University of Heidelberg

Department of Physics and Astronomy

Institute for Theoretical Physics

Bachelor Thesis in Physics

Testing the HEPTopTagger for moderately boosted tops

---

submitted by
Rabea Link
born in Bietigheim-Bissingen
(Germany)

Supervisor
Prof. Dr. Tilman Plehn

This bachelor thesis has been carried out by Rabea Link
at the Institute for Theoretical Physics Heidelberg
under the supervision of Prof. Dr. Tilman Plehn.

November 28, 2016

Heidelberg

Abstract

The HEPTopTagger is a tool that was developed to detect top quarks in decay processes at collider experiments. In this work, we will focus on the performance of the HEPTopTagger in a $t\bar{t}H$ production, with one of the tops decaying hadronically ($t \to bjj$) and the other one decaying leptonically ($t \to b\ell\nu_\ell$). The Higgs decays to bottom quarks ($H \to b\bar{b}$). The goal is to improve HEPTopTagger in the region of the top transverse momenta 150 GeV $< p_{T,\text{top}} <$ 250 GeV with the Mass Jump algorithm.

We compare the perfomance of HEPTopTagger with standard fat jets as input to two different approaches: First, we divide the event into four areas that are determined by the position of the b-jets and feed the Mass Jump jets of each area separately into HEPTopTagger. Second, we use Mass Jump jets from the whole event as input. We find that with the area analysis, we can improve the quality of the top reconstruction but decrease the tagging efficiency and the reconstruction quality of the Higgs, and that with the pure Mass Jump, the analysis is limited by high combinatoric.

Zusammenfassung

Der HEPTopTagger wurde entwickelt, um zerfallende Top-Quarks an Teilchenbeschleunigern nachweisen zu können. Wir testen den HEPTopTagger in $t\bar{t}H$ Prozessen, bei denen eines der Top-Quarks hadronisch zerfällt ($t \to bjj$), und das andere leptonisch ($t \to b\ell\nu_\ell$). Das Higgs zerfällt zu Bottom-Quarks ($H \to b\bar{b}$). Das Ziel ist es, herauszufinden, ob man den HEPTopTagger für transverse Impulse in der Region 150 GeV $< p_{T,\text{top}} <$ 250 GeV mithilfe von Mass Jump Jets verbessern kann.

Dazu werden wir prüfen, wie gut die Leistung des HEPTopTagger im Vergleich zu der Standard-Methode mit „fat jets“ in zwei unterschiedlichen Ansätzen ist: Für den ersten Ansatz teilen wir das Event in vier Sektionen auf, die durch die Position der vier b-Jets definiert werden, und lassen den HEPTopTagger über jede Sektion separat laufen. Für den zweiten Ansatz werden wir das gesamte Event mit Mass Jump Jets gruppieren und dann diese für den HEPTopTagger verwenden. Die Sektionen-Analyse verbessert die Qualität der Top-Rekonstruktion leicht, rekonstruiert das Higgs aber schlechter und die Analyse, bei der nur Mass Jump Jets verwendet werden, ist aufgrund von Kombinatorik nicht zuverlässig.

Contents

1 Introduction

The Standard Model of particle physics [1] describes all the fundamental particles that are known today and their interactions, except for gravity. Even though the Standard Model has been experimentally proven to be a good description of fundamental interactions, there are still open questions, including the search for dark matter and the fact that neutrino oscillations are experimentally established, which suggest that there is new physics beyond the Standard Model (BSM). One way to find new particles are collider experiments like the Large Hadron Collider (LHC) at CERN.

There are several particles, like the Higgs boson or the top quark, which decay to other particles before they can be measured by detectors. There has been great effort and success in reconstructing those particles from the visible final state particles using various algorithms based on physical properties. The challenge of all tagging algorithms is to identify the correct particles while excluding background events to minimize mis-tagging. The first reconstruction algorithm that is based on jet substructure was developed in 1994 and was able to tag W bosons and tops [2]. Since then, the tagging algorithms have been improved constantly. Today, a variety of tagging algorithms exist for different settings and particles to provide the best possible particle reconstruction for experiments at the LHC.

We will focus on the HEPTopTagger in this work. HEP stands for Heidelberg, Eugene and Paris; the locations where the tagger was developed by Tilman Plehn, Gavin P. Salam and Michael Spannowsky [3]. The HEPTopTagger is a jet substructure algorithm that uses mass cuts on the decay products. It focuses in particular on moderate p_T ranges of the top ($p_{T\text{top}} > 200$ GeV).

The goal of this work is to check with simulations whether we can extend the accessible phase space in the $t\bar{t}H$ process towards lower top boosts (150 GeV $< p_{T\text{top}} <$ 250 GeV) using the HEPTopTagger and the Mass Jump algorithm.

In Chapter 2, we briefly discuss the theoretical background needed to understand the work we did. This includes the successes and the limitations of the Standard Model. We focus on the characteristics of top quarks in particular, since they are of special interest to this work.

In Chapter 3, we describe the usual setup of a collider and introduce conventions and ideas used in phenomenology such as transverse momentum, Mandelstam variables, jet clustering and QCD background effects at hadron colliders. Furthermore, we give a detailed description of the mass drop and the Mass Jump algorithm as well as the

HEPTopTagger algorithm. We also list and explain the simulation programs we used.

Then, we focus on improving the HEPTopTagger for $t\bar{t}H$ production with $H \to b\bar{b}$ and semileptonic $t\bar{t}$ decays. We will compare three different approaches: the standard analysis with two fat jets, an area analysis where we divide the event in four areas and a pure Mass Jump analysis, where we only work with clustered Mass Jump jets. We discuss the different analyses in depth in Chapter 4.

Finally, in Chapter 5 we present the results we found. We conclude that the area algorithm improves the quality of the top reconstruction slightly over the standard analysis, but decreases the tagging efficiency and the quality of the Higgs reconstruction. The pure Mass Jump algorithm does not work reliably due to the high combinatoric of the event.

2 Standard Model

The Standard Model is a great success of modern physics. It incorporates quantum electrodynamics, the electroweak theory developed by Glashow, Weinberg and Salam, and quantum chromodynamics. The gauge interactions in the Standard Model follow the mathematical structure of an $SU(3)_C \times SU(2)_L \times U(1)_Y$ group. Here, $SU(3)_C$ represents the symmetry of quantum chromodynamics, left-handed fermions carry weak isospin numbers and $U(1)_Y$ represents the hypercharge symmetry. With the Standard Model, physicists have correctly predicted many particles that were found during the last 40 years. Those were the gluon (1979), the $W^{\pm}$ and the Z bosons (1983), as well as the top quark (1995), τ neutrinos (2000) and the Higgs boson (2012).

2.1 The Particles of the Standard Model

There are three different kinds of elementary particles in the Standard Model: leptons, quarks and gauge bosons. Leptons and quarks are so-called fermions since they have spin $\frac{1}{2}$. Bosons have integer spin.

- Leptons ℓ can be classified into three generations: the first generation is the electron e and the electron neutrino ν_e. The muon μ and the muon neutrino ν_μ are the second generation and the third generation consists of the tau τ and the tau neutrino ν_τ. Left-handed leptons carry weak isospin but no color quantum numbers, explaining why they participate in weak but not in strong interactions.

- The quarks q are divided into three generations: The first generation consists of the up u and down d quark, which can form bound states, for instance protons and neutrons. The strange s and charm c quark form the second generation and the third generation consists of top t and bottom b quark. The top quark is the heaviest particle in the Standard Model. Quarks can interact with all gauge bosons and with the Higgs boson.

- Furthermore, there are gauge bosons in the Standard Model: the photon γ for the electromagnetic force, which mediates interactions between charged particles, the $W^{\pm}$ and the Z boson for the weak interactions between particles with weak isospin, and eight gluons g for the strong interaction between colored particles.

- Finally, there is the Higgs field H, whose vacuum expectation value generates the masses of massive particles in the Standard Model because of spontaneous symmetry breaking.

In the Standard Model, all fundamental interactions can be derived only from the requirement of local gauge invariance, Lorentz invariance and renormalizability. Since, on the one hand, those are very fundamental features, and, on the other hand, the differences between the predictions made with the Standard Model and the measurements deviate by no more than 3σ, it is likely that the Standard Model is correct at the scale that we measure. But there are still a lot of questions that are not answered yet. Therefore, it is expected that the Standard Model has to be expanded by a new theory.

2.2 Limitations of the Standard Model

Even though the Standard Model is extremely successful in describing experiments we see in particle physics, there are several results we cannot currently explain with it. Specifically:

- When looking at the rotational velocities of galaxies, it becomes obvious that there should be more matter than what is seen in γ-ray telescopes. This, as well as other experimental observations, leads to the conclusion that about 85% of the matter density in the universe is non-baryonic [4].

- In the Standard Model, neutrinos are massless and therefore they interact diagonally in flavor space, i.e. they cannot change their flavor. But neutrino oscillations are experimentally established, which leads to the conclusion that at least two neutrinos must have non-zero masses [5].

- When calculating the next to leading order corrections of the Higgs mass, one finds that the Higgs mass should be at a high scale, the so-called Planck scale, rather than the electroweak scale where it was found experimentally. However, the Higgs mass can be renormalised to the electroweak scale with fine-tuning. But if one introduces a heavy particle that couples to the Higgs, there are correction terms to the Higgs mass that cannot be renormalised [6].

These and other pieces of evidence lead to the conclusion that the Standard Model is incomplete and that there is more to discover. The search for new physics beyond the Standard Model (BSM) can be done with high energy experiments, an example being the LHC at CERN.

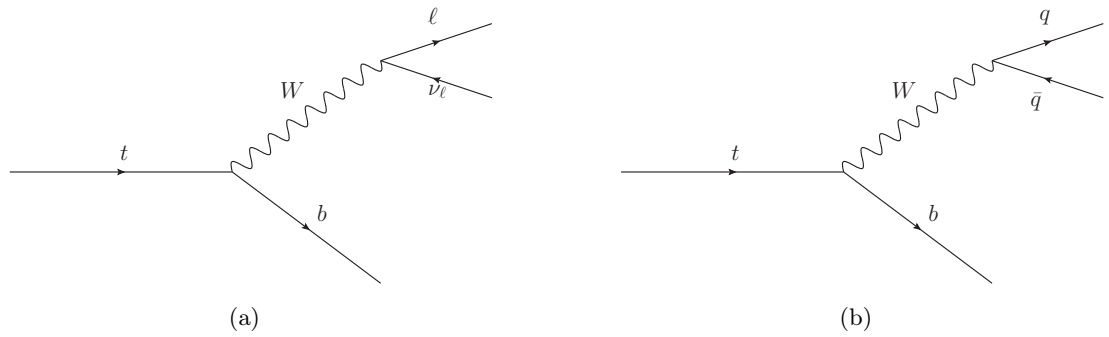


Figure 1: Feynman diagrams for leptonically (a) and hadronically (b) decaying tops. The top decays into a W boson an a bottom quark. Dependent on the leptonic or hadronic decay of the W, the top decay is called leptonic or hadronic.

