Department of Physics and Astronomy

University of Heidelberg

Master thesis

in Physics

submitted by

Torben Karl Schell

born in Heidelberg

2013

# With Top Tagging towards $t\bar{t}H$

This Master thesis has been carried out by Torben Karl Schell

at the

Institute for Theoretical Physics

under the supervision of

Prof. Tilman Plehn

**Zusammenfassung:**

Nach der Entdeckung des Higgsbosons 2012 müssen jetzt dessen Kopplungen an andere Teilchen betrachtet werden. Diese Untersuchungen werden zeigen, ob der im Standardmodell angenommene Mechanismus zur Brechung der elektroschwachen Symmetrie standhält. Besonders die Kopplung an das Top–Quark ist von Bedeutung, da diese die größte Yukawakopplung darstellt.
In dieser Arbeit versuchen wir den HepTopTagger, einen Algorithmus der hadronisch zerfallene Top–Quarks identifiziert, zu verbessern. Die vorgeschlagenen Änderungen reduzieren das Formen des Hintergrunds. Dies ist vor allem für Prozesse mit hoher Multiplizität des Endzustandes, wie z. B. den vollständig hadronischen Zerfallskanal von $t\bar{t}H$, nützlich. Des weiteren untersuchen wir mittels Fox–Wolfram Momenten inwieweit der Tagger durch das Ausnutzen von Winkelkorrelation zwischen den einzelnen Teiljets verbessert werden kann. In Analysen wie beispielsweise der Produktion eines Higgsteilchens zusammen mit einem Top–Quarkpaar ist die Anzahl der zugänglichen Ereignisse durch den minimalen transversalen Impuls der identifizierbaren Top–Quarks beschränkt. Wir versuchen daher den Anwendungsbereich des HEPTopTaggers in Richtung niedriger Impulse auszudehnen. Eine andere Methode um hadronisch zerfallene Top–Quarks zu rekonstruieren, ist das kürzlich vorgestellte Korb-Verfahren. Nachdem wir gezeigt haben, dass wir sowohl Signal- als auch Hintergrundunsicherheiten unter Kontrolle haben, demonstrieren wir, dass eine Untersuchung des vollkommen hadronischen Zerfallskanals von $t\bar{t}H$ am LHC im Bereich des Möglichen liegt. Hierzu verwenden wir die Korb-Methode, um die beiden Top–Quarks zu identifizieren.

**Abstract:**

After the discovery of the Higgs(-like) boson in 2012 one has now to study the couplings of this particle. Such analyses will indicate whether the mechanism of electroweak symmetry breaking assumed in the Standard Model is consistent. Of particular interest is the coupling to the top quark since it has the largest predicted Yukawa coupling.
In this thesis we try to improve the HEPTopTagger, a tool for identification of hadronically decaying top quarks. We propose changes in the algorithm leading to a reduced shaping of backgrounds. This might be useful for studies of high multiplicity final states like the fully hadronic decay channel of $t\bar{t}H$. Furthermore, we analyze possible improvements by using angular correlations of subjets in terms of Fox–Wolfram moments. In studies like Higgs production associated with a pair of top quarks the number of accessible signal events is limited by the minimal transverse momentum of the reconstructable top quarks. We therefore try to extend the HEPTopTagger towards lower transverse momenta.
Another recent tool for top identification are top buckets. After illustrating that we have control over the uncertainties in the signal and background simulations, we show that an analysis of the fully hadronic decay channel of $t\bar{t}H$ at the LHC should be possible by using the bucket approach to identify the two top quarks.

# Contents

# 1. Introduction

The Standard Model of particle physics (SM) is one of the success stories in physics today. After the discovery [1, 2] of a Higgs(-like) boson at the Large Hadron Collider (LHC) — its theoretical fathers have been awarded this year's Nobel prize — we are now at the stage were we have to analyze its couplings to other Standard Model particles to answer the question whether this new particle is really the Standard Model Higgs particle. In other words, is the assumption of a minimal electroweak symmetry breaking mechanism correct or are there any hints of new physics?

