in Fig. 72 and Fig. 73.

A cut on $M_{2,\text{sub}} < 100$ GeV/c$^2$ has been used.
Figure 78: Distribution of $M_\Sigma$ in the two leading fatjets for 1.96 fb$^{-1}$. The vertical line marks the later cut positions.

Figure 79: Shown are $\frac{S}{B}$ and $\frac{S}{\sqrt{S+B}}$ on the left axis and signal and background efficiencies on the right axis. Distributions as function of a lower cut value for $M_\Sigma$ for 1.96 fb$^{-1}$ is shown. Signal S denotes a Z' with 3 TeV/c$^2$ mass and 30 GeV/c width as signal. Background B denotes the QCD simulation.

The recombined subjet mass inside a fatjet is another quantity to cut on,
which is not exactly similar to the mass of the fatjet (see Chapter 9.1.4). The declustering with trimming and filtering shifts the invariant mass of the recombined subjets.

Based on Fig. 79 a cut of $M_{\Sigma subjets per fatjet} > 130 \text{ GeV/c}^2$ is used.

Figure 80: Shown are $S/B$ and $S/\sqrt{S+B}$ on the left axis and signal and background efficiencies on the right axis. Distributions as function of an upper cut value for $M_{\Sigma}$ for 1.96 fb$^{-1}$ is shown. Signal S denotes a Z$'$ with 3 TeV/c$^2$ mass and 30 GeV/c width as signal. Background B denotes the QCD simulation.

Figure 80 shows the effect of an upper cut on the recombined invariant mass of all subjets inside a fatjet. A cut on $M_{\Sigma} < 200 \text{ GeV/c}^2$ is used. The cutflow for the inclusion of substructure techniques is shown in Tab. 9.2:
Only a few background events remain in the signal region. A counting experiment can be performed. The comparison of signal to background ratios before and after the substructure techniques is shown in Tab. 9.2:

<table>
<thead>
<tr>
<th>Cut</th>
<th>QCD</th>
<th>$Z’_{2000.20}$</th>
<th>$Z’_{3000.30}$</th>
<th>$Z’_{4000.40}$</th>
<th>data</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC reweighted</td>
<td>3.84E+10</td>
<td>534</td>
<td>678</td>
<td>692</td>
<td>/</td>
</tr>
<tr>
<td>2 jets, $p_T &gt; 50$ GeV/c, $</td>
<td>\eta</td>
<td>&lt; 2.5$</td>
<td>3.84E+010</td>
<td>530</td>
<td>672</td>
</tr>
<tr>
<td>$H_{T,fat} &gt; 1500$ GeV/c</td>
<td>12020</td>
<td>118</td>
<td>274</td>
<td>62</td>
<td>7889</td>
</tr>
<tr>
<td>$M_{fat} &gt; 120$ GeV/c$^2$</td>
<td>1848</td>
<td>98</td>
<td>210</td>
<td>50</td>
<td>1626</td>
</tr>
<tr>
<td>$</td>
<td>\Delta\eta</td>
<td>&lt; 2.5$</td>
<td>1828</td>
<td>98</td>
<td>210</td>
</tr>
<tr>
<td>$</td>
<td>\Sigma\eta</td>
<td>&lt; 1.5$</td>
<td>1696</td>
<td>92</td>
<td>210</td>
</tr>
<tr>
<td>$</td>
<td>\Delta\phi</td>
<td>&gt; 2$</td>
<td>1696</td>
<td>92</td>
<td>210</td>
</tr>
<tr>
<td>$N_{sub} \geq 3$</td>
<td>214</td>
<td>50</td>
<td>118</td>
<td>30</td>
<td>249</td>
</tr>
<tr>
<td>$50 &lt; M_{2,sub} &lt; 100$ GeV/c$^2$</td>
<td>60</td>
<td>36</td>
<td>68</td>
<td>14</td>
<td>70</td>
</tr>
<tr>
<td>$130 &lt; M_{\Sigma subjets per fatjet} &lt; 200$ GeV/c$^2$</td>
<td>8</td>
<td>24</td>
<td>34</td>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 4: Cutflow for fatjets and subjets for 2 fb$^{-1}$.

Table 5: Signal to background ratios after advanced selection with subjets (second line) in comparison to the basic selection (first line) for three different $Z’$ masses.

<table>
<thead>
<tr>
<th>S/B $Z’_{2000.20}$</th>
<th>S/B $Z’_{3000.30}$</th>
<th>S/B $Z’_{4000.40}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.12</td>
<td>0.03</td>
</tr>
<tr>
<td>3</td>
<td>4.25</td>
<td>0.75</td>
</tr>
</tbody>
</table>

The improvement relative to the analysis not using substructure is substantial for each masspoint. The analysis has been tuned to the $Z’_{3000.30}$ masspoint. Other $Z’$ masses are only shown for comparison. For a full analysis of a range of $Z’$ masses, a custom cutflow for each $Z’$ masspoint would have to be used.
9.3 Results

A short comparison between expectation and data for the $Z'$ analysis will be shown in this chapter.