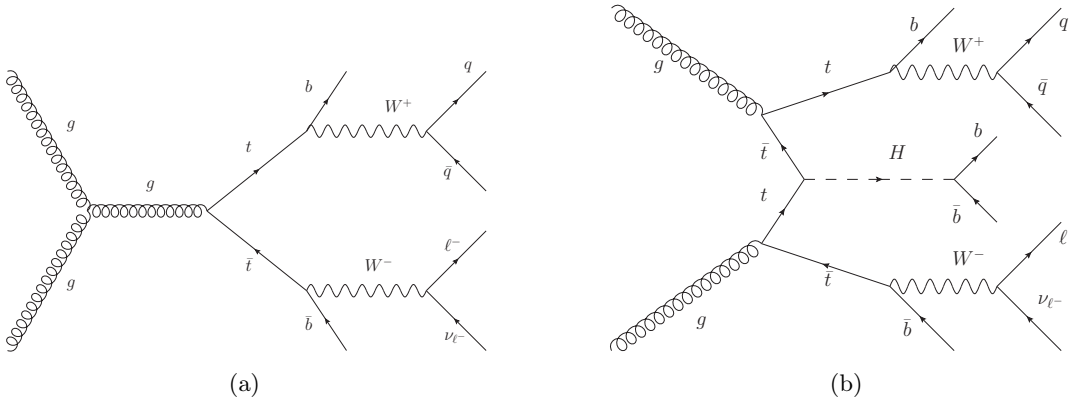


Figure 2: Example processes of the $t\bar{t}$ (a) and $t\bar{t}H$ (b) production from gluons. Tops decay into a W and a b in most cases, and we assume that the Higgs decays into b quarks.

2.3 The Top Quark

The top quark has charge $Q = +\frac{2}{3}$ and weak isospin $T_3 = +\frac{1}{2}$. It makes up a weak isospin doublet with the bottom quark. The top quark is of particular interest, because it is very heavy. With a mass of $m_t = (172.44 \pm 0.13 \pm 0.47)$ GeV [7], it is the heaviest particle in the Standard Model. In almost all cases, the top decays into a W boson and a b quark. Because of the large top mass, it has a very short lifetime of order of 10^{-25} s. Hadronization takes place at about 10^{-23} s, which is why the top is not found in hadrons.

We focus on the $pp \to t\bar{t}H$ associated production of a Higgs boson with a top-antitop quark pair, where one of the tops decays hadronically $t \to Wb$, $W \to jj$ and the other one decays leptonically $t \to Wb, W \to \ell\nu_\ell$; $\ell = e, \mu$. The Higgs decays into a bottom pair, $H \to b\bar{b}$. We focus on this process since the Higgs is not yet detected in the $t\bar{t}H$ channel.

Measuring this process is of interest because it is the largest direct probe of the Yukawa coupling. The Yukawa coupling describes interactions between fields. In the case of the $t\bar{t}H$ process, the Yukawa interaction can be described by the interaction Lagrangian between quarks and the Higgs field ϕ

$$\mathscr{L}_y = -y_d \bar{Q}_L \phi d_R - y_u \bar{Q}_L \tilde{\phi} u_R + \text{h.c.} \tag{1}$$

where $y_{d,u}$ is the Yukawa coupling, Q_L is the left-handed quarks doublet and d_R and u_R are the right-handed quark singlets. The interaction between the tops and the Higgs therefore is a ~ 1 probe of our value of the Yukawa coupling and of our understanding of the electroweak symmetry breaking. Measuring the Yukawa coupling is crucial when searching for possible extensions to the electroweak breaking mechanism [8].

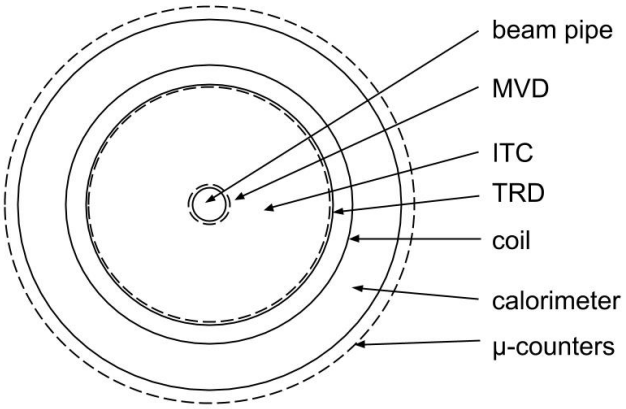


Figure 3: The schematic structure of a detector when looking along the beam axis. Detectors usually have a cylindric form with their center axis identical to the beam axis. This ensures the highest possible detection rate. Figure taken from [9].

3 Phenomenology

Generally speaking, the purpose of phenomenology in high-energy physics is to compare experimental results to theoretical models. Many particles cannot be detected directly. The reason for this is that the lifetime of the particle is too short and it decays before it can be measured by a detector, it forms bound states with other particles (hadronization) or it escapes the detector. It is usually not clear which intermediate steps happened during the collision. To reconstruct quarks and gluons in the collider, one uses jet algorithms that are based on physical quantities of the intermediate states we expect. In this section, we will explain the general setup of a high energy collider, the mathematical tools that are typically used to describe the collision and the jet algorithms we use. This section follows closely [3, 9–12].

3.1 Collider Physics

Detectors: As one can see in Fig. 3, colliders are built of different layers of detectors around a beam pipe. The closest detector layer around the beam is the microvertex detector (MVD), which aims to detect vertices of short-lived particles. The inner tracking chamber (ITC) can record tracks of charged particles. Then, a transition radiation detector (TRD) can efficiently identify electrons. Magnet coils are used to build a solenoidal magnetic field. This makes it possible to measure momenta in the ITC. The calorimeter then measures energies based on particle showers produced by the outcoming particles of the collision. Since muons are likely to pass all the inner detectors, the muon counter

is the last layer of the detector. Even though colliders are well-adapted to measure the outcome of a particle collision, they can only identify the following:

- Jets are constructs that can be related to hadrons and partons, and their energy can be seen in a geometrically localised peak.
- Electrons and muons can be measured in the ITC with their momentum and their sign. The TRD and the muon tracker help to conclude which particle was seen.
- Photons produce a calorimeter signal similar to an electron, but they do not show up in the ITC.
- τ leptons or hadrons containing c or b quarks can be identified in the MVD because of a specific topological signature, the so-called displaced vertices.
- The transverse momenta of all particles can be measured, and if transverse momentum is missing, this calls for neutrinos (or detector effects).

Mandelstam Variables: For $AB \to CD$ scattering events, Mandelstam variables s, t and u are often used to describe the kinematic features of a process. They are defined by

$$\begin{aligned} s &\equiv (p_1 + p_2)^2 = (p_3 + p_4)^2, \\ t &\equiv (p_1 - p_3)^2 = (p_2 - p_4)^2, \\ u &\equiv (p_1 - p_4)^2 = (p_2 - p_3)^2, \end{aligned} \tag{2}$$

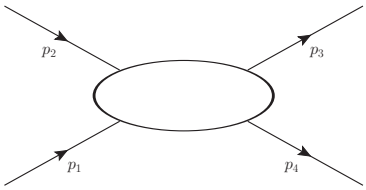


Figure 4: The Feynman diagram for a general $2 \to 2$ process. p_1 and p_2 are the momenta of the incoming particles, p_3 and p_4 are the momenta of the outgoing particles.

where p_1 and p_2 describe the four-momenta of the initial particles and p_3 and p_4 describe the four-momenta of the outgoing particles. Mandelstam variables have several advantages. They fulfill $s + t + u = \sum m_j^2$, where m_j are the masses of the particles, and they are Lorentz invariant.

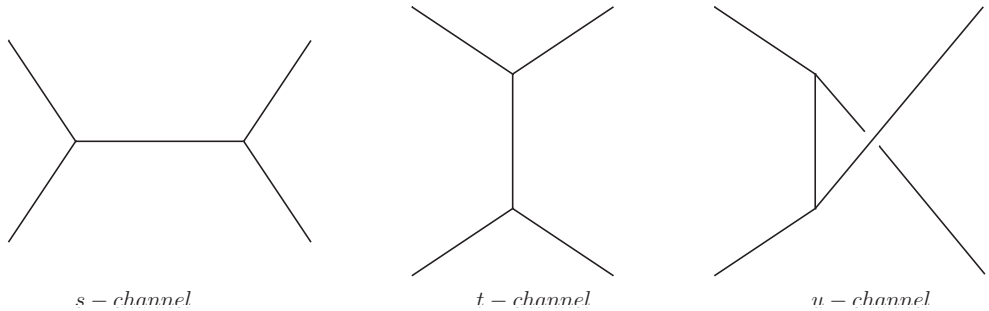


Figure 5: Different Feynman diagrams that illustrate the idea of Mandelstam variables. The incoming particles are always on the left, the outgoing particles on the right side. The matrix element of each process contains the respective mandelstam variable.

Transverse Momentum: The transverse momentum is defined by

$$p_T = \sqrt{p_x^2 + p_y^2}. \tag{3}$$

The reason for the introduction of the transverse momentum is the fact that we do not know the center of mass system of the collision. Since the constituents of the proton, the so-called partons, have different and unknown velocities, it is impossible to calculate the center of mass and to use the usual cms-frame. Furthermore, as we have seen before in Section 3.1, the typical collider is cylindric around the beam pipe. This makes it impossible to measure all the kinematics of the event. In the direction of the beam pipe, which is usually identified with the z-axis, it is impossible to detect any radiation. Note that another advantage of the transverse momentum is the fact that it is invariant under longitudinal Lorentz boosts.

Pseudorapidity: Now we are able to describe momenta in the setting of a collider. Another useful variable is the rapidity y. It is related to the angle between a particle and the beam. It is defined as

$$y = \frac{1}{2} \ln\left(\frac{E + p_z}{E - p_z}\right). \tag{4}$$

The rapidity y is additive under longitudinal Lorentz boosts. In the limit of massless particles, we find the pseudorapidity η give by

$$\eta = \lim_{m \to 0} y = \frac{1}{2} \ln\left(\frac{|\vec{p}| + p_z}{|\vec{p}| - p_z}\right) = \operatorname{artanh}\left(\frac{p_z}{|\vec{p}|}\right) = -\ln\left(\tan\frac{\theta}{2}\right). \tag{5}$$

η only depends on the angle θ, which is the angle between $\vec{p}$ and the z-direction. The pseudorapidity is often used as part of a distance measure ΔR in particle physics, i.e. in jet reconstruction. ΔR depends only on angular coordinates and is defined by

$$\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}, \tag{6}$$

where ϕ denotes the azimuthal angle. It is invariant under boosts in the $\vec{z}$-direction if one assumes massless particles.

Since we do not have any information about the z -axis of our event, it is useful to introduce the $\eta - \phi -$ plane. One can think of the $\eta - \phi -$ plane as cutting along the cylindric detector parallel to its $z -$ axis. The advantage is that now our final state is in a two-dimensional setting, which makes it a lot easier to calculate distances between particles.

