Department of Physics and Astronomy
University of Heidelberg

Bachelor Thesis in Physics
submitted by

Josua Göcking

born in Tübingen (Germany)

2015

Alternative substructure reconstruction for the HEPTopTagger

This Bachelor Thesis has been carried out by Josua Göcking at the
Institute for Theoretical Physics in Heidelberg
under the supervision of
Prof. Dr. Tilman Plehn

Abstract

Top tagging algorithms are crucial tools in the analyses of processes in the Standard Model (SM) and beyond. Improving these algorithms is, hence, of great importance. The HEPTopTagger is a top tagging algorithm using the mass drop (MD) algorithm for reconstructing the substructures of a fat jet. To this alternative approaches like the Soft Drop (SD) and the Mass Jump (MJ) algorithm exist.
The aim of this thesis is to implement these alternatives to the HEPTopTagger, analyze them and decide if it is useful to replace the MD algorithm by one of them. Also an application of MJ to the fully hadronic $t\bar{t}H$ process is done. It turns out replacing the MD algorithm with one of these alternatives is not necessary, since they achieve basically the same results. However, to be able to tag tops with low p_T the MJ algorithm is a useful tool, since this way one is not dependant on the size of the fat jets.

Zusammenfassung

Toptagging-Algorithmen sind ein wichtiges Instrument um Prozesse des Standard Modells (SM) und darüber hinaus zu untersuchen. Diese Algorithmen zu verbessern ist daher von großer Bedeutung. Der HEPTopTagger ist ein Toptagging-Algorithmus, der den mass drop (MD) Algorithmus zur Rekonstruktion von Substrukturen eines Fatjets verwendet. Es existieren aber auch alternative Algorithmen, wie z.B. der Soft Drop (SD) und der Mass Jump (MJ) Algorithmus.
Ziel dieser Arbeit ist es, diese Alternativen zu implementieren, analysieren und festzustellen, ob es sinnvoll ist den MD Algorithmus durch sie zu ersetzen. Zudem wird der MJ Algorithmus auf den voll hadronischen $t\bar{t}H$ Prozess angewandt. Es stellt sich heraus, dass eine Ersetzung des MD Algorithmus mit einem der alternativen Algorithmen nicht notwendig ist, da diese im Wesentlichen dieselben Ergebnisse erzielen. Um jedoch Topquarks zu taggen, die ein niedrigen transversalen Impuls haben, ist es von Nutzen, den MJ Algorithmus zu verwenden, da man so nicht von der Größe der Fatjets abhängig ist.

Contents

1 Introduction and Motivation

The Standard Model of particle physics is one of the greatest success stories in physics of the last decades. Through the prediction of before unknown particles, like the Higgs boson, it has been successfully confirmed. However, since the Standard Model is not able to explain several experimental findings, e.g. neutrino oscillations or dark matter/energy, we know, that it has to be just one aspect of a greater, more comprehensive theory. Therefore, particle physicists are in search of so-called "new physics". In the last decades several theories regarding these topics have been developed and are waiting to be tested.

One way to test the Standard Model or search for physics beyond the Standard Model are collider experiments. In colliders, particles are accelerated to high speeds, close to the speed of light, and are then impacted on each other. In such collisions new particles can be created. However, most of these particles decay before we can measure them. Therefore, strategies to analyze the outcome of the event have to be developed to be able to link the experimental results to the theory.

One important thing here are the tagging algorithms. They allow to reconstruct particles by analyzing their decay products. Practically for every heavy particle, which cannot be directly measured there is a tagging algorithm.

Although the top quark has already been discovered, top tagging algorithms are still of great importance. This is due to the fact, that most processes regarding "new physics", but also interesting Standard Model processes, include top quarks somewhere. Therefore, it is important to be able to reconstruct tops in an event.

There are several problems that occur when tagging top quarks. One problem is, that often the decay products are spread in all directions, which makes it difficult or even impossible to identify them in a QCD busy environment. Also means to improve the tagging efficiency and the quality of reconstruction are always welcome.

The HEPTopTagger is one example for a top tagging algorithm. It aims for the reconstruction of moderately boosted hadronic tops. It uses a substructure analysis based on masses. However, there are, especially in the process of (un)clustering a fat jet, several other methods one can use. Examples are the Soft Drop and the Mass Jump algorithm. Testing if these alternatives yield improvements in tagging efficiency, the quality of reconstruction or other aspects, compared to the so-called mass drop algorithm of the HEPToptagger, is the aim of this study.

We will start by outlining the theoretical backgrounds needed, i.e. the Standard Model and processes of it, which are important for our study. This can be found in section 2.

In section 3, we will describe the physics of colliders and how we can link the results to the theory. There we will also introduce the software which we will use for our analysis.

Section 4 will deal with tagging top quarks. There, we will introduce the HEPTopTagger.

Finally, we will study the alternatives to the used methods in tagging the top quarks in section 5. There we will also have an application of the Mass Jump algorithm on the fully hadronic $t\bar{t}H$ process.

2 Theoretical Background

2.1 The Standard Model

The Standard Model (SM) [1–3] describes elementary particles and their interactions. These particles are considered to be point-like, i.e. they have no spatial extension. One distinguishes these particles by their spin. First of all, there are particles of half-integer spin, so-called fermions. These can again be separated into two groups: Fermions that participate in the strong interaction, namely quarks, and those that don't, which are called leptons. In both of these so-called families, there are 3 generations of particles. For leptons, these generations consist of one particle with electrical charge -1 and its associated neutrino, which is charge- and massless (according to the theory, however the experimental observation of neutrino oscillations indicates, that neutrinos have mass [9]). For quarks, each generation consists of two quarks with electrical charge $+2/3$ and $-1/3$ respectively. All of these particles, quarks and leptons, have an associated anti-particle with opposite charge. Therefore, disregarding differences in electromagnetic and color charge we have 6 leptons and 6 quarks. Next, there are particles with integer spin, so-called bosons. These particles are responsible for the interaction between particles. There are 12 particles that have spin 1: The photon γ, which is responsible for the electromagnetic interaction, the $W^{\pm}$ and the Z, responsible for the weak interaction and 8 gluons g, responsible for the strong interaction.

However one should note, that there are also bosons and fermions, which are not elementary particles. Because of color confinement, quarks cannot exist by themselves, since they carry color charge. Therefore, they have to form color neutral compositions of particles, called hadrons. Hadrons can be either compositions of a quark and an antiquark (mesons) which are bosons or of three quarks/antiquarks to form baryons, which are fermions. The Standard Model also allows for hadrons which consist of more than three quarks like so-called pentaquarks consisting of 5 quarks. Recently the LHCb collaboration at CERN announced [12], that such states were successfully observed.

Lastly, there is the Higgs boson, which carries spin 0 and is therefore a scalar particle. Its existence follows from the Higgs-mechanism which is necessary to describe the electroweak symmetry breaking causing the weak gauge bosons to have mass. If one inserts the Higgs field into the Lagrangian this results in a mass term for the fermions. This mechanism was already established in 1964 by Peter Higgs [4, 5] and Francois Englert [6]. However it was up until July 4, 2012 that the ATLAS and CMS collaborations at CERN could announce

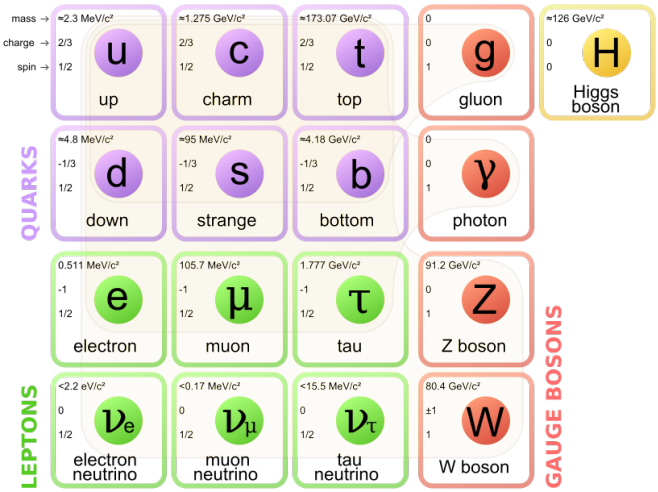

Figure 1: The elementary particles of the Standard Model. Taken from [14].

its discovery. [7, 8].

In Fig. 1, one can see an overview of the previously described elementary particles of the Standard Model.

The SM has been experimentally confirmed by the discovery of all particles which it predicts, e.g. the top quark or the $W^{\pm}$ and Z bosons. However, there are several questions which remain unanswered. For example gravity, since it is extremely weak compared to the other forces, is totally ignored by the SM. Furthermore, the SM can't explain dark matter, dark energy or several other cosmological findings. Also, as already mentioned, the fact that neutrinos do carry mass cannot be explained within the SM.

Therefore, new theories have been developed, which include such aspects. Since the Higgs boson has been observed, the primary goal of collider physics is now to explore these physics beyond the Standard Model (BSM), which is also often referred as "new physics". Most of these BSM theories predict particles, that couple to quarks, which is why the study of quarks is an important aspect of particle physics. This is especially true for top quarks, hence this study will be mostly concerned with improving the top tagging algorithms.

2.2 Top Quarks

As one can see in Fig. 1 the top quark is the heaviest of all SM particles. In fact it is about 40-times heavier than the second heaviest quark, the bottom quark. It took up until 1995 to discover this particle, although it has already been predicted in the 1970's [10, 11].

Due to its high mass, it has a very short life time of order 10^{-25} s. Since hadroniza-

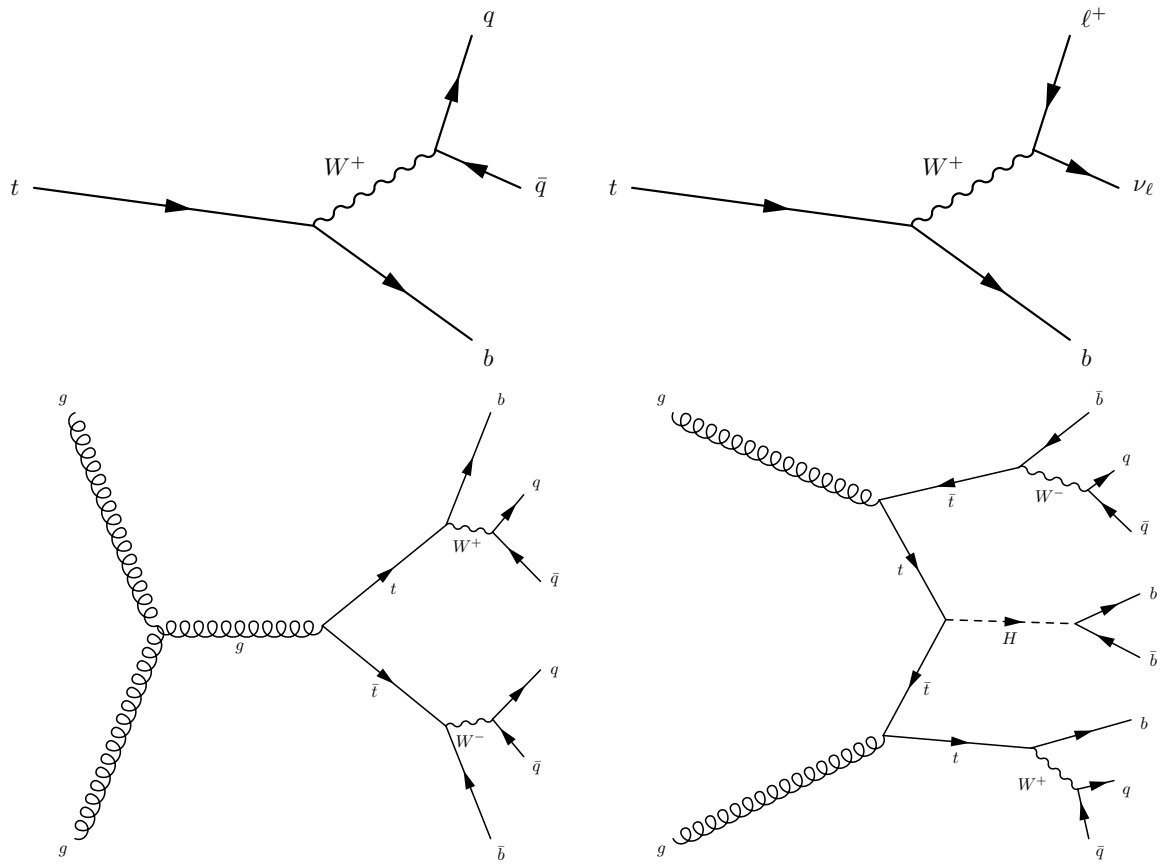

Figure 2: The Feynman graphs for the hadronic and leptonic top decay (upper row), where q describes a quark and ℓ is either an electron or an muon. In the lower row one can see one possible diagram for the fully hadronic top pair production as well as a fully hadronic $t\bar{t}H$-process. Note that q and $\bar{q}$ do not necessarily have the same particle identity.

tion takes place after about 10^{-23} s, the top quark decays before it can hadronize, i.e. there are no hadrons consisting of top quarks. In almost all cases, the top quark decays into a W and a bottom quark. The W decays leptonically or hadronically, i.e. into a lepton and a neutrino or into two quarks respectively. If the W of the top decay decays hadronically/leptonically, the top is called a hadronic/leptonic top quark.