One major task is to study the Higgs-fermion couplings (Yukawa couplings) which are in the Standard Model fixed by the mass of the fermions and the vacuum expectation value of the electroweak symmetry breaking doublet, which can be measured *e.g.* in muon decays. Of particular interest is the top Yukawa coupling, the coupling between the Higgs boson and the heaviest quark. This is even more important since the top quark is the only fermion with a weak scale mass. In principle, this coupling could be extracted from processes like a Higgs boson decaying into two photons, since this process is mediated through a top triangle. However, such an extraction is highly model depend. Any new massive particle with electric charge would modify this coupling. Consequently, one has to use a process which allows for a direct measurement of the top Yukawa coupling (*i.e.* a tree-level process) like the associated production of a Higgs with a top-antitop pair. Because of its small cross section of $\sigma_{t\bar{t}H} = 611$ fb [3] at the LHC running at a center of mass energy of $\sqrt{s} = 14$ TeV, one is focused on the dominant branching ratios, *i.e.* hadronically decaying top quarks and the Higgs decay into a bottom quark and its antiparticle. However, this channel has been removed from the list of possible discovery channels for a Higgs boson early on, since it is ruined by combinatorial backgrounds caused by four $b$ jets in the final state. Nevertheless, recent developments concerning the reconstruction of boosted objects have changed this picture. Any deviation of the top Yukawa coupling from the Standard Model value would signal the presence of new physics.

The task is to reconstruct the Higgs and the top quarks from its decay products (tagging). In this thesis we will study possibilities to improve one particular tool that allows to reconstruct moderately boosted top quarks, the Heidelberg-Eugene-Paris Top Tagger [4] (HEPTopTagger). Furthermore, we demonstrate that an analysis of the fully hadronic decay channel is indeed possible using the most recently proposed technique to tag low $p_T$ top quark via buckets [5].

In Section 2 we give a short overview about the Standard Model of particle physics, Monte Carlo event simulations and techniques we are using throughout our studies. Once these concepts are set, we study in Section 3 possible improvements of the HEPTopTagger. In Section 4.2 we show by an detailed study of QCD backgrounds that an analysis of the fully hadronic decay channel of $t\bar{t}H$ is not impossible at all. We end this thesis by summarizing our results and give an outlook of further studies in Section 5.

The results presented in this thesis were obtained in collaboration with Tilman Plehn, Matt Buckley, Michihisa Takehuchi and Catherine Bernaciak. The studies concerning the HEPTopTagger benefited from discussions with experimentalist, namely Christoph Anders, Gregor Kasieczka and Sebastian Schätzel. The analyses and results in this thesis follow [6, 7].

# 2. Basics & Concepts

We start by giving an overview on the basic concepts that are connected with the topics of this thesis. Of course nothing written down here is new and based on the author's work. Therefore, it is clear that the way things are presented is largely subjected to the various text books and lecture notes the author used to learn the particular subjects. Among those we will follow closely [8, 9, 10, 11].

## 2.1. Standard Model of particle physics

The Standard Model of particle physics [12, 13, 14] is a quantum field theory with the gauge group $SU(3)_C \times SU(2)_L \times U(1)_Y$. In this section we will focus on the electroweak gauge group $SU(2)_L \times U(1)_Y$ consisting of the left-handed $SU(2)_L$ and the hypercharge gauge group $U(1)_Y$ and its breaking [15, 16, 17] to the electromagnetic symmetry group $U(1)_Q$. The color gauge group $SU(3)_C$ will be studied in more detail in the next section.

The fermionic particle content of the Standard Model and the corresponding representations under the gauge groups are summarized in Tab. 2.1. The Lagrangian for the fermion fields is given by

$$\mathcal{L}_f = \sum_L \bar{L} i\gamma^\mu D_\mu^L L + \sum_R \bar{R} i\gamma^\mu D_\mu^R R \,. \tag{2.1}$$

We denote the left-handed doublets and right-handed singlets as $L$ and $R$ and introduce the combined covariant derivatives

$$D_\mu^L = \mathbf{1}\partial_\mu + ig_2 W_\mu^a \frac{\tau^a}{2} + ig_1 Y B_\mu \frac{\mathbf{1}}{2} + ig_3 G_\mu^a \frac{\lambda^a}{2} \tag{2.2}$$

$$D_\mu^R = D_\mu^L|_{\tau=0} \,, \tag{2.3}$$

where for fermions uncharged under $SU(3)_C$, *i.e.* leptons and neutrinos, the $g_3$ term is absent. We denote the Pauli matrices by $\tau^a$ and the Gell-Mann matrices by $\lambda^a$. From this definition it is already obvious that mass terms of the type $\frac{1}{2}\bar{L}R$ are forbidden due to the $SU(2)_L$ symmetry. The gauge fields are $B_\mu$ for the $U(1)_Y$ hypercharge, the three $W^a$ of $SU(2)_L$, and the 8 gluons $G^a$ of $SU(3)_C$. Their dynamics are governed by

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4} B^{\mu\nu} B_{\mu\nu} - \frac{1}{2}\text{Tr}(W^{\mu\nu} W_{\mu\nu}) - \frac{1}{2}\text{Tr}(G^{\mu\nu} G_{\mu\nu}) \,, \tag{2.4}$$

where we introduced the field strength tensors

$$F_{\mu\nu} = \frac{-i}{g} [D_\mu, D_\nu] \tag{2.5}$$

using the covariant derivative for the considered gauge group.