JADE Distance: The JADE distance $d_{j_1 j_2}$ for hadron colliders [13] measures the transverse mass,

$$d_{j_1 j_2} = p_{T,j_1} p_{T,j_2} (\Delta R_{j_1 j_2})^2 \sim m^2_{T,j_1 j_2}. \tag{7}$$

The modified JADE distance $d^{\text{mod}}_{j_1 j_2}$ [14] increases the importance of the geometrical distance between the two subjets:

$$d^{\text{mod}}_{j_1 j_2} = p_{T,j_1} p_{T,j_2} (\Delta R_{j_1 j_2})^4. \tag{8}$$

3.2 Jet Algorithms

Many particles decay or hadronize before one can detect them, thus one needs to reconstruct the original particle from its decay products. Ideally, one would know which measured particles belong together and simply add up their four-momenta to find the properties of the particle that one is actually interested in. But due to final state radiation, it is not even clear how many final state particles one expects to find. Despite those facts it is possible to reconstruct the particles one is interested in by means of jet algorithms.

Jet algorithms use a collinearity measure to find the final state particles that originate from the same quark or gluon. The algorithm iterates over all subjets and clusters the hadronic activity of the event to jets with properties similar to the properties of the original quarks and gluons. With this, a jet algorithm reveals the structure of the event.

There are several algorithms that use different approaches to measure the distance between two subjets. The most used algorithms are the k_T-algorithm [15] , the C/A-algorithm [16] and the anti-k_T-algorithm [17]. Their measures depend on the distance ΔR as defined in Eq. 6 as well as the transverse momentum p_T of the subjets.

For the different reconstruction algorithms, the distance measure between two subjets y_{ij} and the jet-beam distance y_{iB} are defined as follows:

$$k_T \qquad y_{ij} = \frac{\Delta R_{ij}}{R} \min\{p_{T,i}, p_{T,j}\} \qquad y_{iB} = p_{T,i} \tag{9}$$

$$\text{C/A} \qquad y_{ij} = \frac{\Delta R_{ij}}{R} \qquad y_{iB} = 1 \tag{10}$$

$$\text{anti-}k_T \qquad y_{ij} = \frac{\Delta R_{ij}}{R} \min\{{p_{T,i}}^{-1}, {p_{T,j}}^{-1}\} \qquad y_{iB} = {p_{T,i}}^{-1}. \tag{11}$$

The jet algorithm uses those distance measures and does the following:

(1) **Find the minimum** $y^{\min} = \min_{ij}\{y_{ij}, y_{iB}\}$ for all combinations of two subjets in the event.

(2a) If $y^{\min} = y_{ij}$, **merge the subjets** i and j and their momenta and keep only the new subjet. Go back to (1).

(2b) If $y^{\min} = y_{iB}$, remove subjet i and call it a **final state jet**, then go back to (1).

The algorithm ends when there are no subjets left. Merging in this context means simply adding the four-momenta, $p_j = p_{j_1} + p_{j_2}$. The mass of the subjet is then defined as $m_j^2 = p_j^2$. The jet radius R is usually chosen to be within 0.4 and 0.7.

The k_T-algorithm starts clustering with soft constituents, the C/A-algorithm is purely geometric and the anti-k_T-algorithm starts with the hard constituents.

While the jet algorithms using k_T and C/A have a physical interpretation in the intermediate steps, this is not the case for the anti-k_T-algorithm. The advantage there is that the jet shapes are circles, which makes it easier to calculate their area. This is needed to estimate the effect of underlying event. This will be discussed in Section 3.3.

Fat Jets: With the development of larger and more energetic colliders like the LHC, heavier particles such as the top quark or the Higgs boson were produced and had to be reconstructed. It is possible to do this with essentially the same reconstruction algorithms. The only difference is that the radius R is chosen larger than for standard jets.

The reason for this is the following:

$$\begin{aligned} m^2 = (p_1 + p_2)^2 = 2(p_1 p_2) &= 2(E_1 E_2 - |\vec{p_1}||\vec{p_2}| \cos\theta_{12}) \\ &\approx 2 p_{T_1} p_{T_2} (1 - \cos\theta_{12}) \\ &\approx p_{T_1} p_{T_2} \theta_{12}^2, \end{aligned}$$

and with $p_{T_1} \approx p_{T_2} \approx \frac{p_{T_{1+2}}}{2}$ and $\theta_{12} \approx \Delta R$ it follows that

$$m^2 \approx \frac{p_{T_{1+2}}^2}{4} \theta_{12}^2 \quad \Rightarrow \quad \Delta R \approx \frac{2m}{p_T}. \tag{12}$$

Because of the larger masses of Higgs bosons or top quarks, the geometric size of those events is increased, too.

The large geometric size (fat jets are usually with $R = 1.5$ or $R = 1.8$) introduces some new challenges. As we will see in Section 3.3, the effect of underlying event increases drastically with the size of the jet.

3.3 QCD Background Effects

Besides the actual hard event that one is interested in, high energy collisions produce a lot of additional radiation that also needs to be understood to avoid mis-tagging when reconstructing the hard particles. There are different sources of radiation effects:

- **Final state radiation** are dominantly soft and collinear jets that are radiated off the final state particles. It can be simulated with a parton shower [18]. Since we are interested in the four-momenta of the top and the Higgs at parton-level, we need to include the final state radiation in the reconstruction algorithm.

- **Initial state radiation** are dominantly soft and collinear jets. As their name suggests, they are radiated off initial state particles. The reason for the radiation is that the initial partons are colored. This results in additional jets in the final state.

- **Underlying event** is caused by the partons that did not participate in the hard process but still interact with other partons. It results in extra soft QCD activity in the detector. Understanding the effect of underlying event is very important for reconstruction based on jets. The QCD radiation from the underlying event can have a large effect on the mass of the jet. Its effect scales approximately with R^4

when R is the size of the jet [19]:

$$\langle \delta m_j^2 \rangle \approx \Lambda_{UE}\ p_{T,j}\ \left(\frac{R^4}{4} + \frac{R^8}{4608} + \mathcal{O}(R^{12}) \right). \tag{13}$$

Here, Λ_{UE} is the amount of transverse momentum of the underlying event radiation per unit rapidity, $\Lambda_{UE} \approx \mathcal{O}(10)$GeV.

- **Pile-up** results from several proton-proton collisions in one bunch crossing. Since there are bunches of protons in the beams that collide, many collisions can happen in a short period of time and it is not always clear which jet results from which collision. This effect can be quite strong, it depends on the energy and the luminosity of the collider.

The effects described above can be distinguished from jets with so called jet grooming methods. One of the most important jet grooming methods is filtering. It was first implemented in the BDRS tagger [20]. The basic idea is that the constituents of the Higgs tagged fat jet are recombined into smaller C/A subjets of size

$$R_{\text{filter}} = \min\{0.3, \frac{\Delta R_{b\bar{b}}}{2}\}. \tag{14}$$

Obviously, this reduces the effective area and thus the effect of underlying event (Eq. 13). Then, the n_{filter} hardest constituents are chosen for the Higgs reconstruction. In the case of the BDRS tagger [20], it was found that $n_{\text{filter}} = 3$ reconstructs the Higgs mass best.

A very similar jet grooming method is trimming. It recombines the constituents of a fat jet into many small subjets and excludes all soft subjets with a p_T criterion. Then, it recombines the hard subjets, which again reduces the effective area and thus the effect of underlying event.

Pruning uses a different approach. It vetoes the merging of two subjets during the construction of the fat jet if the following two conditions are met:

$$\frac{\min p_{T,j_i}}{p_{T,j}} < z_{\text{prune}} \quad \text{and} \quad \Delta R_{j_1 j_2} > R_{\text{prune}}, \tag{15}$$

where z_{prune} and R_{prune} are free parameters.

3.4 Mass Drop and Mass Jump

The mass drop and the Mass Jump algorithm both make use of the same idea: they search for a large difference between the sum of the masses of two subjets and the mass of the combination of the two subjets. Such a difference in mass occurs when a heavy particle decays into ligher particles. This is obviously the case for our processes $t \to Wb$, $W \to q\bar{q}$ and $W \to \ell\nu_\ell$.

Mass Drop: In the original BDRS-tagger [20], the mass drop algorithm is used to identify the $H \to b\bar{b}$ decay. The HEPTopTagger also uses the mass drop criterion to identify the three prong structure of $t \to bW \to bjj'$.

After the standard clustering algorithm ($k_T, C/A,$ anti-k_T) is applied, the mass drop algorithm undoes the clusterings and checks if a mass drop occurred:

- **Undo the last clustering** of a jet j into j_1 and j_2, such that $m_{j_1} > m_{j_2}$.
- If $m_{j_1} < \theta m_j$, where θ is a free parameter, we conclude that a **mass drop** occurred and keep j_1 and j_2 as subjets. Otherwise, we assume j_2 is background radiation and keep only j_1.
- Those steps are repeated for kept subjets until the mass drops below a **cutoff** μ, where μ is a free parameter, then j_i is added to the output jets.

In our case of the top decay, we find $\frac{m_W}{m_t} \approx 0.46$ and $\frac{m_q}{m_W} \approx 0$. Therefore, the parameter θ should be chosen large enough to incorporate the first decay. By default,θ is chosen as 0.8. μ is by default 30 GeV.

Mass Jump: The Mass Jump algorithm (MJ) [21] is applied during the jet clustering: it prevents the recombination of two jets if their combined jet mass would be significantly larger than the masses of the separated jets. Instead of unclustering already clustered jets like the mass drop, the Mass Jump algorithm uses the constituents of an event and clusters them into jets. The Mass Jump algorithm searches for changes in the mass during the jet clustering. The distance measure for the Mass Jump is the measure for the usual k_T-, C/A- or anti-k_T- algorithms squared:

$$\begin{aligned} d_{j_1 j_2} &= \frac{\Delta R^2_{j_1 j_2}}{R^2} \min\{p^{2n}_{j_1\perp}, p^{2n}_{j_2\perp}\} \\ d_{j_1 B} &= p^{2n}_{j_1\perp}, \end{aligned} \tag{16}$$

where $n = 1$ for the k_T-, $n = 0$ for the C/A- and $n = -1$ for the anti-k_T-algorithm.

The algorithm first labels all candidates "active" and then does the following:

(1) **Find the smallest** $d_{j_a j_b}$ among all active candidates. If it is given by $d_{j_a B}$, label j_a passive, then repeat (1).

(2) **Combine** j_a **and** j_b in the E-scheme ($p_{j_a} + p_{j_b} = p_{j_a+j_b}$).
If $m_{j_a+j_b} < \mu$, replace j_a and j_b by j_{a+b}. Go back to (1).
If $\theta m_{j_a+j_b} > \max\{m_{j_a}, m_{j_b}\}$, label j_a and j_b passive. Go back to (1).
Here, μ (scale threshold) and θ (mass-jump threshold) are again free parameters.

(3) Check if a **Mass Jump** is happening between an active and a passive jet candidate:
Find the closest passive jet candidate j_n to j_a with $d_{j_a j_n} < d_{j_n B}$, then
check if $m_{j_a+j_n} > \mu$ and if $\theta m_{j_a+j_n} > \max\{m_{j_a}, m_{j_n}\}$. If both are fulfilled, label j_a passive. Do the same for j_b. If either j_b or j_a turned passive, go to (1).