In Fig. 2, one can see the Feynman graphs for the hadronic and leptonic top decay on the upper left and the upper right. In these graphs q describes a quark and ℓ describes a lepton. On the lower left, one can see an example of the top pair production, which will be the events we will consider in our study. The g describes a gluon. There are other processes in which tops can be produced, however the production of a top-antitop-pair is the most frequent one. In our study, we will only consider the hadronic top decay.

The reasoning behind this is, that in this way we do not have the problem of missing transverse momentum as it is the case for the leptonic top quark, since we will not be able to register the neutrino at the detector. Furthermore, the hadronic top decay channel has the dominant branching ratio.

2.3 Fully hadronic $t\bar{t}H$

There are several decay channels for the Higgs boson. It was first experimentally observed in the channel $H \to \gamma\gamma$ [7]. However, this channel has a small branching ratio compared to decay channels like for example $H \to W^+W^-$ or $H \to b\bar{b}$, which have the highest branching ratios [15]. The channel $H \to b\bar{b}$ is hard to measure, because of the combinatorial background. Furthermore, channels like the $H \to \gamma\gamma$ or the $H \to ZZ \to 4\ell$ allow for a high precision in the measurement of their energy and momenta and therefore, for a high quality in the reconstruction of the Higgs mass. Also, it allows, to easily remove the background. This is the reason, why this channel, despite its low branching ratio, was first observed.

However, since the Higgs has been discovered, we are interested in exploring the properties it has. The Higgs boson is the only scalar particle in the SM and therefore, also the only particle, where the Yukawa coupling to other fermions occurs. Therefore the $t\bar{t}H$ process is of huge interest in studying the Yukawa coupling. The Feynman graph for this process with hadronic tops can be seen on the bottom right of Fig. 2. This process is one of the most challenging processes one has to deal with in the study of the Higgs boson, because there is still a lot of combinatorial background occurring. These background processes include for example $t\bar{t}b\bar{b}$ and $t\bar{t}Z$ with $Z \to b\bar{b}$. Reconstructing the top quarks allows to rule out many of the combinatorial backgrounds.

In the last part of our study we will examine this process, with fully hadronic top quarks.

3 Phenomenology and Technicalities

Phenomenology describes physical aspects lying between the field of theoretical and experimental particle physics. This includes the aspects relating the experimental results of particle colliders to the theoretical predictions and vice versa. At colliders, like the LHC, we measure the result of particle collisions with detectors. Yet, these signals are not directly from quarks or gluons, since they cannot exist by themselves, due to color confinement. Except for the top quark they will hadronize into baryons and mesons, which will decay in several stages. The top quark decays directly, since its lifetime is smaller than the time it takes for quarks to hadronize. The hadronic decay products of the top quark, however do also hadronize. The other heavy particles $W^{\pm}, Z$ and H also decay directly. We can only measure particles that live long enough to reach the detector and interact with it, e.g. p, $\bar{p}$, $e^{\pm}$ and γ. This is why we have to find a way to relate the measured outcomes to the theory.

In the following we will consider how signals are received at a detector and which entities are used to describe them. Then we will see how to link these experimentally observed objects to the asymptotic final states of QCD. Due to hadronization and hadron decay, they give rise to energy deposits in the calorimeter, which one clusters into jets. For our later purposes it will be also necessary to look at the special case of so-called fat jets. In the last section we will have a look on how such events are generated and introduce the software we are using for that purpose.

3.1 Collider physics

In a collider, particles (usually electrons or protons and/or their antiparticles) are accelerated until a certain energy is achieved. Then those beams are impacted on each other. The result of this collision is registered by the detectors. Usually the hard processes we want to examine do not relate to hadrons, but rather to their constituents, the partons. In Fig. 3 an event at a hadron collider like the LHC is illustrated. The incoming protons collide, their partons are released and emit radiation, the so-called initial state radiation (ISR). Partons interact with each other and this way a lot of other particles are created. This process is divided into the hard process, the process one is interested in, and the underlying event. The remnants of the hard process and the underlying event are the outgoing partons, which also emit radiation, the so-called final state radiation (FSR). These outgoing partons then hadronize and/or decay in several stages. The decay products are

then registered by the detectors, if they interact with them.

3.1.1 The η-ϕ-plane

To analyze the results we obtain, the separation between these occurring structures is an important quantity. However, we are not able to access all the kinematic variables of the process. Considering the beam axis to be in z-direction, it is impossible to access the momentum in this direction p_z. It is therefore important to find a way to describe the kinematics of the process, using only the quantities we are able to measure.

Usually, the shape of the detector is cylindric to be able to register most of an event. We imagine to cut this cylinder open and consider the structures on this plane. The azimuthal angle ϕ of the plane is invariant to longitudinal Lorentz boosts, since it is orthogonal to the beam axis. The polar scattering angle θ (the angle between the beam axis and the three-momenta), however does not have this property.

The rapidity y of a particle with four momentum p^μ and energy E is defined as

$$y \equiv \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right) . \tag{1}$$

It turns out that it is additive under longitudinal Lorentz boosts. Therefore differences in the rapidity are Lorentz invariant. As one can see, the rapidity depends on p_z, which as already mentioned, cannot be accessed in the experiment. However, if we take $m \to 0$, we

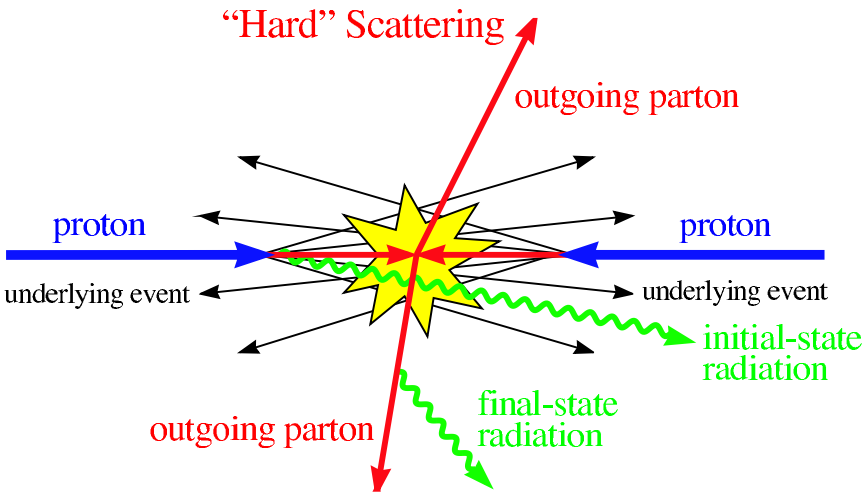

Figure 3: Illustration of an event at a hadron collider, like the LHC. Taken from [18].

get

$$\eta = \lim_{m \to 0} y = \frac{1}{2} \ln \left(\frac{|\vec{p}| + p_z}{|\vec{p}| - p_z} \right) = \operatorname{artanh} \left(\frac{p_z}{|\vec{p}|} \right) = - \ln \left[\tan \left(\frac{\theta}{2} \right) \right] \ . \tag{2}$$

This is the so-called pseudorapidity, which is apparently only dependant on θ. Note that, like the rapidity, differences in the pseudorapidity are Lorentz invariant under longitudinal boosts. Apparently for $m \approx 0$ it is $\eta \approx y$.

In the definition of the pseudorapidity, we can see that for small values of θ (i.e. particles close to the beam axis) η gets large. The same is true for values of $\theta \sim \pi$, where η gets negative. In these cases the particles are so close to the beam axis, that they do not reach the detector and therefore cannot be registered.

Therefore, we will impose $|\eta| < 2.5$ on the particles we examine, to make sure, that we can actually measure these particles at a detector.

Furthermore, in an actual detector exists a certain degree of granulation, i.e. there is a limit up to which the detector can distinguish if two signals are separated. This is due to the fact that the detector is split up into cells. This limit roughly lies at $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$, i.e. for signals which are closer together than this, the detector considers them as one signal. Since this becomes relevant already for jets of low transverse momentum, we need to consider this our analysis.

Using the definition of the pseudorapidity, we can now introduce the η-ϕ-plane and get the following measure of separation between two particles, which is invariant under longitudinal Lorentz boosts

$$\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} \ . \tag{3}$$

Especially for the specification of the size of a jet this definition of separation will become very useful later on.

3.1.2 The Mandelstam variables

The Mandelstam variables s, t and u are Lorentz invariant kinematic variables to describe a $2 \to 2$ scattering processes. With p_1, p_2 being the four momenta of the incoming particles

and p_3, p_4 the ones of the outgoing particles these are defined as:

$$s = (p_1 + p_2)^2 = (p_3 + p_4)^2 \tag{4}$$

$$t = (p_1 - p_3)^2 = (p_2 - p_4)^2 \tag{5}$$

$$u = (p_1 - p_4)^2 = (p_2 - p_3)^2 \tag{6}$$

Apparently s is the square of the center-of-mass (c.o.m.) energy, which is therefore given by $\sqrt{s}$. Usually, one uses this quantity to specify the energies at which the particles are colliding. If the particles are accelerated head-on with an beam energy of 6.5 TeV, as it is approximately the case at the LHC in run II, one can just add the c.o.m. energies of both beams, i.e. $\sqrt{s} = 13$ TeV.

3.1.3 The transverse momentum

Another important quantity to describe the kinematics is the transverse momentum

$$p_T = \sqrt{p_x^2 + p_y^2} = p \sin\theta \ . \tag{7}$$

It serves as a measure of how highly the outgoing particles are boosted. It is apparently invariant under Lorentz boosts in longitudinal direction. As we will see in later sections, for tagging heavy particles it is important to have the decay products close enough to each other, to cluster them into jets. A minimum cut on the transverse momentum yields narrow jet cones. Therefore, in such cases one usually imposes a cut on the transverse momentum. Also $p_T > 0$ ensures that the particles will be detectable, since particles with no transverse momentum clearly won't reach the detector, but go down the beam line.

3.1.4 QCD effects at hadron colliders

There are several effects, occurring in actual colliders, which complicate the analysis of the events. In the following, we will briefly consider them. Including all these effects in our study would be beyond our scope. Experimentalists have, however, found ways to deal with most of these problems.

Underlying Event

The colliding parts of the protons, which are not contributing to the hard process will usually also be registered by the detector. These beam remnants form colorless hadrons, which will be distributed equally in the detector. This effect is called

underlying event (UE). As it is pointed out in Ref. [20] and [21] UE will have an impact on the jet mass, which will increase proportional to R. Therefore, to reduce the effects of UE, it is desirable to keep the jet size R as small as possible.

Multiple particle interactions

In the hard process, we are considering individual partons interacting with each other. Since the colliding hadrons consist of more than one parton, there can be also interactions coming from other partons. This effect is what one understands under multiple particle interactions (MPI).

Pile-up

In a hadron collider the hadrons typically do not come one by one, but in bunches of many hadrons. Therefore it happens, that there is more than one event per bunch crossing and it is difficult to be able to clearly distinguish between the individual events. The events are piling up, which is why this effect is called pile-up.

3.2 Jets and Fat jets

As we have seen, the outgoing quarks and gluons of the hard processes are not observed at the detector. Instead, the partons emit FSR, hadronize and decay in several stages. Our aim is to reproduce the partons from this spray of hadronic activity. The objects we obtain in this way are called jets and the methods used to obtain them are the jet algorithms.

3.2.1 Jet algorithms

The purpose of jet algorithms is to be able to link the hadronic activity in the calorimeter to the partons. In other words we want to know which hard process lies behind the bunch of signals our detector received. The first question is if this is even possible.

It turns out, that it is possible, if we combine the huge amount energy deposits we get in the experiment to a small number of jets using jet algorithms. To terminate the jet algorithm, one has to either impose a cut on the number or the resolution of the subjets.

There are two different kinds of jet algorithms: the cone algorithms and the recombination algorithms. We will restrict ourselves to the discussion of recombination algorithms. Recombination algorithms which are usually used are the purely geometric Cambridge/Aachen (C/A) algorithm [22] and the (anti-) k_T algorithms [23, 24].