To construct massive fermions and gauge boson we introduce a $SU(2)_L$ doublet field that acquires a vacuum expectation value (VEV) and thus breaks the electroweak symmetry spontaneously.

$$\mathcal{L}_\Phi = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi^\dagger \Phi) \tag{2.6}$$

with the famous mexican hat potential

$$V(\Phi^\dagger \Phi) = \mu^2 (\Phi^\dagger \Phi) + \lambda (\Phi^\dagger \Phi)^2 \, . \tag{2.7}$$

In unitary gauge the doublet $\Phi$ reads

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \, , \tag{2.8}$$

where $v$ denotes the vacuum expectation value of $\Phi$ and $H$ is the Higgs field. With this at hand one finds

$$(D_\mu \Phi)^\dagger (D_\mu \Phi) \supset \frac{g_2^2 v^2}{8} (W_\mu^1 W^{1\mu} + W_\mu^2 W^{2\mu}) + \frac{v^2}{8} \left(W_\mu^3, B_\mu\right) \begin{pmatrix} g_1^2 & -g_1 g_2 \\ -g_1 g_2 & g_1^2 \end{pmatrix} \begin{pmatrix} W^{3\mu} \\ B^\mu \end{pmatrix} \, . \tag{2.9}$$

By diagonalization of the mass matrix one receives the two neutral physical fields $A_\mu$ and $Z_\mu$ as well as the charged W bosons.

$$(D_\mu \Phi)^\dagger (D_\mu \Phi) \supset \frac{g_2^2 v^2}{4} (W_\mu^+ W^{+\mu} + W_\mu^- W^{-\mu}) + \frac{g_1^2 + g_2^2}{8} v^2 Z_\mu Z^\mu \, , \tag{2.10}$$

where

$$W_\mu^\pm = \frac{1}{\sqrt{2}} (W_\mu^1 \pm W_\mu^2), \qquad \begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} W_\mu^3 \\ B_\mu \end{pmatrix} \, . \tag{2.11}$$

The corresponding weak mixing or Weinberg angle follows from the coupling constants as

$$\sin\theta_W = \frac{g_1}{\sqrt{g_1^2 + g_2^2}} \quad \cos\theta_W = \frac{g_2}{\sqrt{g_1^2 + g_2^2}} \, . \tag{2.12}$$

The couplings of the gauge bosons to the Standard Model fermions $f$ follow from the covariant derivative in the Lagrangian

$$\mathcal{L}_{\text{f-gauge}} = i \bar{f} \not{D} f = i \bar{f} \not{\partial} f + \mathcal{L}_{\text{f-W}} + \mathcal{L}_{\text{f-Z}} + \mathcal{L}_{\text{f-A}} \, . \tag{2.13}$$

The fermion couplings to the charged $W$ bosons are given by

$$\mathcal{L}_{\text{f-W}} = -\frac{g}{\sqrt{2}} \bar{u}_L \gamma^\mu W_\mu^+ d_L - \frac{g}{\sqrt{2}} \bar{d}_L \gamma^\mu W_\mu^- u_L \, , \tag{2.14}$$

while the coupling to the neutral $Z$ is dictated by

$$\mathcal{L}_{\text{f-Z}} = \frac{g}{\cos\theta_W} \left( q_f \left[ \bar{f}_L \gamma_\mu f_L + \bar{f}_R \gamma_\mu f_R \right] - \sin(\theta_W) T_3 \bar{f}_L \gamma_\mu f_L \right) Z^\mu \, . \tag{2.15}$$

Using the Gell-Mann-Nishijima formula $q_f = T_3 + Y/2$ for the electromagnetic charge we can write down the fermion-photon coupling

$$\mathcal{L}_{\text{f-A}} = q_f \bar{f} \gamma_\mu f A^\mu \, . \tag{2.16}$$

(2.17)

The introduction of $\Phi$ also allows the construction of fermion mass terms via

$$-g_Y \bar{L}\Phi R \to -g_Y \frac{v}{\sqrt{2}} \bar{L}^d R^d = -m_d \bar{L}^d R^d \; . \tag{2.18}$$