(4) **No Mass Jump** occurred, replace j_a and j_b with their combination and go back to (1).

The algorithm terminates if no more active jets are left. R acts as an upper bound to the jet radius, μ is an indirect lower bound. For $\theta = 0$ or $\mu = \infty$, the Mass Jump algorithm is the standard sequential clustering.

One important difference between the mass drop and the Mass Jump is the fact that the Mass Jump avoids the cascade mass drops that are used in the mass drop algorithm. Therefore, in our example of the top decay, it is not necessary to choose $\theta > \frac{m_W}{m_t}$, which allows for stricter cuts and might decrease the mis-tagging. This is also the main reason why we want to try implementing the Mass Jump into our analysis. The $t\bar{t}H$ decay is a very busy environment, and therefore it might be useful to apply stricter cuts.

3.5 Top and Higgs Tagging

Tagging tops and Higgs reliably in a busy environment is very important for the measurement of the Yukawa coupling and therefore for the search of new physics BSM. In the following, we will present the BDRS Higgs tagger [20] and discuss the HEPTopTagger [12] in detail.

BDRS Tagger: The BDRS Higgs tagger [20] was one of the first jet substructure algorithms. It uses a C/A jet algorithm and a mass drop criterion. The key component

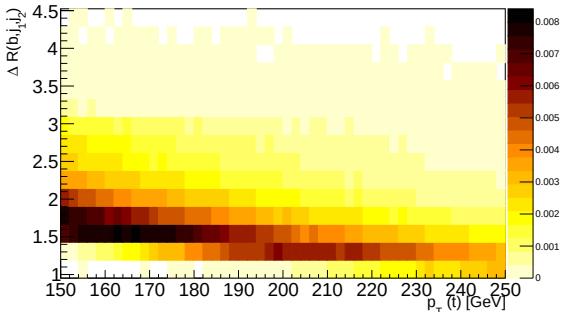


Figure 6: The distance between the decay products of the hadronic top is plotted versus the p_T of the top. With increasing p_T of the top, the distance between its decay products decreases.

of the BDRS tagger are three conditions:

$$\frac{\min m_{j_i}}{m_j} < 0.67, \qquad \frac{\min(p_{T,j_i}^2)\Delta R_{j_1 j_2}^2}{m_j^2} \sim \frac{\min p_{T,j_i}}{\max p_{T,j_i}} > 0.09, \qquad p_{T,j_i} > 30 \text{ GeV}. \tag{17}$$

The first condition makes sure that the drop in jet mass is large enough for a heavy particle decay, the second one assures the splitting is symmetric and finally it checks if the subjets are hard enough. Before the BDRS tagger reconstructs the Higgs mass from the relevant splittings, a filtering stage is applied to remove soft radiation, underlying event, and pile-up. The conditions above combined with the filtering make it possible to determine whether a jet comes from a typical soft or collinear QCD splitting or if it is from the Higgs. Since the Higgs (at that time not yet discovered) was expected to decay dominantly into bottom jets, the immediate subjets of the Higgs are expected to be bottom jets. Note that the BDRS tagger does not require the Higgs mass for the reconstruction.

HEPTopTagger: The HEPTopTagger [12] uses a generalised BDRS setup. It aims at moderate p_T ranges. It can reconstruct tops with $p_{T,\text{top}} \gtrsim 200$ GeV. The reason for this estimate is illustrated in Fig. 6 and Fig. 7. They show the distance between the decay products of the hadronic top $\Delta R(b, j_1, j_2)$ on parton-level. This distance is defined in the following way: We find the smallest distance between two of the three particles and combine those two particles by adding their four-momenta. The maximum distance is calculated as the distance between the combined particles to the third particle. One can see that the higher the transverse momentum of the top, the closer together are its decay products. With $p_{T,\text{top}} \gtrsim 200$ GeV, we find that the distance between the decay products is in most cases smaller than 1.8. If the decay products are close enough together, they

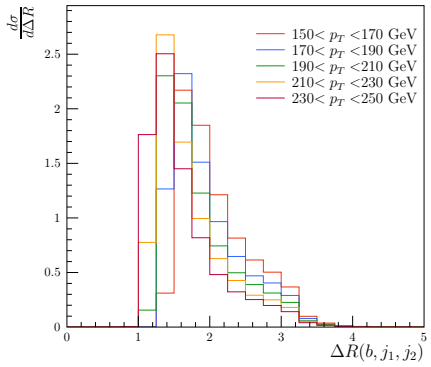


Figure 7: The same result as in Fig. 6 is obtained when looking at the ΔR for different p_T ranges of the top. We find that the peak of the distribution changes from $\Delta R = 1.7$ for $p_T < 170$ GeV to $\Delta R = 1.3$ for $p_T > 230$ GeV.

will be clustered inside one fat jet and only in this case the top can be reconstructed. The HEPTopTagger does the following steps:

- First, we **define a C/A fat jet** with $R_{\text{fat}} = 1.5$ or 1.8.
- Then, we **identify all hard subjets with a mass drop criterion**: we undo the last clustering of jet j into the two subjets j_1 and j_2 such that $m_{j_1} > m_{j_2}$. We require $m_{j_1} < f_{\text{drop}} m_j$, where $f_{\text{drop}} = 0.8$, to keep both subjets; otherwise assume j_2 is soft radiation and keep only j_1. We continue decomposing subjets until there are no jets left with $m_j > m_{\text{min}} = 30$ GeV.
- Next, we **find possible top candidates** by iterating through all triplets of three hard subjets. We filter them and use the 5 hardest constituents to calculate their combined jet mass, allowing for two gluons from final state radiation of the decay products. Then, we re-cluster these subjets into three assumed top decay jets and reject triplets outside $m_{123} = m_{\text{rec}} \in [150, 200]$ GeV. This reduces the effective jet area and thus the influence of underlying event.
- Finally, we **accept a top candidate** if the masses of the p_T-ordered subjets

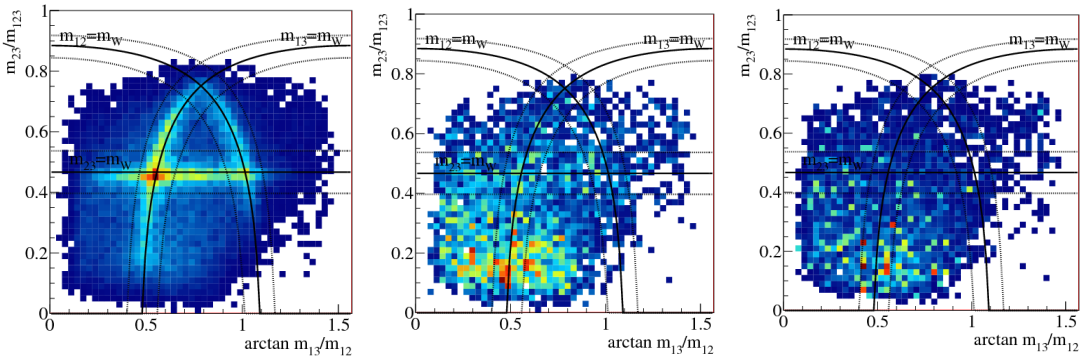


Figure 8: Events for the $t\bar{t}$ (left), Wjj (center) and pure QCD jets (right) plotted in the $\arctan(m_{13}/m_{12})$ vs (m_{23}/m_{123}) plane. Areas with a higher population density appear red. Figures taken from [3].

j_1, j_2, j_3 satisfy one of the following criteria:

$$
\begin{aligned}
&0.2 < \arctan\frac{m_{13}}{m_{12}} < 1.3 \quad \text{and} \quad R_{\text{min}} < \frac{m_{23}}{m_{123}} < R_{\text{max}}, \\
&R_{\text{min}}^2\left(1+\left(\frac{m_{13}}{m_{12}}\right)^2\right) < 1-\left(\frac{m_{23}}{m_{123}}\right)^2 < R_{\text{max}}^2\left(1+\left(\frac{m_{13}}{m_{12}}\right)^2\right) \text{and } \frac{m_{23}}{m_{123}} > 0.35, \\
&R_{\text{min}}^2\left(1+\left(\frac{m_{12}}{m_{13}}\right)^2\right) < 1-\left(\frac{m_{23}}{m_{123}}\right)^2 < R_{\text{max}}^2\left(1+\left(\frac{m_{12}}{m_{13}}\right)^2\right) \text{and } \frac{m_{23}}{m_{123}} > 0.35.
\end{aligned}
\tag{18}
$$

Here, $R_{\text{min,max}} = (1 \mp f_W)\frac{m_W}{m_t}$ is defined by the parameter f_W, which is by default 0.15.

- From all triplets that passed until this point, we **accept the top candidate** with its mass $m_{123} = m_{\text{rec}}$ closest to m_t.
- Finally, we **do a consistency check** by requiring $p_{T,t} > 200$ GeV.

The requirements in Eq. 18 are chosen because of the following: We know that for three jets j_1, j_2 and j_3 with respective masses m_i and combined masses m_{ij}:

$$
m_{\text{top}}^2 = m_{123}^2 = (p_1+p_2+p_3)^2 = m_{12}^2 + m_{23}^2 + m_{13}^2. \tag{19}
$$

$$
\Rightarrow 1 = \frac{m_{12}^2}{m_{123}^2} + \frac{m_{23}^2}{m_{123}^2} + \frac{m_{13}^2}{m_{123}^2}, \tag{20}
$$

which is the equation for a surface of a sphere. This is why we consider $\arctan(\frac{m_{13}^2}{m_{12}^2})$ and $\frac{m_{13}^2}{m_{12}^2}$; they represent mass planes. Therefore, we can conclude that if a top existed in the decay, we will find a mass drop where $\frac{m_{ij}^2}{m_{123}} \approx \frac{m_W^2}{m_t}$. Thus, we consider $|\frac{m_{ij}}{m_{123}}/\frac{m_W}{m_t}|$.

This leads to A-shaped bands, which can be seen in Fig. 8. We find indeed, that the $t\bar{t}$ events follow the expected A-shaped bands, whereas the figures for the Wjj and pure QCD jets show no specific shape.

The HEPTopTagger is based only on jet masses and does not use p_T-cuts. It uses filtered subjets, which improves the top tagging performance. Note that the HEPTopTagger also reconstructs the four-momenta of the tops, which is crucial when one wants to use the HEPTopTagger in the search for new physics.