To decide if two signals are arising from the same parton these algorithms use a suitable measure. It turns out, that using the following measure d_{ij} with the parameter n, the

angular separation R as defined in Eq. (3) and the transverse momentum p_T defined in Eq. (7) is a convenient choice for recombination algorithms:

$$d_{ij} = \frac{\Delta R_{ij}^2}{R^2} \min\left(p_{T,i}^{2n}, p_{T,j}^{2n}\right), \; d_{iB} = p_{T,i}^{2n} \,, \tag{8}$$

where d_{iB} is the beam distance.

The three algorithms do now only differ in the choice of the parameter n. For k_T one chooses $n = 1$, for C/A $n = 0$ and for anti-k_T it is $n = -1$. Before we discuss them in more detail we first consider how the algorithms work:

1. Find the minimal distance for all pairs of ij: $d_{\min} = \min_{ij}(d_{ij}, d_{iB})$

2. (a) If $d_{\min} = d_{ij} < d_{\text{cut}}$, join i and j together. Return to step 1.

   (b) If $d_{\min} = d_{iB} < d_{\text{cut}}$, consider i as beam radiation and disregard it. Return to step 1.

   (c) If $d_{\min} > d_{\text{cut}}$, consider all remaining as jets.

Here we introduced a cutoff distance d_{cut}, which determines the resolution of the algorithm. Such an algorithm is called an exclusive jet algorithm. In an inclusive jet algorithm we use the beam distance as cutoff distance, i.e. in step 2 (b) instead of dropping i we consider it as a jet and go back to step one (step 2 (c) does of course not exist in such an algorithm).

All jet algorithms try to relate the signals in the calorimeter to the partons of the hard process. Yet they do this in different ways. While the C/A algorithm is purely geometric, the k_T algorithm begins with the soft constituents and the anti-k_T algorithm takes the hard constituents first to form a jet. From this it is apparent, that the anti-k_T algorithm seems to have in contrast to the other algorithms no physical interpretation at all. However the jets of this algorithm have a circularly shape (see Fig. 4). This simplifies the computation of the area of the jets and since the impurities from underlying event are proportional to this area these effects can be easily removed. In Fig. 4 one can see the shapes of the clustered jets for these algorithms. Throughout our study we will use the purely geometric C/A algorithm.

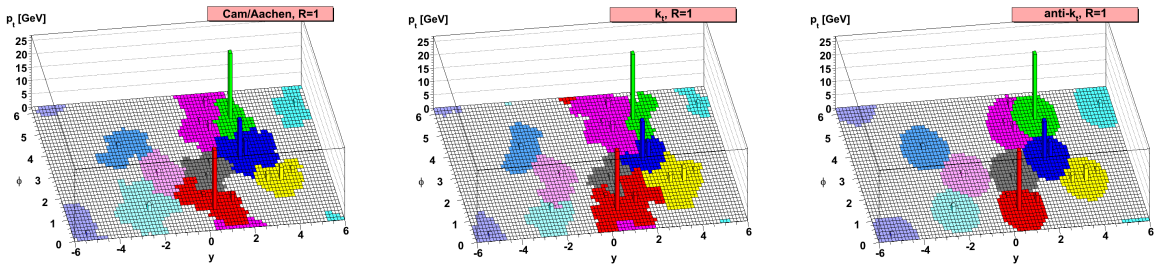

Figure 4: Shapes of the clustered jets for the different jet algorithms. Taken from [24].

3.2.2 Fat jets

Usual choices for the jet size are $R \in [0.4, 0.7]$. The ATLAS Collaboration for example takes $R = 0.4$ and $R = 0.6$ for their analyses [25]. As we have seen in our discussion of UE it is useful to keep the jet size as small as possible to reduce these effects. However, if we are interested in jets originating from the decay of heavy particles, as it is the case for the top quark, we have to include all decay products in our jet. Such jets are called fat jets.

In our discussion of the HEPTopTagger in the next section, it will turn out, that to include top decay products for low p_T in one jet the jet size has to become huge (see Fig. 5). However, in this case we won't be able to identify the decay products in a QCD busy environment. However, if we take a small jet size $R < 1$ we won't be able to include all decay products in our jet for the most top quarks. As as compromise one usually chooses $R = 1.8$, building jets that include the decay products of the top quarks with $p_T > 200$ GeV.

However, as one can see in Fig. 5 most of the top quarks have $p_T < 200$ GeV and can therefore not be considered using fat jets. Therefore it is important to find other ways of clustering to include also the top quarks in this regime. Just using higher jet sizes is not helpful, since then impurities of UE will become huge. In the last section of our analysis, we will consider one possibility to avoid this.

Although, we discussed this problem on the example of the top quark, this is also true for other heavy particles like the Higgs, the $W^{\pm}$ or the Z.

To be able to analyze the massive splittings inside fat jets is crucial for tagging algorithms. The C/A algorithm turns out to be a good choice for achieving this, which is why we will use it throughout our study.

3.3 Event Generation

To simulate events comparable to those at the LHC, Monte Carlo event generators are used. It is important to be able to consider the occurring effects in the event generation, to compare the obtained results to the experimental data. Therefore, one has to compute the cross section of the hard process one is interested in, the parton showers for the in and outcoming partons, the interaction of the partons, the hadronization and the decays of unstable particles.

In the following, we will briefly review these effects and how they can be computed by the event generators.

Hard process

One is interested in hard processes with large momentum transfer (e.g. the production of heavy particles). The cross section is an area, which serves as a mesure for the probability of a certain scattering event to occur. To compute it, it is helpful to exploit the QCD factorization theorem, which allows to factorize the scattering into a parton level hard scattering encapsulated with the parton distribution functions (PDFs). The cross section of a scattering of hadrons A and B into FX, with F being the final state we are interested in and X additional scattering remnant including FSR, is therefore

$$\sigma(AB \to FX) = \int_0^1 \mathrm{d}x_1 \int_0^1 \mathrm{d}x_2 \sum_{a,b} f_{a,A}(x_1, Q^2) f_{b,B}(x_2, Q^2) \hat{\sigma}(ab \to FX) \,, \quad (9)$$

where $f_{a,A}$ and $f_{b,B}$ are the PDFs, i.e. the probability density describing the probability of finding a parton inside the hadron with momentum $k = x_i p$, with p being the total momentum of the hadron and $x_i \in [0, 1]$. Q^2 is the factorization scale and $\hat{\sigma}(ab \to FX)$ is the cross section of the hard process. The computation of this integral is a complicated task, which can be solved numerically using Monte Carlo methods.

Parton shower

In the hard process we are mostly dealing with partons, which can emit gluon radiation. These gluons carry color charge and therefore can also emit soft and/or collinear radiation and/or decay into quark-antiquark pairs. This way a huge amount of particles will be generated. This effect is described by the parton shower.

The parton shower can be simulated by a Markov chain, which probabilistically chooses if at a certain time a parton is added or not. This is implemented in the parton shower algorithm.

Hadronization

As the event evolves downwards in momentum, it will at some point reach scales of order Λ_{QCD}, where QCD becomes strongly interacting and perturbation theory won't be valid anymore. Therefore, one has to change to the hadronization model, which describes the confinement of colored partons into colorless hadrons. One key property of the hadronization model is, that the partons won't hadronize by themselves, but together with the other partons. One should note, that since the transition to color confinement is still not completely understood, this model is not directly derived from QCD. The hadronization of a certain colored system is, however, independent of how that system was produced, therefore this poses no problem.

Hadron decays

Usually the produced hadrons are unstable and decay into other particles. One might think, that this is easy to simulate, since one just has to implement the experimental data of the hadron decays. However, the information one obtains in this way is often insufficient. Therefore, additional to the experimental data theoretical assumptions have to be made. In event generators like PYTHIA8, which we will discuss in the next section, sophisticated models for doing this have been developed and implemented.

3.4 Software

In the previous section, we saw how events are generated and analyzed. In the following, we will introduce the Software, that we use for generating as well as analyzing such events.

PYTHIA8

PYTHIA8 [27] is a Monte Carlo event generator suited for high energy physics events. It contains the Standard Model as well as several other models of new physics. It includes hard and soft interactions, parton distributions, initial and final state parton showers, multiparton interactions, decays and several other aspects. Since version 8, it is written in C++. In the first parts of our study we will use it for generating events, while for the last part, we will only use it for the parton shower.

MADGRAPH/MADEVENT

MADGRAPH/MADEVENT [28] also serves to generate events using Monte Carlo methods. Here MADGRAPH creates the matrix element of a hard process and MADEVENT generates the events at the parton level and computes the cross sections. In the last part, we will use it to generate the hard events.

FASTJET

FASTJET [29] delivers tools for analyzing and constructing (clustering) jets. It includes the previous described jet algorithms. We will especially use it for clustering the parton level events into jets.

Furthermore, we will use the HEPTopTagger for tagging the top quarks. Also for the last part, we will use the BDRS Higgs Tagger, which works in a similar way. In the next section we will describe tagging algorithms in more detail.

4 Tagging Top Quarks

For tagging top quarks we use the HEPTopTagger [30–34] (HEP=Heidelberg-Eugene-Paris). It identifies top quarks inside fat jets and was originally used to study the $t\bar{t}H$ process. Depending on the p_T of the top quark, the decay products will usually be far apart. Therefore, when clustering fat jets we have to choose the jet size high enough to include all decay products in the jet. We get this size if we compute the three R distances of this decay on the parton level: For the first distance we find the combination with the smallest ΔR_{ij}. The second length we get by combining i and j and calculating the R distance to the third constituent. The maximum of these two lengths is the approximate partonic initial size ΔR_{bjj} of a C/A jet including all three main decay products. As one can see in Fig. 5 a high initial jet size has to be taken to be able to resolve top decays at a sufficient p_T-range. However, as mentioned in the previous section, with higher jet sizes the effect of underlying event will also increase. Therefore, as a compromise, we're going to use $R = 1.8$ throughout this study for tagging top quarks. This allows us to tag top quarks with $p_T > 200$ GeV.

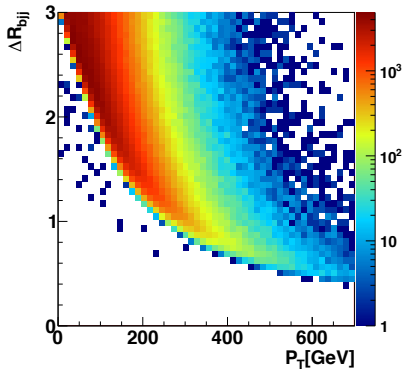

Figure 5: Plot of ΔR_{bjj} against p_T for a SM $t\bar{t}$-sample on parton level.Taken from [**?**].

4.1 The HEPTopTagger

4.1.1 The algorithm

In the following the HEPTopTagger algorithm will be introduced. Before one starts the algorithm one has to define a fat jet with $R = 1.8$ using the C/A algorithm as the starting point for the HEPTopTagger. The algorithm then proceeds in the following steps:

1. **Finding the hard substructures:** To find the hard substructures coming from the bottom of the $t \to W^+ + b$ decay or one of the jets from the $W^+ \to j + j$ decay by iterative unclustering the jet j into two subjets j_1 and j_2 with $m_{j_1} > m_{j_2}$. To analyze if the splitting is related to two top decay products rather than ISR or UE check the mass drop criterion

$$m_{j_1} < \theta m_j \ , \tag{10}$$

where θ is the mass drop threshold. Since $m_W/m_t \approx 0.47$ and $m_j/m_W \to 0$ a threshold of $\theta = 0.8$ turns out to be sufficient to resolve the massive splittings. If this criterion is fulfilled both jets are kept, otherwise the lighter one is regarded as soft radiation or coming from underlying event and therefore dropped. All kept jets are then decomposed further. There is a cutoff mass m_{cut}, i.e. for $m_{j_i} < m_{\text{cut}}$ with $i = 1, 2$ the unclustering is stopped and the jets are kept. Usually one chooses $m_{\text{cut}} = 30$ GeV. This serves to ensure that one does not decompose the fat jet too far, possibly dropping FSR from the decay products, which is not what we want.

2. **Filtering:** Iterate over every combination of three hard substructures and apply a filtering using the criterion $R_{\text{filt}} = \min(0.3, \Delta R_{jk}/2)$. In this way, the effects of underlying event are reduced, since the effective jet size is downscaled. Take the $N_{\text{filt}} = 5$ hardest filtered constituents and compute their jet mass. Now recluster them into three top decay jets and keep the ones which fulfill $m_{123} \equiv m_{\text{rec}} \in [150, 200]$ GeV. If no such triple exists, reject the fat jet.

3. **Check the mass criteria:** Sort the decay jets j_1, j_2, j_3 by p_T.