On 1.96 fb$^{-1}$ of data with an assumed error of 6%, $8.4 \pm 1.6(MC \text{stat})$ events were expected for QCD.

Assuming a cross section of 0.35 pb for three $Z'$-scenarios of 1, 2 and 3 TeV/$c^2$ mass with 1% width, the results will be shown.

![Graph showing $\bar{t}t$ invariant mass reconstruction](image)

Figure 81: Distribution of the invariant mass of the recombination of all subjets inside the two leading fatjets for 1.96 fb$^{-1}$. Comparison to QCD background and three different $Z'$ masses are shown. All $Z'$ scenarios are stacked on QCD only and not on each other.

Ten events were found in data which is in agreement with a background expectation, as seen in Fig. 81.
Figure 82: A modified frequentist CLs Limit at 95 % confidence level has been calculated for three different Z’ masses.

A modified frequentist CLs Limit at 95 % confidence level has been calculated for three masspoints with efficiencies of:

- Z’ of 2 TeV/c²: 0.0043(5)
- Z’ of 3 TeV/c²: 0.05 ± 0.0005
- Z’ of 4 TeV/c²: 0.0094(2)

results in the limits in Fig. 82. The observed limits are:

- Z’ of 2 TeV/c²: 1.22 pb
- Z’ of 3 TeV/c²: 0.10 pb
- Z’ of 4 TeV/c²: 0.53 pb

The comparison to the CMS fullhadronic analysis on roughly 0.9 fb⁻¹ (see Chapter 8.2) shows four things:

1. A Z’ limit for 4 TeV/c² mass could be set in this analysis, that could not be set in [41].
2. The limit for the 3 TeV/c² masspoint could be improved roughly by a factor of two, corresponding to the difference in luminosity for the analyses.
3. A different cutflow for each masspoint, like in [41], allows for improved limits. The lower mass range might especially profit from the use of not fully boosted topologies for the search.
4. For very high masses, neither topcolour, nor KK-gluon models can be excluded with the current luminosity.
10 Summary and outlook

In this thesis, collimated structures ut of highly boosted top decays have been studied. Different algorithms to resolve the substructure in single jets containing all partons of the top decay have been investigated.

First, the dependence of the angular distance to the boost of three different combinations of top decay products has been studied for highly boosted tops. It has been shown that for transverse momenta in the order of $\approx 300 \text{ GeV}/c$ substructure becomes an issue for analyses with top quarks and W bosons.

Second a thorough study on algorithms for resolving substructure inside merged jets has been performed. Several improvements and combinations thereof have been investigated. Pruning, filtering and trimming have been tested in detail. Pruning has been found to be advantageous for studies of W bosons while disadvantageous for studies of more complex substructures in this thesis. Filtering has been shown to have an impact on the subjet energy scale. Migration from high mass tail top quarks in fatjets to the expected mass values has been observed due to filtering and trimming.

It has been shown that substructure algorithms are capable of efficiently finding the expected number of partons in a jet and that the finding of the expected jet multiplicity corresponds to a good quality of reconstruction.

Third an extensive illustration of the processes of clustering and declustering highly collimated structures inside single jets was shown in this thesis. The complete clustering in every step and the complete declustering were presented.

Fourth an analysis on very massive $Z'$ particles of $2 - 4 \text{ TeV}/c^2$ mass was done to quantify the improvement of using substructure methods in comparison to conventional use of jet algorithms. Substantial improvements in the signal to background ratio have been demonstrated and a limit on a $Z'$ of $3 \text{ TeV}/c^2$ mass of a cross section times the branching ratio into tops of $0.1 \text{ pb}$ could be achieved with a signal efficiency of 5%.

The methods used in this thesis are possibly applicable to several exotic scenarios at very high energies. The higher the difference in mass between SM particles and particles of possible new physics, the higher boost for SM particles is to be expected. Since no new physics has been found so far, substructure methods become increasingly attractive for searches in very high energy regions.

Substructure studies could benefit from including hadron level generator information and dedicated subjet data formats that allow for the traditional noise cut variables like number of photodiodes triggered in the HCAL or the electromagnetic energy fraction of the jet to be applied.
Further study is also in order for the behaviour of noise events where large parts of the detector are “lit” and basically an infinite number of subjets can be clustered. Low energy jets consisting only of evenly distributed pileup are also an issue for the algorithms used so far. Close to even distributions are also declustered down to constituent level.