3.6 Event Generation

To compare the different analyses, we use Monte Carlo simulation programs. As the name suggests, Monte Carlo generators use random sampling to calculate results. There are two problems when it comes to calculating the outcome of a collider experiment. First, one usually has to solve an integral with many integration variables, which might not be possible with ordinary numerical integration methods. The second point is that colliders usually have a non-trivial geometrical area where they can measure particles, which also needs to be taken into account. Monte Carlo methods sample the integrand N times and then take the average:

$$\bar{f} = \int_0^1 dx_1 \cdots \int_0^1 dx_n f(x_1, \cdots, x_n) \approx \frac{1}{N} \sum_{i=1}^{N} f(x_1(i), \cdots, x_n(i)). \tag{21}$$

In the case of a hadron-hadron collision, one needs to evaluate the following integral:

$$\sigma(AB \to cX) = \sum_{a,b} \int dx_a dx_b \left[f_{a/A}(x_a) f_{b/B}(x_b) + (A \leftrightarrow B \text{ if } a \neq b) \right] \sigma(ab \to cX), \tag{22}$$

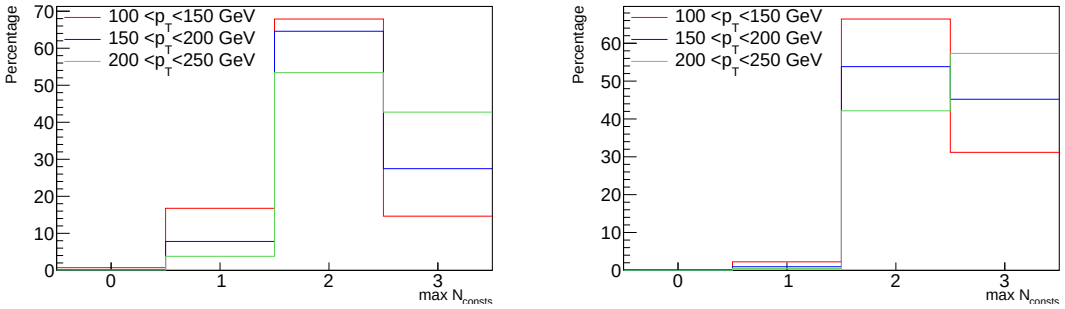
where A and B are hadrons with a and b being their respective partons. x_a is the longitudinal momentum fraction of a and b respectively, $f_{a/A}$ is the parton density of a in A, the so-called parton distribution function.

MadGraph5: MadGraph5 [22] was used to generate the hard processes needed in our analysis. It finds all processes with the initial and final states one inputs into the program. It then calculates the cross section at tree-level and next to leading order (NLO) for many processes and outputs the events it generated. One can use MadGraph5 for event generation within the Standard Model as well as for theories beyond the Standard Model.

Pythia8: We then shower the events generated with MadGraph5 using Pythia8 [23]. Pythia8 can simulate initial and final state radiation, i.e. the parton shower, hadronization and the decay of hadrons. The output file of Pythia8 contains the same information as one would expect from an actual collider experiment with perfect detectors.

Rivet: To analyse the output of our simulated process, we use Rivet (Robust Independent Validation of Experiment and Theory) [24]. Rivet provides many analyses for Monte Carlo simulations and one can build one's own analyses in Rivet with C++. It is possible to compare the reconstructed particles with parton-level particles in order to see how efficient and trustworthy the analysis is in the reconstruction of particles.

FastJet: Furthermore, we use FastJet [25] to build and analyse jets. FastJet has many different sequential recombination algorithms and features built in to cluster jets. The tools are also helpful for jet substructure analysis. Furthermore, it contains tools to calculate the area of a jet to estimate background effects such as underlying event and pile-up.



(a) $R = 1.8$ (b) $R = 2.5$

Figure 9: Number of top decay products captured inside a fat jet for different p_T ranges and different sizes of the fat jets.

4 Analysis

The main goal of this work is to check whether we can extend the accessible phase space for the $t\bar{t}H$ process towards lower boosts using the HEPTopTagger and the Mass Jump algorithm. Therefore, we include cuts on the parton-level p_T of the hadronically decaying top that require 150 GeV $< p_{T,\text{top}} <$ 250 GeV in all our analyses. Obviously, this can never be applied in an experiment. But since we are only interested in the question if our idea works in principle, the parton-level cut is justified. We assume a collision energy of $\sqrt{s} = 13$ TeV for the generation of the events.

The problem with the $t\bar{t}H$ process is its high combinatoric. The most challenging part is the combinatorial background from the signal itself, since we don't know which of the b-jets originate from which particle. Our final state consists of (at least) four b-jets, and it is not possible to determine which ones originate from the tops and which from the Higgs. Furthermore, there are many systematic uncertainties on the signal and background predictions that limit the sensitivity.

4.1 Kinematics of the $t\bar{t}H$ Process

The general idea of the standard tagging algorithm is the following: we use the lepton from the leptonically decaying top to trigger the event. Then, we find a fat jet that is tagged with the HEPTopTagger, and a fat jet which includes two b-jets for the Higgs reconstruction. We want to find parameters that work best when analysing the decay. This is, for example, the cone size R of the fat jet, which is related to the p_T of the top. It should be large enough to capture all top decay products. To determine the best

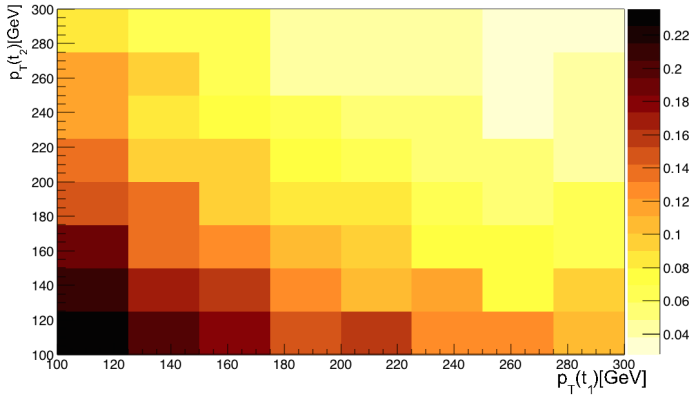


Figure 10: Overlap probability of the tops. To determine if the tops overlap, we define triangles for the decay products of each top. If a decay product of one top is inside the area of the triangle from the other top, it is considered an overlap.

possible parameters, it is useful to look at the $t\bar{t}H$ decay kinematics at parton-level.

Fig. 9 shows the number of decay products N_{consts} that are captured by the fat jet for different jet cone sizes ($R = 1.8$ and $R = 2.5$) and at different transverse momenta of the hadronic top. We want to capture three hard constituents, because we need all decay products to reconstruct the top with the HEPTopTagger.

As one would expect, Fig. 9 shows that with larger fat jets it is likelier that all three decay products of the top are captured inside the fat jet. The probability to capture all decay products increases with the transverse momentum of the top. This is expected since a higher p_T is related to a higher energy of the top, and the decay products of a boosted top are collimated.

Another interesting observation when studying the parton-level kinematics is the fact that the total number of events increases at lower p_T. Even though the possibility to capture all three decay products for 100 GeV $< p_T <$ 150 GeV is very small, we would have a lot more data from the LHC if it was possible to tag the top reliably in this p_T range.

Generally, we can see that using larger fat jets increases the probability to find all three decay products of the top. But using larger areas for the reconstruction of the tops has negative effects: we will have to deal with more QCD background. Moreover, we will be in trouble if the areas of the decay products of the tops or the Higgs overlap and we capture decay products of the two tops in one fat jet.

To estimate how often this overlap will happen and how it is related to the p_T of

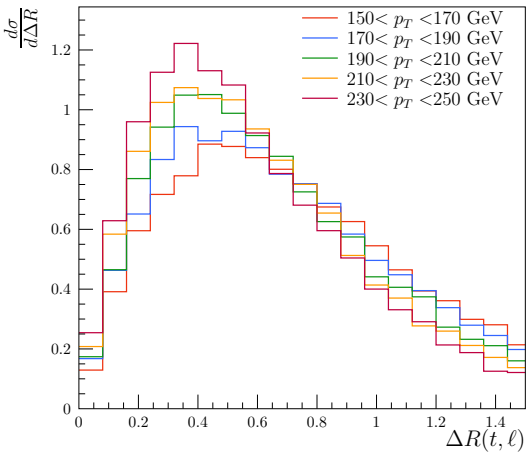


Figure 11: Distance between the leptonically decaying top and the lepton.

the top, we defined triangles for each top in the $\eta - \phi -$ plane containing the decay products of each top. We then find for each triplet of top decay products respectively the triangle with the smallest circumference. There is not just one possibility because of the periodicity in ϕ. With those triangles we can then check if one of the decay products of one top is inside the area of the other top or if the areas overlap somewhere. If one of the above is true, we then say that the tops overlap.

The probability for this to happen can be seen in Fig. 10. Obviously, this is only an estimation since in the actual analysis, we cluster jets and not triangles and we also have to deal with the Higgs, but we think the result is still useful for the determination of the top $- p_T$ ranges we want to consider.

We can see in Fig. 10 that the larger the p_T of the tops, the lower the possibility for the tops to overlap. This is what we expect and we find that for $p_T > 150$ GeV of both tops, the probability for the overlap is below 15%, a reasonable compromise.

It is also interesting to look at the distances between the different particles at parton-level. This helps us to determine the parameters we choose for the analyses.

In Fig. 6 and Fig. 7 one can see the maximum distance between the three decay products of the hadronically decaying top.

One can see that the more energetic the top is, i.e. the higher the transverse momentum of the top is, the smaller is the maximum distance between its decay particles. The plots Fig. 6 and Fig. 7 also confirm that with a jet radius of $R = 1.8$ we can cluster all decay products of the hadronically decaying top in one jet in many cases.

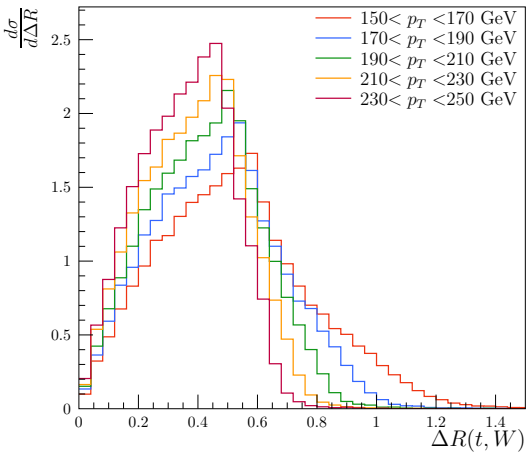


Figure 12: Distance between the hadronically decaying top and the W boson. We can see that the higher the transverse momentum, the smaller the distance between the top and the W.

Furthermore, we can look at the distances between the tops and other particles. This is especially interesting when we want to divide the event into four areas, depending on the position of the b-jets, as we will do in Section 4.2.2. We will for example identify the leptonic b as the b closest to the lepton. The question we should ask first is whether we can assume that the decay products of the different particles are close enough to tag them reliably inside an area. This result can be seen in Fig. 11 and Fig. 12.

We find that the lepton and the top it originates from are, in most cases, relatively close. The higher the transverse momentum of the top, the smaller is the distance. The peak changes from R = 0.5 to R = 0.3 when the p_T of the top changes from 150 GeV to 250 GeV. This is promising since we hope to find the particles related to each other close together, and we can indeed confirm that the lepton is, in most cases, close to the leptonically decayed top.

In Fig. 12, we find that for most events the distance between the top and the W is smaller than 1, which is very promising for the standard fat jet analysis and the area analysis. As before, the distance decreases with increasing transverse momentum of the top and we are likely to find the hadronically decayed top close to the W.