   If we assume $p_i^2 = 0$ with $i \in \{1, 2, 3\}$, we get

$$m_t^2 \equiv m_{123}^2 = (p_1 + p_2 + p_3)^2 = (p_1 + p_2)^2 + (p_1 + p_3)^2 + (p_2 + p_3)^2 = m_{12}^2 + m_{13}^2 + m_{23}^2 \ . \tag{11}$$

   This describes the surface of a sphere with radius m_t in (m_{12}, m_{13}, m_{23}). Therefore if we fix m_{123}, we can get two more variables, which describe the kinematics: By choosing m_{23}/m_{123} and $\arctan m_{13}/m_{12}$ we get for m_{12}/m_{123}

$$1 = \left(\frac{m_{12}}{m_{123}}\right)^2 \left(1 + \left(\frac{m_{13}}{m_{12}}\right)^2\right) + \left(\frac{m_{23}}{m_{123}}\right)^2 \ . \tag{12}$$

   For $m_{123} = m_t$ the constraint $m_{12} = m_W \pm 15\%$ becomes $m_{12}/m_{123} \in [R_{\min}, R_{\max}]$.

Using this we can impose further constraints on the subjets:

$$
\begin{aligned}
&0.2 < \arctan\frac{m_{13}}{m_{12}} < 1.3 \qquad \text{and} \quad R_{\text{min}} < \frac{m_{23}}{m_{123}} < R_{\text{max}} \\
&R_{\text{min}}^2\left(1+\left(\frac{m_{13}}{m_{12}}\right)^2\right) < 1-\left(\frac{m_{23}}{m_{123}}\right)^2 < R_{\text{max}}^2\left(1+\left(\frac{m_{13}}{m_{12}}\right)^2\right) \quad \text{and} \quad \frac{m_{23}}{m_{123}} > 0.35 \\
&R_{\text{min}}^2\left(1+\left(\frac{m_{12}}{m_{13}}\right)^2\right) < 1-\left(\frac{m_{23}}{m_{123}}\right)^2 < R_{\text{max}}^2\left(1+\left(\frac{m_{12}}{m_{13}}\right)^2\right) \quad \text{and} \quad \frac{m_{23}}{m_{123}} > 0.35\,,
\end{aligned}
\tag{13}
$$

where $R_{\text{min,max}} = (1 \mp f_W)m_W/m_t$ introduces the parameter f_W, with $f_W = 0.15$. These constraints restrict the masses to lie on certain bands in the $\arctan(m_{13}/m_{12})$-m_{23}/m_{123}-plane, which together have the shape of an A. This is illustrated in Fig. 6. The constraints for $\arctan m_{13}/m_{12}$ and m_{23}/m_{123} in Eq. (13) serve as an additional cut to reduce backgrounds.

If at least one of the constraints given in Eq. (13) is fulfilled, the masses lie on the A-shaped area and we accept them as top candidates.

4. **Choose candidate:** If there is more than one candidate per fat jet fulfilling these criteria, pick the one whose $m_{123} \equiv m_{\text{rec}}$ is closest to m_t.

5. **Consistency cut:** Demand for the reconstructed top to fulfill $p_{T,t} > 200$ GeV for consistency.

There are several extensions to this algorithm like the low-p_T mode or the optimalR mode. Since they are not relevant for our purposes we'll neglect these here. They can be looked up in the appendix of [31], from where we also took the steps above.

4.1.2 Performance

We test this algorithm with hadronic $t\bar{t}$ and QCD dijet Monte Carlo samples for the LHC at $\sqrt{s} = 13$ TeV generated with PYTHIA8 without multiple interactions. As starting point, we take C/A jets with $R_{\text{fat}} = 1.8$ constructed with FASTJET obeying $|\eta_{\text{fat}}| < 2.5$ and $p_{T,\text{fat}} > 200(600)$ GeV. For the signal we require that the fat jets can be matched to parton level tops within $\Delta R = 0.8$. The signal efficiencies are then given by the number of tags divided by the number of fat jets fulfilling these criteria. The results can be seen in Tab. 1, where $\epsilon_{S,B}$ describe the signal and background efficiencies respectively.

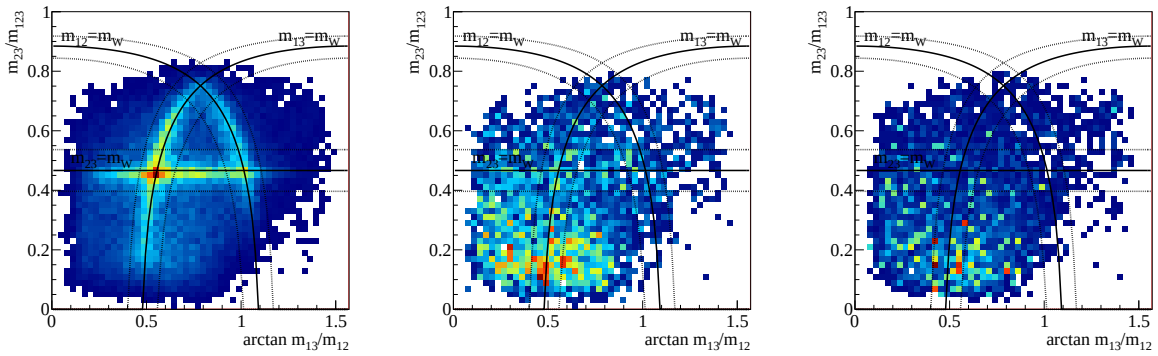

Figure 6: Distribution of the events in the m_{13}/m_{12}-m_{23}/m_{123} plane. From the left to the right: $t\bar{t}$, W+jets and pure QCD-jets. The plots are taken from [31]
.

A study of the quality of the reconstructed top quarks compared to the parton level tops can be found in section 5, where we will compare them to alternative algorithms for the substructure reconstruction.

$p_{T,\text{min}}$ [GeV]	ϵ_S	ϵ_B
200	0.3065	0.0189
600	0.4432	0.0481

Table 1: Efficiencies for the HEPTopTagger using the mass drop criterion with $\theta = 0.8$.

5 Alternative Substructure Reconstruction

In the previous section we introduced the HEPTopTagger. It uses the mass drop algorithm to uncluster the fat jets.

However, there are several other approaches to reconstruct these substructures. To see if these alternative approaches should be preferred for top tagging is the aim of this section.

We will study the so-called Soft Drop and Mass Jump algorithms and compare the efficiency of the top tagging to the one we get using the mass drop algorithm. Furthermore, we will study the quality of reconstruction and see if there are any improvements using these alternative algorithms.

In the following sections, we will therefore introduce these algorithms. Then we will give the results of our study and compare them to the mass drop algorithm. In the end we will conclude whether it is recommendable to adapt one of these algorithms to the HEPTopTagger.

5.1 Soft Drop

5.1.1 The Soft Drop algorithm

In this section, we study the Soft Drop algorithm (SD) [36]. Instead of masses SD uses a combination of the angular distance ΔR_{ij} and the asymmetry of the splitting in terms of the transverse momenta. The SD algorithm checks in every unclustering step if the following relation (soft drop criterion) is fulfilled:

$$\frac{\min(p_{T_1}, p_{T_2})}{p_{T_1} + p_{T_2}} > z_{\text{cut}} \left(\frac{\Delta R_{12}}{R} \right)^{\beta} \tag{14}$$

where we define R with the fat jet size R_{fat}. z_{cut} and β are two adjustable parameters.

If Eq. (14) is fulfilled, both jets are kept and decomposed further. Otherwise the softer jet is dropped, as it is the case for the mass drop criterion.

The dependance on the parameters z_{cut} and β is illustrated in Fig. 7. For $\beta < 0$ one asks for symmetric and angular separated splittings and thereby removes soft and collinear radiation. $\beta = 0$ achieves a pure cut on the asymmetry and therefore removes only soft radiation. $\beta > 0$ keeps collinear emission and removes wide angle radiation, which is not serving our purposes. However, if one chooses z_{cut} sufficiently small it effectively removes the dependance on the angular separation.

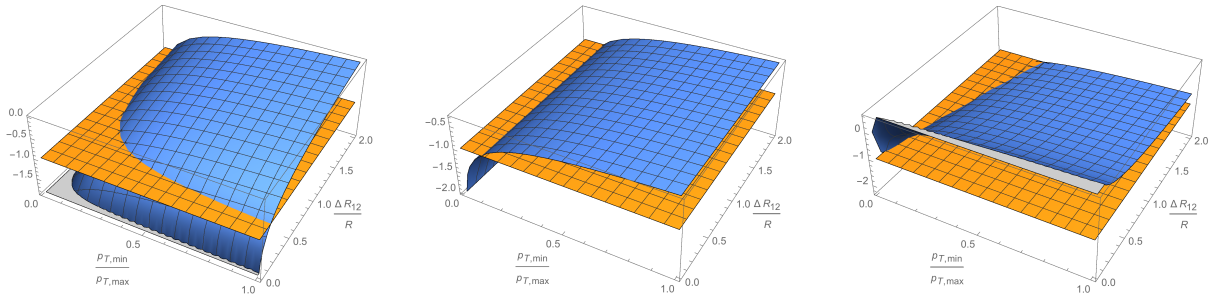

Figure 7: Soft Drop condition for $\beta = -1$, $\beta = 0$, and $\beta = 1$ (left to right). The yellow plane illustrates a cut $z_{\text{cut}} = 0.1$. The soft drop criterion is fulfilled if the blue surface lies above the yellow one.

To implement this algorithm into the HEPTopTagger we replace Eq. (10) by Eq. (14). The rest of the algorithm remains unchanged.

5.1.2 Performance

As before, we test this algorithm with hadronic $t\bar{t}$ and QCD dijet Monte Carlo samples for the LHC at $\sqrt{s} = 13$ TeV generated with PYTHIA8 without multiple interactions. As starting point, we take C/A jets with $R_{\text{fat}} = 1.8$ constructed with FASTJET obeying $|\eta_{\text{fat}}| < 2.5$ and $p_{T,\text{fat}} > 200(600)$ GeV. For the signal we require that the fat jets can be matched to parton level tops within $\Delta R = 0.8$. The signal efficiencies are then given by the number of tags divided by the number of fat jets fulfilling these criteria.

In Fig. 8 one can see the parton level distributions for $\Delta R/R$ and $\min(p_{T_1}, p_{T_2})/(p_{T_1} + p_{T_2})$. Here and in future plots N_{rel} describes the relative abundance, i.e. the abundance after normalizing the histograms.

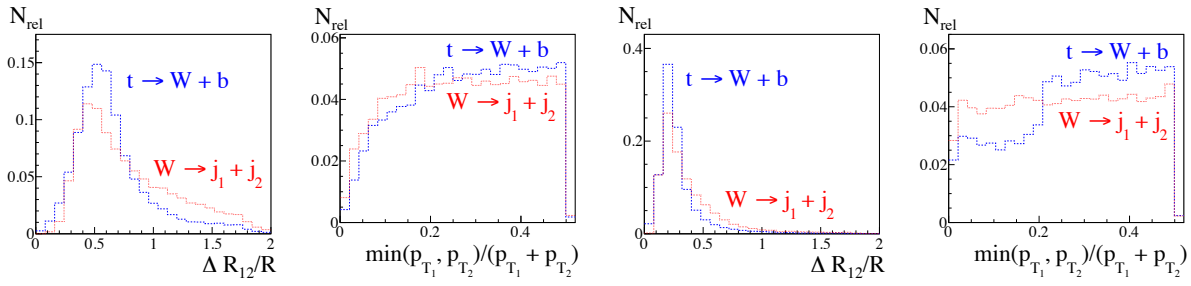

Figure 8: Parton level distributions for $\Delta R_{12}/R$ and $\min(p_{T_1}, p_{T_2})/(p_{T_1} + p_{T_2})$ for $p_T > 200$ GeV (left) and $p_T > 600$ GeV (right). Events with $\Delta R_{12}/R > 1$ cannot be captured in a fat jet of size $R = 1.8$.