The goal for the far future could be the implementation of subjet algorithms as standard jets to be freely used by any analysis with dedicated jet corrections and a physical object group dedicated solely for this purpose.
# 11 Figures

## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Top fatjet</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>SM particle zoo</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>Generated $M_{Z'}$</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>Generated $p_{T,t}$</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>Generated $p_{T,W}$</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>1D-boost</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>Transverse separation impulse</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>W kinematics peak</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>W kinematics upper limit</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>W kinematics lower limit</td>
<td>23</td>
</tr>
<tr>
<td>11</td>
<td>Top kinematics peak</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>Top kinematics upper limit</td>
<td>25</td>
</tr>
<tr>
<td>13</td>
<td>Top kinematics lower limit</td>
<td>26</td>
</tr>
<tr>
<td>14</td>
<td>Top-W kinematics</td>
<td>27</td>
</tr>
<tr>
<td>15</td>
<td>LHC layout</td>
<td>29</td>
</tr>
<tr>
<td>16</td>
<td>Luminosity 2010</td>
<td>31</td>
</tr>
<tr>
<td>17</td>
<td>Luminosity 2011</td>
<td>32</td>
</tr>
<tr>
<td>18</td>
<td>CMS detector</td>
<td>33</td>
</tr>
<tr>
<td>19</td>
<td>Pixeldetector</td>
<td>35</td>
</tr>
<tr>
<td>20</td>
<td>Tracker readout</td>
<td>36</td>
</tr>
<tr>
<td>21</td>
<td>ECAL</td>
<td>37</td>
</tr>
<tr>
<td>22</td>
<td>HCAL</td>
<td>38</td>
</tr>
<tr>
<td>23</td>
<td>Muon system</td>
<td>39</td>
</tr>
<tr>
<td>24</td>
<td>Substructure</td>
<td>48</td>
</tr>
<tr>
<td>25</td>
<td>Clustering fatjet</td>
<td>49</td>
</tr>
<tr>
<td>26</td>
<td>Clustering step 1</td>
<td>50</td>
</tr>
<tr>
<td>27</td>
<td>Clustering step 2</td>
<td>50</td>
</tr>
<tr>
<td>28</td>
<td>Clustering step 3</td>
<td>51</td>
</tr>
<tr>
<td>29</td>
<td>Clustering step 4</td>
<td>51</td>
</tr>
<tr>
<td>30</td>
<td>Clustering step 5</td>
<td>52</td>
</tr>
<tr>
<td>31</td>
<td>Clustering step 6</td>
<td>52</td>
</tr>
<tr>
<td>32</td>
<td>Clustering step 7</td>
<td>53</td>
</tr>
<tr>
<td>33</td>
<td>Clustering step 8</td>
<td>53</td>
</tr>
<tr>
<td>34</td>
<td>Clustering step 9</td>
<td>54</td>
</tr>
<tr>
<td>35</td>
<td>Clustering step 10</td>
<td>54</td>
</tr>
<tr>
<td>36</td>
<td>Clustering step 11</td>
<td>55</td>
</tr>
<tr>
<td>78</td>
<td>Invariant mass of all subjets recombination in a fatjet</td>
<td>103</td>
</tr>
<tr>
<td>79</td>
<td>Cut motivation, $M_\Sigma &gt; X$</td>
<td>103</td>
</tr>
<tr>
<td>80</td>
<td>Cut motivation, $M_\Sigma &lt; X$</td>
<td>104</td>
</tr>
<tr>
<td>81</td>
<td>Invariant mass of recombination of all subjets in leading two fatjets</td>
<td>106</td>
</tr>
<tr>
<td>82</td>
<td>Bayesian limit setting</td>
<td>107</td>
</tr>
</tbody>
</table>
References


K. Schindl, (Ed.),

[17] [ LHC-B Collaboration ],


[19] G. Dellacasa et al. [ ALICE Collaboration ],


http://web.physik.rwth-aachen.de/ klein/dcms.pdf


[23] [CMS Collaboration], CMS-PAS-QCD-10-013.


[25] G. L. Bayatian et al. [ CMS Collaboration ],

[26] [ CMS Collaboration ], “Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET,”


[30] M. Cacciari,
[hep-ph/0607071].


115

[36] M. Rubin,
[arXiv:0905.2124 [hep-ph]].


[38] J. M. Butterworth, A. R. Davison, M. Rubin, G. P. Salam,
[arXiv:0810.0409 [hep-ph]].


[40] Carsten Magass, https://twiki.cern.ch/twiki/bin/viewauth/CMS/BoostedHiggsAlgorithm

[41] CMS-PAS-EXO-11-006 [ CMS Collaboration ]
12 Acknowledgements

First I want to thank my professor Peter Schleper for his inquisitive questioning and all the numerous advise he’s given me during my thesis. The hours we discussed on the topic of jet substructure and the sometimes hard to grasp definitions were never wasted.

Another thanks goes do my advisor, Dr. Hartmut Stadie, who could always best the beast that ROOT sometimes proves to be and enabled me to learn the intrinsics of jets and C++ in reasonable time.

I also want to thank Dr. Christian Autermann and Dr. Christian Sander for sometimes enlightening discussions either about limit setting or boosted jet kinematics.

Also many thanks to my colleagues, Holger Enderle, Ulla Gebbert, Martin Goerner, Nadja Hühdepohl, Henning Kirschenmann, Friederike Nowak, Eike Schlieckau, Matthias Schröder and Markus Seidel who deserve gratitude for making my thesis mostly a fun time and helping each other with the little problems of everyday programming.