As discussed previously, η is an important parameter within the detector. However, our detector can only detect particles that are within $|\eta| < 2.5$. Therefore, we want to check if the decay products of the top are not too far spread, as that would be a problem. For each event we found the largest η from the b quark and the two jets of the hadronically decaying top. The result can be seen in Fig. 13. We conclude that we can find most final state particles within $|\eta| < 2.5$.

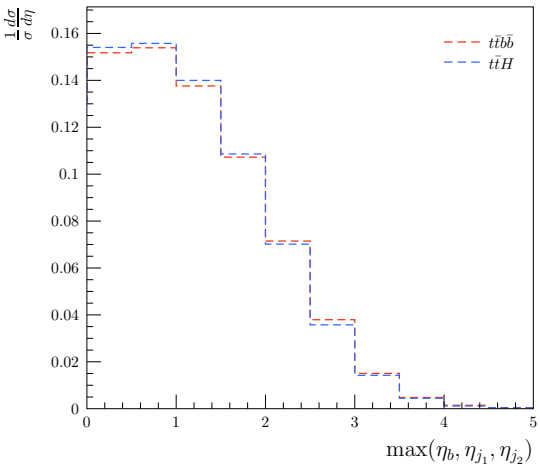


Figure 13: This plot shows the maximum η of the decay products of the hadronically decaying top. Most of the time the decay products are within $|\eta| < 2.5$.

4.2 Description of the Algorithms

In this section, we will introduce the different algorithms we used to tag the top and the Higgs. The fat jet analysis is currently the standard approach for $t\bar{t}H$ processes, the area analysis and the pure Mass Jump analysis are alternative approaches. The main advantage of those alternative approaches it that they are not using fat jets and are thus not limited by the relatively small area of a fat jet.

4.2.1 Fat Jet Algorithm

For the fat jet algorithm, we assume that the decay products of the hadronically decaying top can be clustered together in a fat jet and that the decay products of the Higgs are also inside one fat jet. Therefore we search with the HEPTopTagger for the b-jet and two quarks in the fat jets and with a Higgs tagger for two b-jets. The leptonically decaying top is not reconstructed but instead is used for the triggering of the event. This reconstruction algorithm is similar to the one in Ref. [14].

The tagging algorithm for the fat jet analysis then does the following:

- We require at least two C/A fat jets with $R = 1.8$ and $|\eta| < 2.5$ and at least one lepton from the leptonically decayed top with $|\eta| < 2.5$ and $p_{T,\ell} > 20$ GeV to **trigger the event**. The tagging efficiency of the lepton is assumed to be 100%. If one of the above is not found, we veto the event.

- Then we let the **HEPTopTagger** run over both fat jets. If we find at least one top, we continue. If no top is found, we veto the event. If two top candidates were found, we accept the one with its mass closer to the top mass.

- We use a **Higgs tagger** to find the Higgs in one of the fat jets. Again, if no Higgs was found, we veto the event and if two Higgs candidates were found we accept the one with its mass closer to the Higgs mass.

- Finally, we require that the two **Higgs subjets are b-tagged**, if this is not true then we veto the event. We assume a b-tagging efficiency of 70% and a mis-tagging probability of 1%.

For the Higgs tagger we used the same tagging algorithm as in Ref. [14]:

- We **undo the last clustering** of the fat jet, which gives us two subjets j_1 and j_2 with $m_{j_1} > m_{j_2}$.

- Next, we **use a mass drop criterion**. If $m_{j_1} > 0.9 m_j$, we assume that j_2 either comes from underlying event or from soft QCD emission, which is why we only keep j_1 and discard j_2. If the above is not true, we keep both j_1 and j_2.

- We **further decompose** all jets recursively until the mass of the new subjet is below the threshold, $m_{j_i} < 40$ GeV. Then we add this subjet to the list of relevant substructures.

- Once all relevant substructures are found, we **order all possible pairs** by the modified JADE distance (Eq. 8).

- We **filter** the leading pair and keep it for the Higgs reconstruction.

Note that the Higgs tagger was developed before the Higgs was discovered in 2012 and does therefore not use the Higgs mass as a criterion. This was done because we want to be able to compare our results to the ones found in [14], even though it would improve our tagger if we include a cut on the Higgs mass.

4.2.2 Area Analysis

For the area analysis, we use a different approach. Instead of plugging fat jets into the HEPTopTagger, we will use jets that are build with Mass Jump. To avoid the limitations of the fat jets, the hadronic activity of the event was divided into four areas, depending on which b quark they were closest to. However, this will increase the effect of

underlying event. The Mass Jump algorithm improves the tagging of heavy particles in dense environments. We assume that we can tag the top with the Mass Jump jets from one of the areas and then reconstruct the Higgs from the remaining b-jets.

The area algorithm then does the following:

- We again **trigger the event** requiring at least one lepton with $|\eta| < 2.5$ and $p_{T,\ell} > 20$ GeV.
- Then, we **find all b-tagged jets** assuming a tagging efficiency of 70% and a mis-tagging probability of 1%. We cluster the b hadrons in anti-k_T jets with $R = 0.4$ and $p_T > 20$ GeV. We require at least four b-jets, if there are less we veto the event.
- Then, we select the four hardest b-jets and use this information to **divide the event into four areas**. This is done by finding the smallest ΔR (Eq. 6) distance between each particle and the four b-jets.
- We use the **Mass Jump algorithm** (Section 3.4) to cluster the hadronic activity in each area separately into jets, excluding the area where the lepton was found.
- Then, we run the **HEPTopTagger** over the three sets of Mass Jump jets separately to find at least one top. If no top is detected, we veto the event.
- Since one of the b-jets was used for the top reconstruction, and the one from the leptonic top was already excluded, there are two b-jets left. They are used for the **reconstruction of the Higgs**.

We again have no cut on the Higgs mass since we want to compare the results of the area analysis with the fat jet analysis.

4.2.3 Pure Mass Jump Analysis

A problem with the area analysis is that it assumes that all decay products of the top are closest to the b quark from the top, but another b quark, probably from the Higgs, might be in between. In this case it is not possible to reconstruct the top. Therefore, we try a different approach. We Mass Jump the whole event and use it as input to the HEPTopTagger.

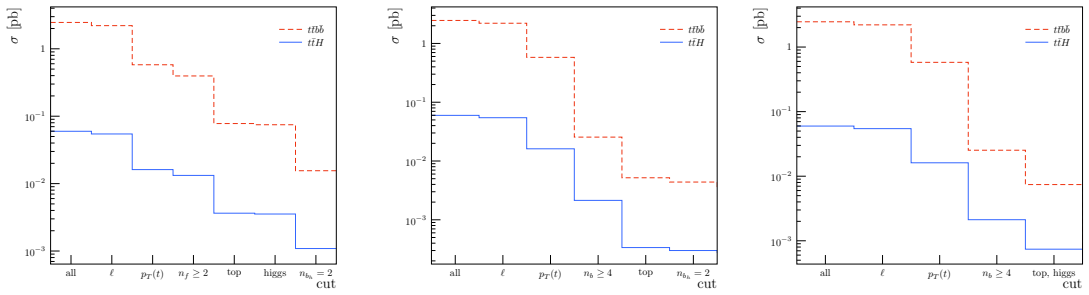
The basic idea is that we can use the Mass Jump jets to find several top and Higgs candidates and then use the top candidate that fulfils Eq. 12 the best.

- First, we **cluster the event** in anti$-k_T$ jets with $R = 0.4$ and $p_T > 20$ GeV.

- Then, we find at least **4 b-tagged jets** with a tagging efficiency of 70% and a mis-tagging probability of 1%.
- We select the four hardest b-jets and **remove the b-tagged jet** from the leptonically decaying top with $\Delta R(b, t_\ell) < 0.3$ and we remove the **lepton**.
- Then, we cluster the rest of the event with **Mass Jump** ($\mu = 30, \theta = 0.7, r_{\max} = 1.5$).
- We **search for the top and the Higgs** in the Mass Jump jets. We run the HEPTopTagger over one b-jet and all non-b-tagged jets and iterate over all b-jets. Then, we choose the candidate for which $\Delta R - \frac{2m_t}{p_T}$ is minimal because of Eq. 12. We require that in the unused Mass Jump jets two b-tagged jets are left, which then are used for the reconstruction of the Higgs.

For the same reasons as above, we have no cut on the Higgs mass.

Note that we use parton-level information when we remove the leptonic part from the event. Therefore, the pure Mass Jump analysis is rather the study of an idea than an actual analysis. In the case where one wants to use this analysis for actual data, one would have to come up with a solution of how to exclude the leptonic part from the event. For example by choosing the b closest to the lepton, which was found to be the correct in about 50% of the events.



(a) fat jet analysis (b) area analysis (c) pure MJ analysis

Figure 14: The cutflow shows how many events are kept after a certain requirement in the algorithm. It is normalised to the cross section.

5 Results

In this section, we will present the results we found. From [14], we know that the most important background for the semileptonic $t\bar{t}H$ decay is $t\bar{t}b\bar{b}$. We will only consider this background here and neglect all other possible backgrounds.

First, we will discuss how well the different analyses exclude background events and how they perform on $t\bar{t}H$ events. Then, we will discuss the quality of the top and the Higgs reconstruction with the help of the mass and the transverse momenta distributions and by comparison to the parton-level particle on an event to event basis.

5.1 Cutflow

The cutflow indicates how many events are rejected by each algorithm. It simply shows the number of events that pass a criteria in the analysis. Obviously, we would like the $t\bar{t}H$ events to pass the analysis an the $t\bar{t}b\bar{b}$ events to be rejected, since we don't know which events have a Higgs included when we look at actual measurement from the LHC. The results can be seen in Fig. 14.

Note that the cutflow plots are normalised to the cross section of the process, which is larger for $t\bar{t}b\bar{b}$ than for $t\bar{t}H$. This is why they don't start at the same number even though event and background were generated with one million events.

The cuts in the fat jet analysis cutflow represent the following: First, we have all particles, then we require one lepton. The third requirement demands the transverse momentum of the parton-level hadronic top to be within 150 and 250 GeV, which is a cut that is using parton-level information. Next, we need at least two fat jets. We

	signal efficiency [%]	background efficiency [%]
fat jet analysis	27.56	19.61
area analysis	15.67	20.36
pure MJ analysis	39.51	38.09

Table 1: Top tagging efficiencies for the different analyses in percent.

want to find at least one top and one Higgs. Finally, we require two b-jets to be in the reconstructed Higgs in order to accept the reconstruction of the whole event.

For the area analysis, we start with the same requirements: first we plot all events, then we cut on the lepton and the $p_{T,\text{top}}$ between 150 and 250 GeV, again using MC truth. Then we require, in difference to the fat jet analysis, 4 b-tagged jets. Next we want to find at least one top and require that two b-tagged jets reconstruct the Higgs.

The pure MJ analysis cuts on the same requirements as before, which are the lepton, the p_T of the hadronic top and the four b-tagged jets that should be found in the event. Since we then reconstruct the top and the Higgs in one final step, this is the last cut of the pure MJ analysis.