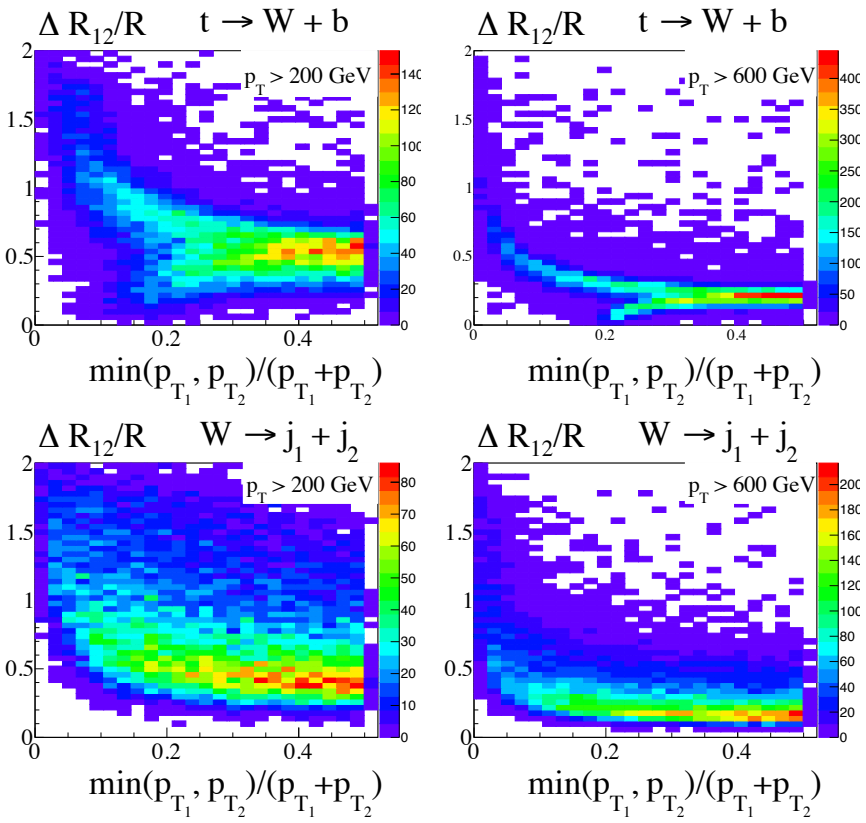

Figure 9: Correlation of $\Delta R_{12}/R$ and $\min(p_{T_1}, p_{T_2})/(p_{T_1} + p_{T_2})$ for $p_T > 200$ GeV (right) and $p_T > 600$ GeV (left).

In both splittings the angular separation peaks at $\Delta R/R \approx 0.5$. Note, that for $\Delta R/R > 1$ it is not possible to capture the event in a fat jet of size $R = 1.8$. As one can see, this is the case only for a minority of events. Also one can see, that for $p_T > 600$ GeV the jets lie closer together, because they are higher boosted. The asymmetry distributions are almost flat above 0.1. This already indicates, that it will be unlikely to find a value for z_{cut} that allows a clear separation of signal and background.

However, for $p_T > 600$ GeV the asymmetry distribution of the top decay has a step at about 0.2, i.e. lower values are suppressed. To make sense of this we calculated the kinematics of the $t \to W + b$ process, where this occurs. It turned out, that if we assume $p_z = 0$ and calculate the transverse momenta of the decay products, imposing $p_T = 600$ GeV, we get $\min(p_{T_1}, p_{T_2})/(p_{T_1} + p_{T_2}) > 0.2$. Thus this step occurs for purely kinematic reasons. The correlation of the angular separation and the asymmetry distribution can be seen in Fig. 9.

To get an estimate of an appropriate value for $z_{\rm cut}$ at fixed β in Fig. 10, the quantity

$$z = \frac{(\min(p_{T,1}, p_{T,2})/(p_{T,1} + p_{T,2})}{(\Delta R_{12}/R)^{\beta}} \tag{15}$$

is plotted for three different values of β. It turns out, that changes in the choice of β effectively result in a sqeezing of the z distribution. Therefore, we expect the performance to be almost independent of β, especially for sufficiently small values of $z_{\rm cut}$, as it has already been noticed earlier. Furthermore, we expect to have high efficiencies for values of $z_{\rm cut} < 0.1$ and decreasing efficiencies at higher values since for all values β the distribution peaks at $z_{\rm cut} \sim 0.1(0.05)$ for $p_T > 200(600)$ GeV.

In the following we run the HEPTopTagger using the mass drop and the soft drop criterion and compare the results. For the soft drop criterion, we vary the parameter $z_{\rm cut}$ in the range $[0, 0.3]$. We do this for the values $\beta = \pm 0.5, \pm 1.0, \pm 1.5$. The granulation of

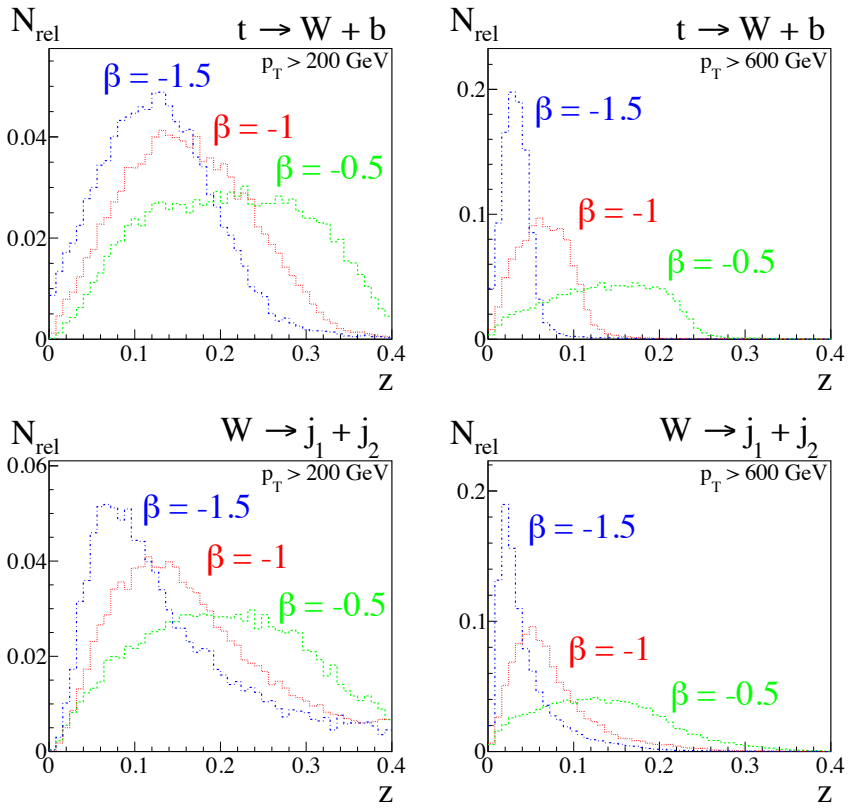

Figure 10: $z = (\min(p_{T,1}, p_{T,2})/(p_{T,1} + p_{T,2}))/(R_{12}/R)^{\beta}$ for $p_T > 200$ GeV (left) and $p_T > 600$ GeV (right).

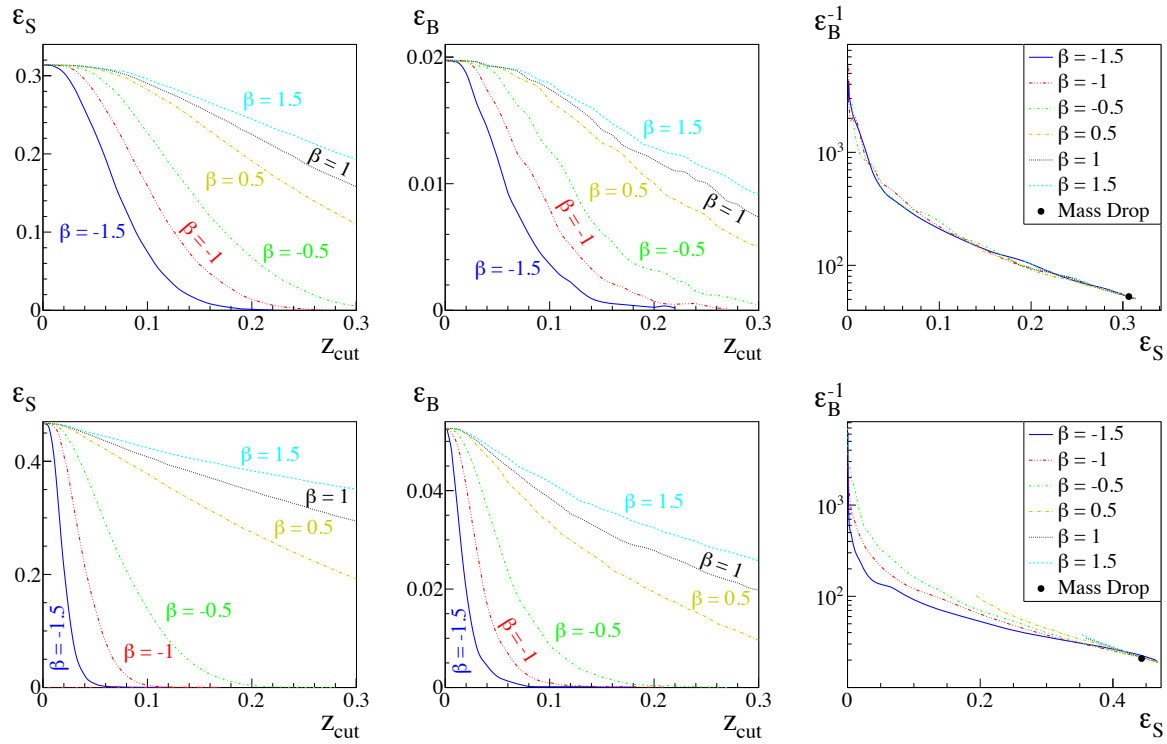

Figure 11: The signal and the background efficiencies in dependance of z_{cut} (left) and the ROC curve (right), for $p_T > 200$ GeV (upper row) and $p_T > 600$ GeV (lower row).

the detector $\Delta\phi \times \Delta\eta = 0.1 \times 0.1$ (see section 3.1.1) is taken into account.

To graphically represent our results, we use Receiver Operating Characteristic (ROC) curves. These curves relate the signal efficiency and the mistagging rate. On the x-axis, we have the signal efficiency and on the y-axis the inverse of the mistagging rate. In Fig. 11, the ROC curve and the curves visualizing the dependance of the signal efficiency and the mistagging rate on z_{cut} are shown. One can see, that for all parameter choices, we typically receive tagging efficiencies, which are lower or about the same size as the ones using the mass drop criterion. For similar signal efficiency, we get a comparable background rejection at $p_T > 200$ GeV as well as at $p_T > 600$ GeV, if we are using the mass drop criterion. As predicted, there is no significant dependence on the choice of β for $p_T > 200$ GeV. For $p_T > 600$ GeV there are, however, some changes in the background rejection. Also one can see, that the signal efficiencies are optimal for values of $z_{\text{cut}} < 0.1$, as we expected.

To analyze the quality of the reconstruction and to compare it with the one using the default HEPTopTagger (i.e. the mass drop criterion), we choose $z_{\text{cut}} = 0.01$, to have comparable signal efficiencies. The particular signal efficiencies and mistagging rates are

listed in Tab. 2.

β	ϵ_S	ϵ_B
mass drop	0.3065 (0.4432)	0.0189 (0.0481)
-0.5	0.3137 (0.4653)	0.0197 (0.0519)
-1.0	0.3135 (0.4614)	0.0197 (0.0502)
-1.5	0.3124 (0.4159)	0.0196 (0.0400)

Table 2: Efficiencies for $z_{\text{cut}} = 0.01$ at $p_T > 200(600)$ GeV.

In Fig. 12, one can see the mass distribution of the reconstructed top quarks and of the mistagged jets. We also show the deviation in angular distance ΔR, transverse momentum $\Delta p_T/p_T = (p_T^{\text{true}} - p_T^{\text{rec}})/p_T^{\text{rec}}$ and the three-momentum $|\Delta\vec{p}|/|\vec{p}| = (|\vec{p}_{\text{true}}| - |\vec{p}_{\text{rec}}|)/|\vec{p}_{\text{rec}}|$ of the reconstructed top quarks to the ones at the parton level. Apparently, there is no significant peak in the distributions of the masses for the mistagged jets, i.e. there is no preferred mass value for the background. We note, that there is no significant deviation in the quality of the reconstruction between mass drop and soft drop, regardless of the chosen value for β. Therefore we conclude, that soft drop and mass drop work equally well for this choice of parameters.

In Ref. [36] it was claimed, that the SD algorithm may also work without an additional filtering step. For their study of the W boson with $p_T > 500$ GeV this might be true, yet for the top quarks we do not expect such a behaviour. A performance study including multiple interactions and pile-up goes beyond the scope of this study and has to be decided on experimental data. However, we did a rough estimate and compared both criteria

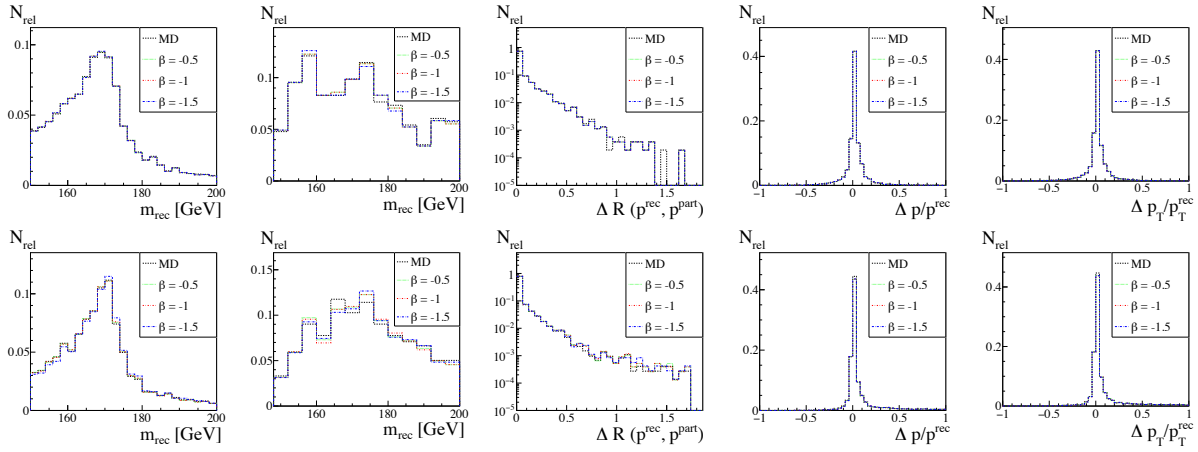

Figure 12: Quality of the reconstructed top quarks with $p_T > 200$ GeV (upper row) and $p_T > 600$. (lower row) for the parameter choice $z_{\text{cut}} = 0.01$.

without using the additional filtering step. It turned out, that no qualitative difference to the results presented above could be found.