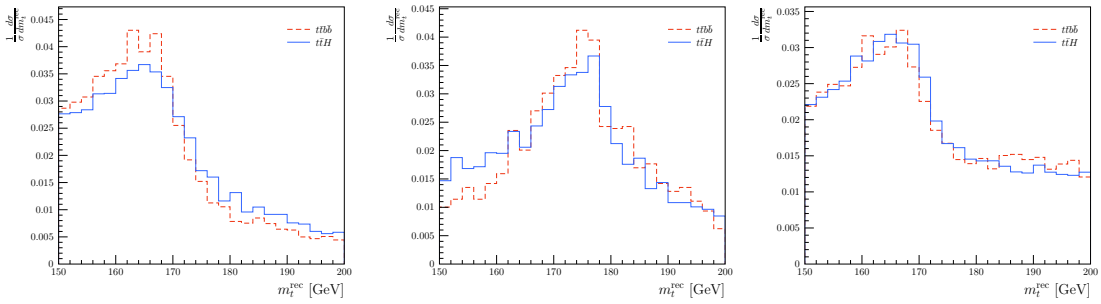
We find that for all analyses, the rejection at each cut seems to be about the same for event and background. The reason for this is that the structure of the $t\bar{t}H$ and the $t\bar{t}b\bar{b}$ process are very similar, which makes it hard to determine whether a Higgs was present in the event or not. This problem would be less important if we include a cut on the Higgs mass.

The cutflow plots also show that the area analysis rejects more events than the fat jet analysis and the pure MJ analysis – which both end up with about the same amount of events at the end of the analysis.

We are also interested in the tagging efficiency of the top, which shows the percentage of the top we were able to tag with the HEPTopTagger. In Tab. 1, we show the tagging efficiencies for the $t\bar{t}H$ signal and the $t\bar{t}b\bar{b}$ background. The pure MJ analysis has by far the highest tagging efficiency. For the total reconstruction efficiency, we find similar results.

	total efficiency [%]
fat jet analysis	1.82
area analysis	0.45
pure MJ analysis	3.90

Table 2: Total reconstruction efficiencies for the different analyses in percent.



(a) fat jet analysis (b) area analysis (c) pure MJ analysis

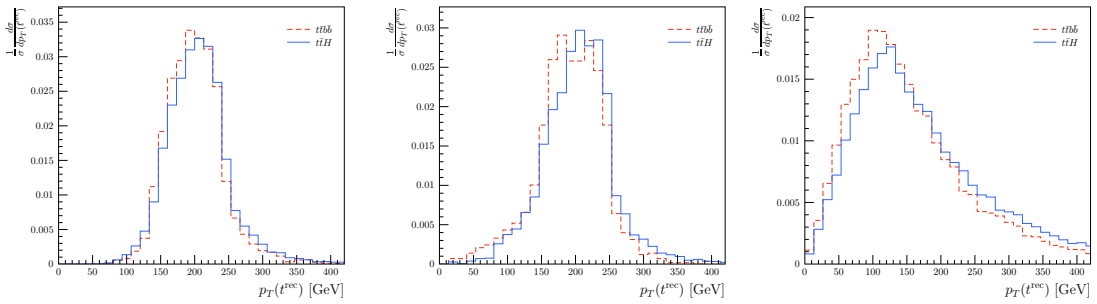
Figure 15: The reconstructed top mass distribution for the different analyses for $t\bar{t}H$ and $t\bar{t}b\bar{b}$.

5.2 Top Reconstruction

For the reconstruction of the top quark, we always used the HEPTopTagger. But the quality of the top reconstruction depends on the jets used as input for the HEPTopTagger, as one can see in Fig. 15 and Fig. 16. The plots look the same for signal and background, which is expected since both $t\bar{t}H$ and $t\bar{t}b\bar{b}$ contain a top.

We expect a mass peak at $m_t = 173$ GeV. However, we find that only the area analysis is able to reconstruct the top mass correctly, both the fat jet analysis and the pure MJ analysis have a peak at 166 GeV. This shows that in both analyses, some of the particles that originate from the top are missing in the reconstruction. In the area analysis, we have a better chance of finding the correct particles compared to the fat jet analysis, where it is more likely that some decay products of the top are not inside the fat jet and therefore cannot be reconstructed. The pure MJ analysis often seems to identify wrong triplets as top. For the p_T plots, we find that the fat jet analysis and the area analysis both peak at around $p_{T,\text{top}} \approx 200$ GeV, which is what we expect. Keep in mind that we set a parton-level cut on the p_T of the top, 150 GeV $< p_{T,\text{top}} <$ 250 GeV. However, we find that for the pure MJ analysis, the p_T distribution looks very different, with a peak below 150 GeV and many events outside the intended p_T range. This again indicates that with the pure MJ analysis, the HEPTopTagger frequently identifies wrong triplets.

Note that the area analysis has the lowest tagging efficiency, but correctly reproduces the known kinematic features. This suggests that when the top is found within one area, it is likely that all the final state particles that originate from the top are correctly reassigned and used for the top reconstruction. This is in contrast to the pure MJ analysis, where we found a high tagging efficiency but it seems that, in many cases, constituents were used for the reconstruction of the top that did not originate from the



(a) fatjet analysis (b) area analysis (c) pure MJ analysis

Figure 16: The reconstruction of the transverse momentum of the top is shown for the different analyses for $t\bar{t}H$ and $t\bar{t}b\bar{b}$.

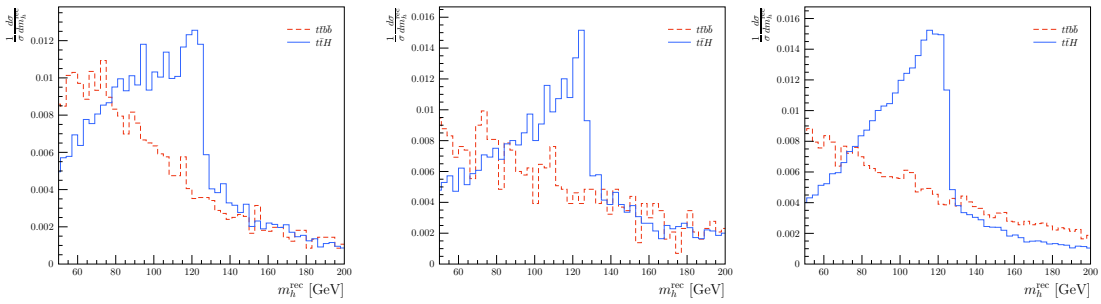
top. The reason for this is the high combinatorics of the event.

5.3 Higgs Reconstruction

In all three analyses, the Higgs boson is reconstructed from the b-jets that were not used for the top reconstruction. We always reconstruct the top first and use the remaining event to reconstruct the Higgs from two unused b-tagged jets. The results can be seen in Fig. 17 and Fig. 18.

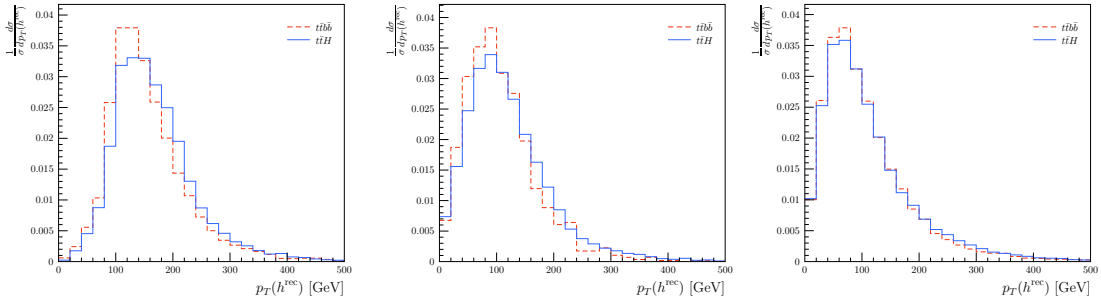
For all analyses, we can clearly see the Higgs peak in the $t\bar{t}H$ process and no peak for the background process $t\bar{t}b\bar{b}$ in Fig. 17. Furthermore, we notice that all peaks are at m_{Higgs} = 125 GeV, as we would expect. But looking at the different mass distributions it also becomes obvious that the area analysis and the pure MJ analysis work best for the reconstruction of the Higgs: the peak is narrow and high at m_{Higgs} = 125 GeV. The peak of the pure MJ analysis looks smoother than the area analysis Higgs peak, which is due to the fact that more events passed the pure MJ analysis. The Higgs peak of the fat jet analysis is broader and has many events at masses below the Higgs mass, and a sharp cut-off for masses above 125 GeV. All peaks have a sharp cut-off for masses above 125 GeV.

When looking at the p_T distributions in Fig. 18, we find that the area analysis and the pure MJ analysis peak at around 100 GeV, whereas the peak of the fat jet analysis is shifted to higher p_T and peaks at 140 GeV. Here, few reconstructed Higgs have p_T smaller than 100 GeV. But for both the area analysis and the pure MJ analysis, there are many reconstructed Higgs with very low p_T. This leads to the conclusion that the fat jet analysis works best for the reconstruction of the Higgs p_T. Another interesting



(a) fat jet analysis (b) area analysis (c) pure MJ analysis

Figure 17: The Higgs mass distribution for the different analyses for $t\bar{t}H$ and $t\bar{t}b\bar{b}$.



(a) fat jet analysis (b) area analysis (c) pure MJ analysis

Figure 18: The transverse momentum of the reconstructed Higgs for $t\bar{t}H$ and $t\bar{t}b\bar{b}$.

observation is the fact that the distributions for the Higgs reconstruction from the $t\bar{t}b\bar{b}$ event are shifted towards smaller p_T in the fat jet analysis and in the area analysis. However, in the pure MJ analysis, the distributions differ only slightly. The reason for this is again that the pure MJ analysis often falsely identifies particles as decay products of the top or the Higgs due to the combinatorics.

5.4 Comparison to Parton-Level

In Fig. 19 we plotted the difference between the reconstructed and the parton-level top. This indicates how accurate the reconstruction which each analysis is. The left figure in Fig. 19 shows that the fat jet analysis is the most accurate when it comes to the reconstruction of the position of the top. The area analysis is slightly worse and the pure MJ analysis is the least accurate in reconstructing the top from the tested algorithms. But none of the comparison distributions are very far off, they all seem to reconstruct

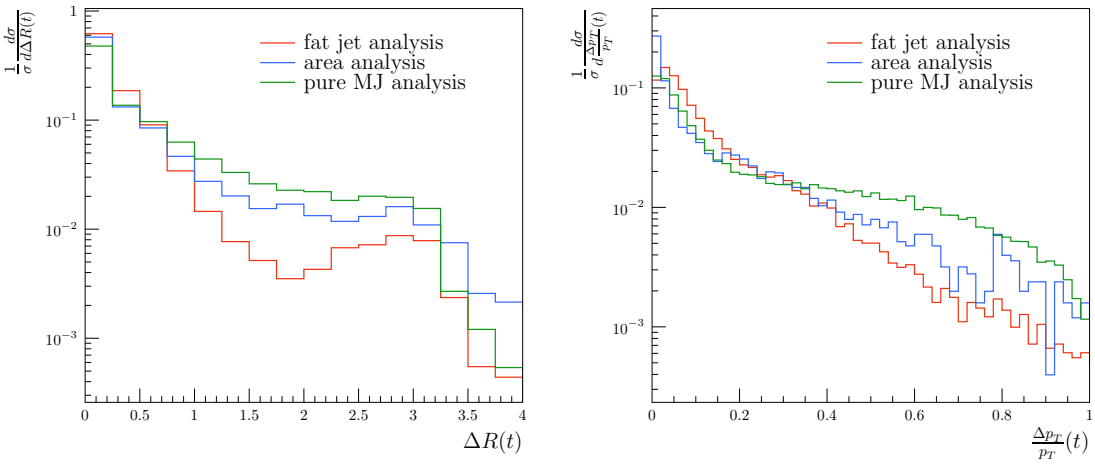


Figure 19: ΔR (on the left) and the percentage difference of the transverse momentum (on the right) between the reconstructed and the parton-level top.

the position of the top accurately.