5.1.3 Conclusions

In this section it was shown, that the replacement of the mass drop criterion used by the default HEPTopTagger with the soft drop criterion does not yield a significant improvement. Neither a higher efficiency was received nor better qualities of reconstruction could be achieved. Therefore, we can conclude this section by noting that there are no benefits in implementing this algorithm to the HEPTopTagger.

5.2 Mass Jump

5.2.1 The Mass Jump algorithm

Another alternative to the mass drop unclustering is the Mass Jump algorithm (MJ) [37]. The special property of this algorithm is, that it does the jet clustering and unclustering in one step. Before we start the algorithm, we label all objects as active jets. Then the algorithm proceeds as follows [37]:

1. Look for the smallest distance d_{ij}, defined as in Eq. (8) between the active jets. If $d_{ij} = d_{iB}$ call i *passive* and repeat step 1.

2. Combine jets i and j. If this jet is below the cutoff mass $m_{ij} < m_{\text{cut}}$ take this combination and go back to step 1. Otherwise, check if the mass jump criterion $\theta \cdot m_{i+j} > \max[m_i, m_j]$ is fulfilled. If so, label them *passive* and go back to step 1.

3. Since a mass jump can also occur between active and passive jets, it is important to include the following substeps:

   a) Look for the passive jet n closest to i. Check that it is not isolated, i.e. $d_{in} < d_{nB}$.

   b) Regard both jets as active and check if they would be recombined.

   c) Check if their recombined mass is higher than the cutoff mass and if the mass drop criterion is fulfilled.

   If these criteria are fulfilled, call i *passive* and do the same for j. If i or j are labeled passive, go back to step 1.

4. If no mass jump has been found combine i and j and go back to step 1.

The Mass Drop algorithm starts with the fat jet and unclusters it into its substructures. The Mass Jump algorithm instead clusters the hadronic activity into the subjets and thereby does the clustering and unclustering in one step.

In Fig. 13, the differences between mass drop and mass jump are graphically shown. The mass drop algorithm starts with a fat jet and removes soft radiation (1) until a mass drop occurs (2). Then it decomposes both jets further (3, 5) until another mass drop occurs (4), or until the cutoff mass is reached (6). In contrast, the mass jump algorithm does standard clustering until the jet mass exceeds the cutoff mass (a). Then MJ continues the clustering (b) until a mass drop occurs (c) at which the jets are labeled passive. The

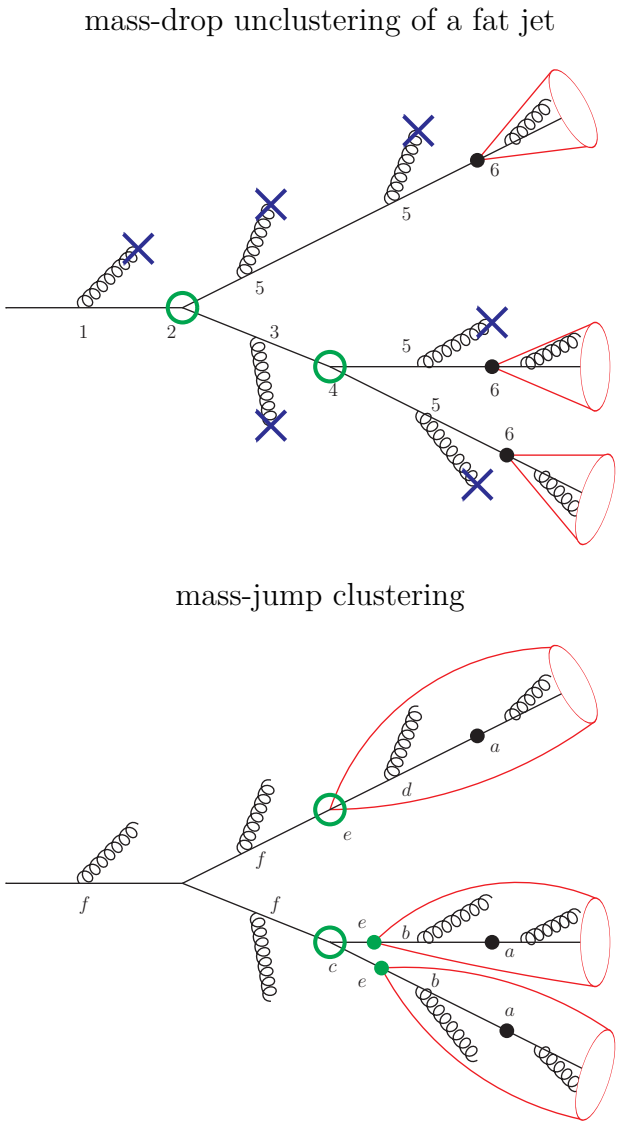

Figure 13: Illustration of the basic differences between mass drop unclustering (top) and mass jump clustering (bottom) for a schematic clustering sequence. Top: Starting from the constituents at the right hand side sequential recombination clusters these jets into the one fat jet on the left hand side. This fat jet is then unclustered again using the mass drop criterion. Bottom: Mass jump clusters the constituents on the right hand side into the unclustered jets. The straight lines symbolize hard prongs while the wiggled lines stand for soft radiation. Black dots signify the jet mass threshold m_{cut} and green circles mark a mass drop/mass jump. The final subjets are denoted by the red cones. Taken from [37].

clustering continues for active jet candidates (d). It is also possible that the active jet gets recombined with an passive jet (e). The remaining particles are also clustered, yielding additional jets (f).

The description of the algorithm makes it clear that the MJ algorithm does not reject a jet. As one can also see in Fig. 13, MJ keeps more radiation in the final subjets. Furthermore, it is expected, that there is more freedom in the choice of the parameter θ. A lower θ in MJ doesn't necessarily mean, that the substructures cannot be resolved anymore, since the jets are not dropped if the mass drop criterion is not fulfilled. Since $m_W/m_t \approx 0.47$, we expect MD to break down for values of $\theta \leq 0.4$, while MJ might still be able to work in this regime, since it is not dependent on the $t \to W^+ + b$ decay. However, for lower θ MJ will also break down, because at some threshold the mass drop criterion will lead to MJ clustering the object into one jet, combining all active jets with each other, because no mass jump ever occurs. Additional to the default mass drop algorithm, we also implemented a changed mass drop (cMD) algorithm, which if no mass drop occurred, keeps the softer jet, but does not decompose it further. Since we expect the original mass drop algorithm to achieve low efficiencies for small values of θ, we want to examine what changes, if we modify mass drop in this way and also compare this to the MJ algorithm. We expect the changed mass drop algorithm to be similar in its dependance on θ as the MJ algorithm. However, this changed mass drop will keep every substructure of the fat jet and therefore, we expect it to have even more freedom in the choice of the parameter θ. However since cMD keeps all jets it is also expected to have a lot more mistaggings.

Implementing the MJ algorithm in the HEPTopTagger is a bit more complicated than what we did when we studied the soft drop algorithm. As before the HEPTopTagger gets the clustered fat jets as an input. But for the MJ algorithm, we take the constituents of these fat jets and cluster them according to the MJ algorithm, which is a FASTJET-contrib (see [38]).

This way, we do the clustering of the fat jet two times. First, we cluster the event to fat jets of $R = 1.8$ in the C/A algorithm using Fastjet and then we take their constituents. This procedure is kind of redundant and excludes an important advantage the MJ algorithm might have: Since the MJ algorithm does the clustering and unclustering in one step, it should not be necessary to cluster the event into fat jets. Therefore, it should be possible to cluster the whole event using the MJ algorithm and giving these jets as hard substructures to the HEPTopTagger. In the last part of our study, we will consider the fully hadronic

$t\bar{t}H$ and thereby try to exploit this property of the MJ algorithm.

5.2.2 Performance

To test this algorithm we again use hadronic $t\bar{t}$ and QCD dijet Monte Carlo samples for the LHC at $\sqrt{s} = 13$ TeV generated with PYTHIA8 without multiple interactions. As starting point, we take C/A jets with $R_{\text{fat}} = 1.8$ constructed with FASTJET obeying $|\eta_{\text{fat}}| < 2.5$ and $p_{T,\text{fat}} > 200(600)$ GeV. For the signal we require that the fat jets can be matched to parton level tops within $\Delta R = 0.8$. The signal efficiencies are then given by the number of tags divided by the number of fat jets fulfilling these criteria. We test especially the dependance on the parameter θ and therefore run the algorithm for several values of θ and calculated the efficiencies.

algorithm	ϵ_S	ϵ_B
mass drop	0.3065 (0.4432)	0.0189 (0.0481)
changed mass drop	0.3137 (0.4670)	0.0197 (0.0495)
mass jump	0.3046 (0.4217)	0.0195 (0.0525)

Table 3: Efficiencies and mistagging rates for $\theta = 0.8$ at $p_T > 200(600)$ GeV for the different algorithms.

In Tab. 3, one can see the efficiencies as well as the mistagging rates at a mass drop threshold of $\theta = 0.8$. Apparently in these settings the MJ algorithm does not achieve a better efficiency than the mass drop algorithm (both default and changed).

The ROC curve as well as the signal and background efficiencies in dependance of θ are shown in Fig. 14. Again we took the granularity of the detector into account. From these curves one can see, that MJ remains for smaller θ much longer stable, i.e. it still achieves relatively high efficiencies. The same holds for the changed mass drop algorithm. However at about $\theta = 0.35$ the efficiency drops much faster for the MJ algorithm than for the changed mass drop algorithm. Even at $\theta \sim 0$ the changed mass drop algorithm still achieves good efficiencies. However, this way the mistags are also a lot higher, which makes sense, since changed mass drop keeps every subjet. In this regime, therefore, all these algorithms are not able to work well anymore.

To get a better idea of what happens at low θ and why MD is much more inefficient than MJ in this area, we analyze the hard substructures for events where the MJ algorithm tags the top quark and the mass drop algorithm does not.

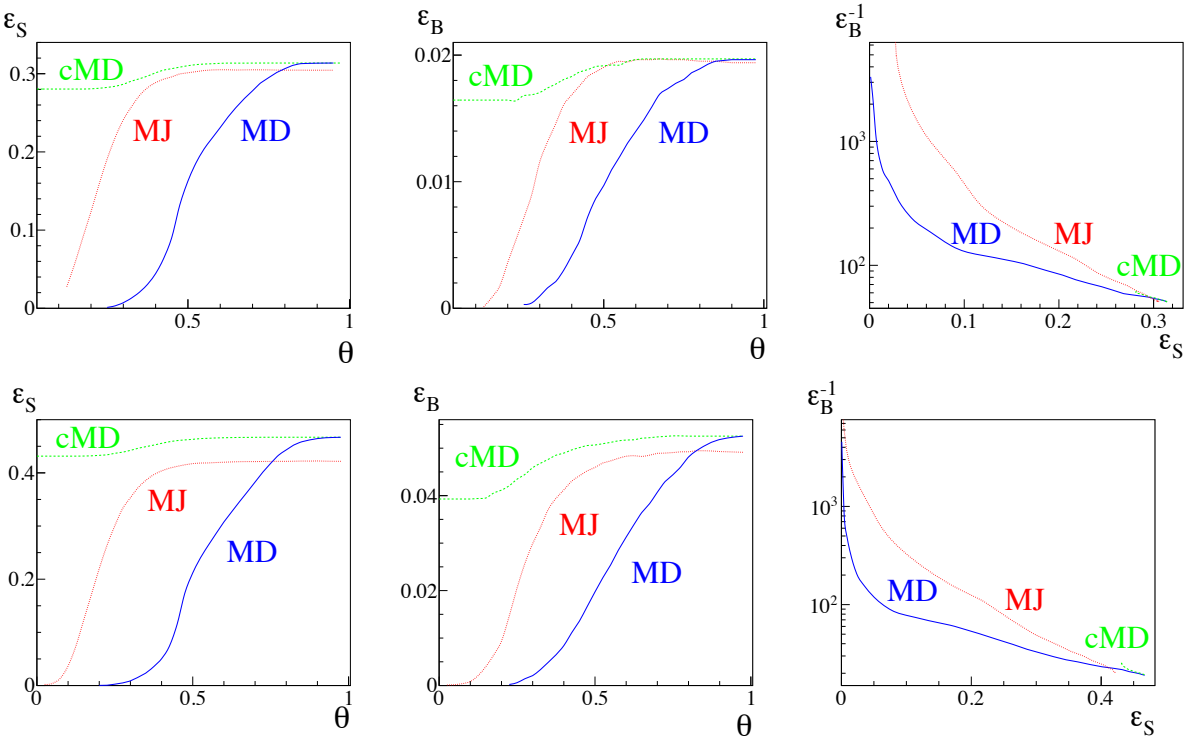

Figure 14: Signal and background efficiencies in dependance of θ (left) and the ROC curve (right), for $p_T > 200$ GeV (upper row) and $p_T > 600$ GeV (lower row).

In Fig. 15, the combined masses of the hard substructures are shown. On the left, one can see them for $\theta = 0.8$, where MD works well and for $\theta = 0.4$, where it breaks down. On the right, we show the same plot with only $\theta = 0.4$ and divided by the reconstructed top

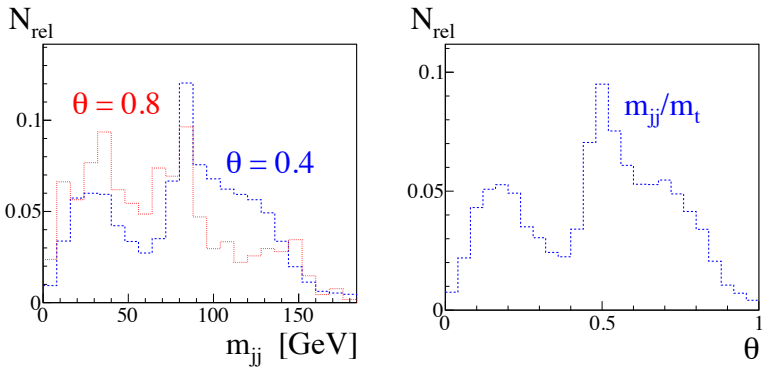

Figure 15: Analysis of events, where MJ tagged the top, but MD didn't. Left: Distribution of the combined masses of pairs of hard substructures for $\theta = 0.4, 0.8$ Right: The same distribution with $\theta = 0.4$ divided by the reconstructed top mass (right).

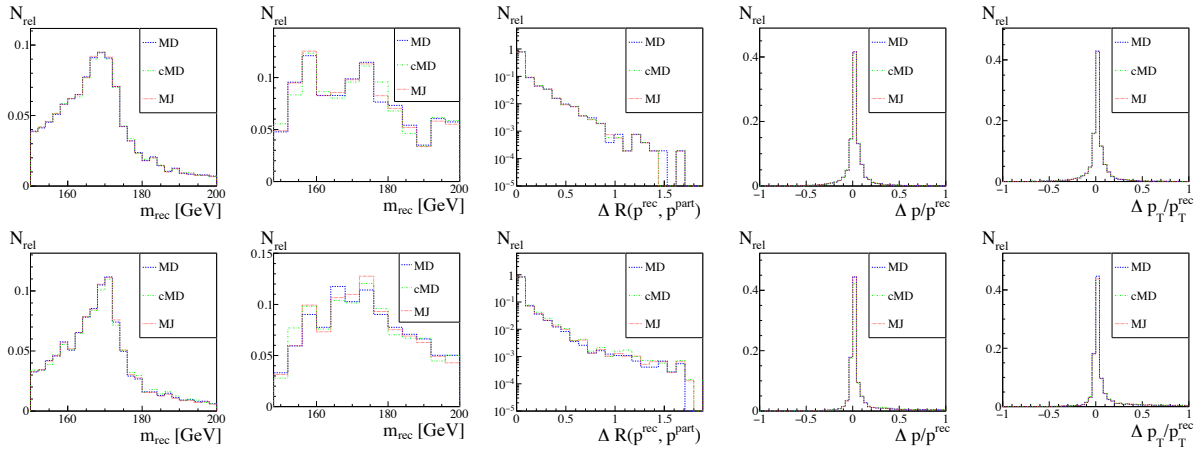

Figure 16: Quality of the reconstructed top quarks at $\theta = 0.8$ with $p_T > 200$ GeV (upper row) and $p_T > 600$ GeV (lower row)

mass.

One can see, that for low θ there is a peak at the combination of the hard substructures, which lies around the mass of the W. For $\theta = 0.8$ this peak is less prominent. This implies, that at low θ MD is not able anymore to resolve the mass splitting of $t \to W + b$ and only keeps the harder jet, i.e. the jet coming from the W. This becomes even more convincing when considering the plot on the right. The ratio of the combined masses of the substructures peaks at about 0.5. For $\theta = 0.4$ the mass drop criterion will be not fulfilled for these structures and the softer jet, will be dropped. For sufficiently small values of θ even the MJ algorithm fails to achieve high efficiencies, probably, because then θ is so low that even the mass jump of the W-decay is not resolved anymore and MJ clusters the everything into one jet. This is both in agreement with our expectations stated before.

Fig. 16 shows the mass distribution for the tagged and the mistagged top quarks as well as the deviations from the angular distance ΔR, the transverse momentum $\Delta p_T/p_T = (p_T^{\text{true}} - p_T^{\text{rec}})/p_T^{\text{rec}}$ and the three-momentum $\Delta|\vec{p}|/|\vec{p}| = (|\vec{p}_{\text{true}}| - |\vec{p}_{\text{rec}}|)/|\vec{p}_{\text{rec}}|$ of the reconstructed to the parton level top quark. The plots were made for at $\theta = 0.8$ with $p_T > 200$ GeV and $p_T > 600$ GeV. As one can see, for the default mass drop threshold $\theta = 0.8$, the quality of the reconstruction is basically the same. Therefore if we use this threshold, we do not get better results using the MJ algorithm.

5.2.3 Conclusions

This section showed, that a replacement of the mass drop algorithm by the MJ algorithm is only to be considered if one wants to have more freedom in the choice of the mass drop threshold θ. Since at $\theta = 0.8$ neither the efficiency nor the quality of the reconstruction have improved using the MJ algorithm it is not necessary to replace the mass drop algorithm by the MJ algorithm. Furthermore, the MJ algorithm takes about 1.27 times more computation time, which becomes apparent if one compares the complexity of the MJ algorithm with the one of the mass drop algorithm. Therefore, if one wants to achieve fast results and/or wants to study a lot of events, the MJ algorithms has the disadvantage of taking more time for basically the same results. Hence, we conclude that the mass drop algorithm of HEPTopTagger should not be replaced by the MJ algorithm.

However, as already mentioned, we disregarded an important property of the MJ algorithm by clustering the event into fat jets before using the MJ algorithm. This is not necessary, since the MJ algorithm can do the clustering by itself. Therefore, if one wants to get rid of the restriction $R = 1.8$ for fat jets, which is too small for top quarks with lower p_T, the MJ algorithm can be useful. We will try to exploit this advantage of the MJ algorithm in the next section.

5.3 Application of Mass Jump to fully hadronic $t\bar{t}H$

In the previous sections, we have seen that in tagging top quarks there is always this problem in finding a compromise between being able to include all the top quarks and keeping the jet size small enough to keep UE impurities small. In this section we will try to circumvent the need of a fat jet, which will allow us to target less boosted tops.

The Mass Jump algorithm seems to contain this property. As mentioned in the previous section, we left one important property of the MJ algorithm out of our analysis: The fact, that MJ clusters the hadronic activity of the event into its hard substructures. This way it is not needed to cluster the event into fat jets before applying MJ.

To examine this property of the MJ algorithm we will, in the following, apply it to the fully hadronic $t\bar{t}H$ process analyze the mass distribution for the Higgs as well as some backgrounds.

It should be mentioned, that there are already existing methods to deal with top quarks of low p_T. One possibility is proposed in Ref. [39]. There one first clusters the event into small jets and assigns them to three buckets, two for the hadronic tops and one for the two bottom tags of the Higgs decay or some combinatorial background. To do this one defines the measure

$$\Delta_{B_i} = |m_{B_i} - m_t| \quad \text{with} \quad m_{B_i}^2 = \left(\sum_{j \in B_i} p_j \right)^2 . \tag{16}$$

For each event one now permutes over all possible groupings of three buckets $\{B_1,\ B_2,\ B_{ISR}\}$. By minimizing this measure, one can get the wanted splitting into two top buckets and one bucket coming from either the Higgs or a combinatorial background.

5.3.1 The BDRS Higgs Tagger

For tagging the Higgs we use the BDRS Higgs Tagger [35]. It was developed for tagging the Higgs boson in the $H \to b\bar{b}$ channel and delivered the inspiration to construct a similar algorithm for tagging top quarks - the HEPTopTagger. Therefore the BDRS-Tagger works in a similar way as the HEPTopTagger.

In contrast to the HEPTopTagger the Higgs Tagger does not check diverse mass criteria to pick a candidate, but minimizes the difference between the combined mass of the bottom pairings and the Higgs mass $m_H = 125$ GeV.

5.3.2 Analysis

We generate fully hadronic $t\bar{t}H$ and $t\bar{t}b\bar{b}$ Monte Carlo samples for the LHC at $\sqrt{s} = 13$ TeV using MADGRAPH/MADEVENT and PYTHIA8 without multiple interactions. As a cut on p_T we set $p_{T,t} > 100$ GeV for the top quarks. Further we impose $p_{T,j} > 10$ GeV, $p_{T,b} > 10$ GeV, $\Delta R_{jj} > 0.1$ and $\Delta R_{bb} > 0.1$.

In our analysis, we first cluster the hadronic activity into jets of size $R = 0.4$ and $p_T > 20$ GeV. Then, we tag the bottom quarks by taking the jet which is closest to the parton level bottom quark. If we get 4 bottom tags, we divide the event into 4 areas, where each granulated hadron is assigned to the closest bottom quark.

Next, we cluster these areas using the MJ algorithm and run the HEPTOPTAGGER over these areas. The difference in the use of the HEPTOPTAGGER here is, that it takes the MJ jets directly as hard substructures. In contrast to the previous section, MJ does not take already clustered jets but clusters directly the granulated activity.

If the HEPTOPTAGGER achieves two top tags we remove the areas where it tagged the tops and let the BDRS Higgs Tagger run over the combination of the remaining areas, after we clustered them using MJ. In this way, by first tagging the tops and removing the related areas from the event, it should be possible to successfully reduce the combinatorial backgrounds.

We are only considering the $t\bar{t}b\bar{b}$ background. Taking all backgrounds into account would be beyond the scope of this study, which is only meant to be a test of concept. In later studies especially the top mistagging should be examined.

5.3.3 Results

When running the analysis, we again calculated the efficiency of the tagging. We accepted only events with $p_{T,t} > 100$ GeV. Then we computed the efficiency ε_{4b} for achieving four bottom tags, the efficiency ε_t for a top tag, the efficiency ε_{2t} for two top tags and the efficiency ε_H for a Higgs tag. Altogether these give the total efficiency $\varepsilon_{\text{tot}} = \#\{\text{Higgs tags}\}/\#\{\text{events}\}$.

These efficiencies were computed for the $t\bar{t}H$ as well as the $t\bar{t}b\bar{b}$ process. To examine, if and how much this improves the usual top tagging, we also computed these efficiencies by running the analysis using the C/A algorithm instead of MJ to cluster the event into fat jets.

The results are shown in Tab. 4.

Clustering	process	ε_{4b}	ε_t	ε_{2t}	ε_H	ε_{tot}
MJ	$t\bar{t}H$	0.432	0.524	0.123	0.708	0.0196
C/A	$t\bar{t}H$	0.432	0.322	0.07	0.271	0.00261
MJ	$t\bar{t}b\bar{b}$	0.185	0.601	0.142	0.588	0.0093
C/A	$t\bar{t}b\bar{b}$	0.186	0.356	0.092	0.087	0.00053

Table 4: Efficiencies of the top and Higgs tagging, for the $t\bar{t}H$ and the $t\bar{t}b\bar{b}$ process by using MJ or C/A for the clustering.

These results show, that, for low p_T, using the MJ clustering algorithm achieves significantly higher efficiencies than the C/A clustering, as we expected it to be. Therefore, regarding the tagging efficiency, MJ seems to really be a improvement to the top tagging algorithms in this regime.