When it comes to the reconstruction of the transverse momentum of the top, we find a different result: we calculated the percentage difference between the parton-level and the reconstructed transverse momentum, $\frac{|p_{T,\text{true}}-p_{T,\text{rec}}|}{p_{T,\text{true}}}$, and plotted the result in Fig. 19 on the right. Here, the area analysis has clearly the best result. The fat jet analysis and the pure MJ analysis are both worse at reconstructing the transverse momentum of the top correctly. Keep in mind that those two analyses have relatively high top tagging efficiencies. This leads to the conclusion that the fat jet analysis and the pure MJ analysis are more likely to identify a wrong triplet due to combinatorics.

This leaves us with the same conclusion as before: the area analysis is overall as good at reconstructing the top as the fat jet analysis, while being slightly better at reconstructing the position of the top. The pure MJ analysis is worse but still accurate at reconstructing the top.

For the reconstructed Higgs, we did the same analysis. The results can be seen in Fig. 20.

For the distance between the reconstructed and the parton-level Higgs in the $\eta-\phi-$ plane we find that the pure MJ analysis works best. The area and the fat jet analysis have about the same quality of reconstruction. The pure MJ analysis reconstructs the Higgs better at the cost of the top reconstruction.

For the difference in the transverse momentum, we find that none of the analyses

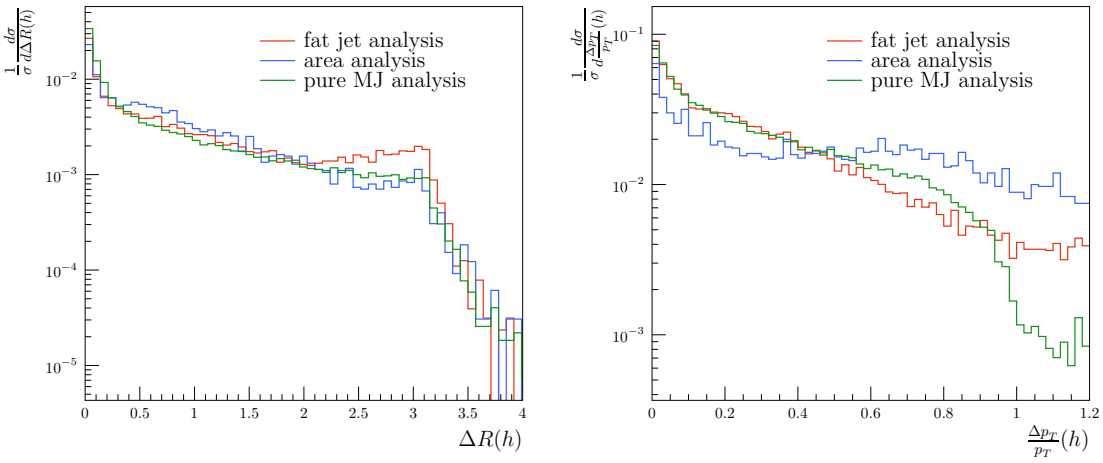


Figure 20: ΔR (on the left) and the percentage difference of the transverse momentum (on the right) between the reconstructed and the parton-level Higgs.

reconstruct the transverse momentum of the Higgs very well. However, the fat jet analysis and the pure MJ analysis reconstruct the transverse momentum of the Higgs better than the area analysis.

Even though the area analysis worked well for the properties of the top, it does not give satisfying results for the Higgs reconstruction. This could be improved with a cut on the Higgs mass. The pure MJ analysis reconstructs the Higgs better than the fat jet analysis, but has problems with the reconstruction of the top, which could be improved with stricter cuts on the Mass Jump jets.

6 Conclusion

Our goal was to check if it is possible to extend the HEPTopTagger for semileptonic $t\bar{t}H$ to lower p_T ranges by using the Mass Jump algorithm. We did this by comparing the reconstruction of the hadronic top and the Higgs in the standard fat jet analysis with two other approaches. In the fat jet analysis, we cluster our event into two fat jets, which we use as input for the HEPTopTagger. Then, we reconstruct the Higgs from the remaining subjets with a Higgs tagger. One of the alternative approaches is the area analysis. Here, we divided the whole event into four areas: each particle is assigned to the b-jet it is closest to. We exclude the area that contains the lepton and Mass Jump the remaining areas. Then, the HEPTopTagger runs over each area separately to find a top. The remaining two b-jets are then used for the Higgs reconstruction. In the other alternative approach, the pure MJ analysis, we exclude the leptonic part, e.g. the lepton and the b-jet from the leptonic top based on parton-level information. Then, we cluster the whole event with Mass Jump and find the best candidate for the top by minimising $\Delta R - \frac{2m}{p_T}$ (Eq. 12). The Higgs is reconstructed from the two unused b-jets.

For the top reconstruction, we found that the area analysis was the only analysis to correctly reconstruct the mass. It reliably reconstructed the transverse momentum and the position. The results from the fat jet analysis were slightly worse. The quality of the top reconstruction from the pure MJ analysis was significantly worse. For the reconstruction of the Higgs, we found that all mass peaks were at the Higgs mass, but the peak was narrower for the area analysis and for the pure MJ analysis than for the fat jet analysis. However, the area analysis did not reconstruct the position of the Higgs reliably. The pure MJ analysis worked well for the reconstruction of the position of the Higgs boson. Reconstructing the transverse momentum of the Higgs is a challenge for all analyses.

In conclusion, we can say that the pure MJ analysis has a high tagging efficiency but often identifies wrong triplets for the top reconstruction due to high combinatorics. The area analysis offers a slightly better result than the fat jet analysis for the reconstruction of the top, but is does not reconstruct the Higgs properly and has a lower tagging efficiency than the fat jet analysis.

We expect that one finds better results after including the cut on the Higgs mass in all analyses. This would help especially for the area analysis, since there the Higgs reconstruction was not satisfying. Furthermore, one could try harsher cuts for the Mass Jump parameters to decrease mis-tagging. This could improve the pure MJ analysis.

Acknowledgments

I would like to thank Tilman Plehn for the possibility to write my bachelor thesis in his group and for the support he gave me. It has been a great experience for me to work in a scientific environment and a good insight in the every-day work of physicists. Furthermore, I would like to thank Torben Schell and Karl Nordström for answering all my questions and supporting me in what I did. Finally, I would like to thank the whole group for the kindness and openness towards me.

References

[1] David J. Griffiths. *Introduction to elementary particles*. Physics textbook. Wiley-VCH, Weinheim, 2., rev. ed. edition, 2008.

[2] Michael H. Seymour. Searches for new particles using cone and cluster jet algorithms: A Comparative study. *Z. Phys.*, C62:127–138, 1994.

[3] Tilman Plehn and Michael Spannowsky. Top Tagging. *J. Phys.*, G39:083001, 2012.

[4] Mariangela Lisanti. Lectures on Dark Matter Physics. In *Theoretical Advanced Study Institute in Elementary Particle Physics: New Frontiers in Fields and Strings (TASI 2015) Boulder, CO, USA, June 1-26, 2015*, 2016.

[5] G. Barenboim. Neutrino Physics. In *Proceedings, 2012 European School of High-Energy Physics (ESHEP 2012): La Pommeraye, Anjou, France, June 06-19, 2012*, pages 161–179, 2014.

[6] Anthony Zee. *Quantum field theory in a nutshell*. Princeton University Press, Princeton, NJ [u.a.], 2. ed. edition, 2010. Includes bibliographical references and index.

[7] Vardan Khachatryan et al. Measurement of the top quark mass using proton-proton data at $\sqrt{(s)} = 7$ and 8 TeV. *Phys. Rev.*, D93(7):072004, 2016.

[8] http://pdg.lbl.gov/2014/reviews/rpp2014-rev-top-quark.pdf.

[9] Vernon Barger and Roger Phillips. *Collider physics*. Frontiers in physics ; 71. Addison-Wesley Publishing, Reading, Mass., updated ed. edition, 1997. Includes bibliographical references and index.

[10] Matthew D. Schwartz. *Quantum field theory and the standard model*. Cambridge University Press, Cambridge [u.a.], 2014.

[11] Tilman Plehn. *Lectures on LHC Physics*, volume 886. Lect. Notes Phys., 2015.

[12] Gregor Kasieczka, Tilman Plehn, Torben Schell, Thomas Strebler, and Gavin P. Salam. Resonance Searches with an Updated Top Tagger. *JHEP*, 06:203, 2015.

[13] W. Bartel et al. Experimental Studies on Multi-Jet Production in e+ e- Annihilation at PETRA Energies. *Z. Phys.*, C33:23, 1986. [,53(1986)].

[14] Tilman Plehn, Gavin P. Salam, and Michael Spannowsky. Fat Jets for a Light Higgs. *Phys. Rev. Lett.*, 104:111801, 2010.

[15] S. Catani, Yuri L. Dokshitzer, M. H. Seymour, and B. R. Webber. Longitudinally invariant K_t clustering algorithms for hadron hadron collisions. *Nucl. Phys.*, B406:187–224, 1993.

[16] Yuri L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber. Better jet clustering algorithms. *JHEP*, 08:001, 1997.

[17] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The Anti-k(t) jet clustering algorithm. *JHEP*, 04:063, 2008.

[18] Stefan Höche. Introduction to parton-shower event generators. In *Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders (TASI 2014) Boulder, Colorado, June 2-27, 2014*, 2014.

[19] Mrinal Dasgupta, Lorenzo Magnea, and Gavin P. Salam. Non-perturbative QCD effects in jets at hadron colliders. *JHEP*, 02:055, 2008.

[20] Jonathan M. Butterworth, Adam R. Davison, Mathieu Rubin, and Gavin P. Salam. Jet substructure as a new higgs-search channel at the large hadron collider. *Phys. Rev. Lett.*, 100:242001, Jun 2008.

[21] Martin Stoll. Vetoed jet clustering: The mass-jump algorithm. *JHEP*, 04:111, 2015.

[22] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014.

[23] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An Introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015.

[24] Andy Buckley, Jonathan Butterworth, Leif Lonnblad, David Grellscheid, Hendrik Hoeth, James Monk, Holger Schulz, and Frank Siegert. Rivet user manual. *Comput. Phys. Commun.*, 184:2803–2819, 2013.

[25] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur. Phys. J.*, C72:1896, 2012.

Erklärung

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

Heidelberg, den 28. November 2016,