However, the mistagging rate using the MJ clustering is a lot higher than for the C/A clustering. The reason for this is, that the background achieves a high Higgs tagging rate for the MJ clustering. Therefore, later studies should also focus on the mistagging rate and find ways to reduce this high Higgs tagging rate of the background.

In Fig. 17, the mass distribution of the reconstructed Higgs bosons and top quarks as well as the deviation of their reconstructed p_T from the one on parton level, are shown. Apparently the mass distribution of the Higgs peaks around 125 GeV. In this region the background is decreasing, which allows a separation of signal and background. Also the deviation of the transverse momentum peaks around zero, indicating a good reconstruction.

Therefore, we can conclude from this, that the MJ clustering algorithm is in fact suited for the study of top quarks at $p_T < 200$ GeV. Using the MJ clustered jets, the HEP-TopTagger and the BDRS Higgs Tagger achieve significantly higher efficiencies than the usual fat jet clustering. However, to compare it to other alternatives, like the bucket algorithm, is beyond the scope of this study.

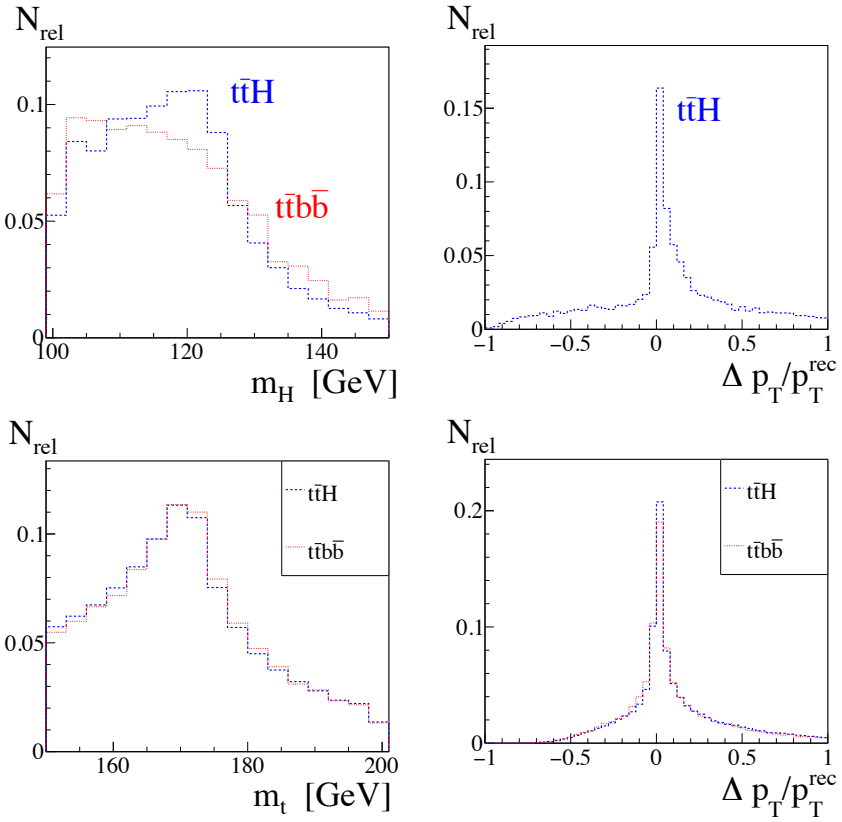

Figure 17: Mass distribution and difference in p_T for the reconstructed Higgs (upper row) as well as the reconstructed top quarks (lower row).

6 Conclusion

In this thesis we studied two alternative ways for unclustering fat jets using the HEPTopTagger, Soft Drop and Mass Jump. It was shown, that both algorithms, at least used in this way, do not lead to significant improvements. The tagging efficiency as well as the quality of reconstruction were comparable to the mass drop algorithm of the HEPTopTagger. In Tab. 5, the efficiencies for all the tested algorithms are summarized.

algorithm	ϵ_S	ϵ_B
mass drop	0.3065 (0.4432)	0.0189 (0.0481)
Soft Drop:		
$\beta = -0.5$	0.3137 (0.4653)	0.0197 (0.0519)
$\beta = -1$	0.3135 (0.4614)	0.0197 (0.0502)
$\beta = -1.5$	0.3124 (0.4159)	0.0196 (0.0400)
mass jump	0.3046 (0.4217)	0.0195 (0.0525)

Table 5: Efficiencies for the examined algorithms at $p_T > 200(600)$ GeV for optimized parameter choices: $z_{\text{cut}} = 0.01$ for SD and $\theta = 0.8$ for MD and MJ.

If one uses the Mass Jump algorithm for clustering instead of the usual clustering into fat jets and gives the result directly as top candidates to the HEPTopTagger this might be an improvement. Besides that, we don't recommend to include these alternatives in the HEPTopTagger.

Also the fully hadronic $t\bar{t}H$ process was studied using the Mass Jump algorithm. It turned out, as expected, that for $p_T < 200$ GeV the MJ algorithm is still very effective, since it delivers an alternative to the usual fat jet clustering.

We were able to successfully reconstruct the masses of the Higgs and the top quarks of this process.

In future projects one can have a closer look on these improvements, the MJ algorithm offers by comparing them to other alternatives, like the bucket algorithm. This might achieve great progress in tagging tops at low p_T.

Furthermore, since the mistagging rate of using the MJ clustering is relatively higher than for C/A clustering, later studies should examine this and try to solve this problem. One possibility to do this would be to extend the restriction on the mass window. Especially at masses below 120 GeV there are lots of mistags.

References

[1] S. L. Glashow, "Partial Symmetries of Weak Interactions," Nucl. Phys. **22**, 579 (1961).

[2] S. Weinberg, "A Model of Leptons," Phys. Rev. Lett. **19**, 1264 (1967).

[3] A. Salam, "Weak and electromagnetic interactions," in Elementary particle theory: relativistic groups and analyticity, N. Svartholm, ed., p.367. Almqvist & Wiskell, 1968. Proceedings of the eighth Nobel symposium.

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

[5] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons," Phys. Rev. Lett. **13**, 508 (1964).

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

[7] G. Aad *et al.* [ATLAS Collaboration], "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," Phys. Lett. B **716**, 1 (2013) [arXiv:1207.7214 [hep-ex]].

[8] S. Chatrchyan *et al.* [CMS Collaboration], "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," Phys. Lett. B **716**, 30 (2012) [arXiv:1207.7235 [hep-ex]].

[9] Y. Fukuda *et al.* [Super-Kamiokande Collaboration], "Evidence for oscillation of atmospheric neutrinos," Phys. Rev. Lett. **81**, 1562 (1998) [hep-ex/9807003].

[10] M. Kobayashi and T. Maskawa, "CP Violation in the Renormalizable Theory of Weak Interaction," Prog. Theor. Phys. **49**, 652 (1973).

[11] F. Abe *et al.* [CDF Collaboration], "Observation of top quark production in $\bar{p}p$ collisions," Phys. Rev. Lett. **74**, 2626 (1995) [hep-ex/9503002].

[12] R. Aaij *et al.* [LHCb Collaboration], "Observation of J/p Resonances Consistent with Pentaquark States in $^{0}_{b}$ $J/K^{-}p$ Decays," Phys. Rev. Lett. **115**, 072001 (2015) [arXiv:1507.03414 [hep-ex]].

[13] T. Ohl, "Drawing Feynman diagrams with Latex and Metafont," Comput. Phys. Commun. **90**, 340 (1995) [hep-ph/9505351].

[14] Wikipedia Commons, `https://upload.wikimedia.org/wikipedia/commons/0/00/Standard_Model_of_Elementary_Particles.svg` (September 9, 2015)

[15] M. Carena, C. Grojean, M. Kado, V. Sharma "Status of Higgs boson physics" `http://pdg.lbl.gov/2013/reviews/rpp2013-rev-higgs-boson.pdf`

[16] Donald H. Perkins, *Introduction to High Energy Physics*, Cambridge University Press, 4th edition, 2000.

[17] T. Plehn, "Lectures on LHC Physics," Lect. Notes Phys. **886** (2015).

[18] T. Han, "Collider phenomenology: Basic knowledge and techniques," hep-ph/0508097.

[19] Cheuk-Yin Wong *Introduction to high-energy heavy-ion collisions* World Scientific, 1994

[20] M. Dasgupta, L. Magnea and G. P. Salam, "Non-perturbative QCD effects in jets at hadron colliders," JHEP **0802**, 055 (2008) [arXiv:0712.3014 [hep-ph]].

[21] T. Plehn and M. Spannowsky, "Top Tagging," J. Phys. G **39**, 083001 (2012) [arXiv:1112.4441 [hep-ph]].

[22] Y. L. Dokshitzer, G. D. Leder, S. Moretti and B. R. Webber, "Better Jet Clustering Algorithms," JHEP **9708**, 001 (1997); M. Wobisch and T. Wengler, "Hadronization corrections to jet cross sections in deep-inelastic scattering," arXiv:hep-ph/9907280.

[23] S. Catani, Y. L. Dokshitzer, M. H. Seymour and B. R. Webber, "Longitudinally invariant K_t clustering algorithms for hadron hadron collisions," Nucl. Phys. B **406**, 187 (1993).

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

[25] G. Aad *et al.* [ATLAS Collaboration], "Measurement of inclusive jet and dijet production in pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector," Phys. Rev. D **86**, 014022 (2012) [arXiv:1112.6297 [hep-ex]].

[26] A. Buckley, J. Butterworth, S. Gieseke, D. Grellscheid, S. Hoche, H. Hoeth, F. Krauss and L. Lonnblad *et al.*, "General-purpose event generators for LHC physics," Phys. Rept. **504**, 145 (2011) [arXiv:1101.2599 [hep-ph]].

[27] T. Sjöstrand *et al.*, "An Introduction to PYTHIA 8.2," Comput. Phys. Commun. **191**, 159 (2015) [arXiv:1410.3012 [hep-ph]].

[28] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer and T. Stelzer, "MadGraph 5 : Going Beyond," JHEP **1106**, 128 (2011) [arXiv:1106.0522 [hep-ph]].

[29] M. Cacciari and G. P. Salam, "Dispelling the N^3 myth for the k_t jet-finder," Phys. Lett. B **641**, 57 (2006); [arXiv:hep-ph/0512210]. M. Cacciari, G. P. Salam and G. Soyez, `http://fastjet.fr`

[30] T. Plehn, G. P. Salam and M. Spannowsky, "Fat Jets for a Light Higgs," Phys. Rev. Lett. **104**, 111801 (2010) [arXiv:0910.5472 [hep-ph]].

[31] T. Plehn, M. Spannowsky, M. Takeuchi and D. Zerwas, "Stop Reconstruction with Tagged Tops," JHEP **1010**, 078 (2010) [arXiv:1006.2833 [hep-ph]].

[32] T. Plehn, M. Spannowsky and M. Takeuchi, "How to Improve Top Tagging," Phys. Rev. D **85**, 034029 (2012) [arXiv:1111.5034 [hep-ph]].

[33] C. Anders, C. Bernaciak, G. Kasieczka, T. Plehn and T. Schell, "Benchmarking an even better top tagger algorithm," Phys. Rev. D **89**, no. 7, 074047 (2014) [arXiv:1312.1504 [hep-ph]].

[34] G. Kasieczka, T. Plehn, T. Schell, T. Strebler and G. P. Salam, "Resonance Searches with an Updated Top Tagger," JHEP **1506**, 203 (2015) [arXiv:1503.05921 [hep-ph]].

[35] J. M. Butterworth, A. R. Davison, M. Rubin and G. P. Salam, "Jet substructure as a new Higgs search channel at the LHC," Phys. Rev. Lett. **100**, 242001 (2008) [arXiv:0802.2470 [hep-ph]].

[36] A. J. Larkoski, S. Marzani, G. Soyez and J. Thaler, "Soft Drop," JHEP **1405**, 146 (2014) [arXiv:1402.2657 [hep-ph]].

[37] M. Stoll, "Vetoed jet clustering: The mass-jump algorithm," JHEP **1504**, 111 (2015) [arXiv:1410.4637 [hep-ph]].

[38] `http://fastjet.hepforge.org/svn/contrib/contribs/ClusteringVetoPlugin/tags/1.0.0/`

[39] M. R. Buckley, T. Plehn, T. Schell and M. Takeuchi, "Buckets of Higgs and Tops," JHEP **1402**, 130 (2014) [arXiv:1310.6034 [hep-ph]].

Acknowledgements

First of all, I would like to thank my supervisor Tilman Plehn for making this project possible in the first place and giving me the opportunity to work in his group and thereby get first insights into the everyday life of a physicist. Furthermore, I have to thank Torben Schell for guiding me through this project and helping me whenever I needed it. Finally, I thank the whole group for the kindness and openness you offered.

Erklärung

Ich versichere, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 26.10.2015,

_____

Josua Göcking