For up-type masses we need a doublet with hypercharge -1 the minimalistic choice is

$$-g_Y \bar{L}\tilde{\Phi} R \to -g_Y \frac{v}{\sqrt{2}} \bar{L}^u R^u = -m_u \bar{L}^u R^u \; , \tag{2.19}$$

with the charge conjugated of the $\Phi$ doublet $\tilde{\Phi} = \Phi^C = i\tau_2 \Phi^* \to \frac{1}{\sqrt{2}} \begin{pmatrix} v \\ 0 \end{pmatrix}$.
Including all three generations leads to the slightly more complicated form

$$-\bar{\mathbf{L}}'_d g_d \Phi \mathbf{R}'_d - \bar{\mathbf{L}}'_u g_u \tilde{\Phi} \mathbf{R}'_u \to -\bar{\mathbf{L}}'_u M_u \mathbf{R}'_u - \bar{\mathbf{L}}'_d M_d \mathbf{R}'_d \; , \tag{2.20}$$

where $\mathbf{L}'$ and $\mathbf{R}'$ are vectors in flavor space, while $g_u$, $g_d$ and therefore $M_u$ and $M_d$ are matrices. In general, the flavor eigenstates (3 generations) have not to coincide with the physical mass eigenstates. The physical fields can be found by diagonalizing the mass matrix according to $M_i = U_i D_i V_i^\dagger$ with $U_i^\dagger U_i = \mathbf{1} = V_i^\dagger V_i$. Therefore, we find the mass eigenstates $\mathbf{L}'_i = U_i \mathbf{L}_i$ and $\mathbf{R}'_i = V_i \mathbf{R}_i$ This mixing matrix will show up in some couplings as well. While neutral weak currents are not effected (no flavor changing neutral currents), charged currents are modified by the CKM matrix $U_{\text{CKM}} = U_u^\dagger U_d$. This $3 \times 3$ matrix allows for a CP violating complex phase and can be parametrized in the so called Wolfenstein parametrization as

$$U_{\text{CKM}} = \begin{pmatrix} 1 - \frac{\lambda^2}{2} & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1 - \frac{\lambda^2}{2} & A\lambda^2 \\ A\lambda^3(1 - \rho - i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4) \; . \tag{2.21}$$

Recent values for the Wolfenstein parameters can be found in [18].

| | | |
|---|---|---|
| Leptons | $(1, 2, Y_L)$ | $\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L, \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L$ |
| | $(1, 1, Y_R)$ | $e_R$, $\mu_R$, $\tau_R$ |
| Quarks | $(3, 2, Y_L^q)$ | $\begin{pmatrix} u \\ d' \end{pmatrix}_L, \begin{pmatrix} c \\ s' \end{pmatrix}_L, \begin{pmatrix} t \\ b' \end{pmatrix}_L$ |
| | $(3, 1, Y_R^q)$ | $u_R$, $d'_R$, $c_R$, $s'_R$, $t_R$, $b'_R$ |

Table 2.1.: Fermionic particle content of the Standard Model and their representations under the gauge group $SU(3)_C \times SU(2)_L \times U(1)_Y$. The hypercharges are $Y_L = -1$, $Y_R = -1$, $Y_L^q = 1/3$, $Y_R^{u,c,t} = 4/3$, and $Y_R^{d,s,b} = -2/3$.

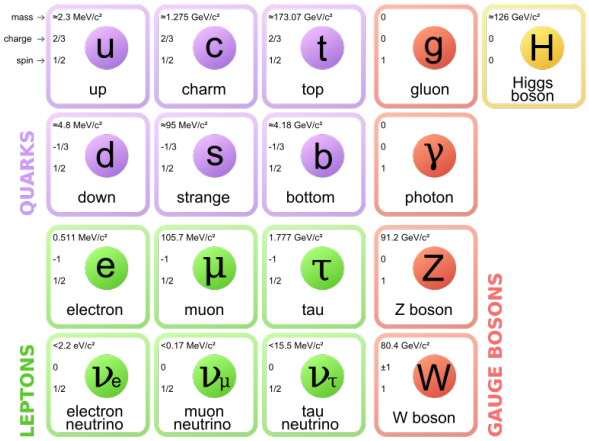


Figure 2.1.: Overview on the particle content of the Standard Model. Taken from [19].

The theory lined out above assumes massless neutrinos which is known to be incorrect since there is experimental prove of neutrino flavor oscillations. This fact can be implemented by *e.g.* postulating heavy right-handed neutrinos. However, this is not of interest for this thesis. Therefore, we will put this complication aside and refer to the literature [20]. Interactions between quarks and gluons are described by Quantum Chromodynamics (QCD) which will be discussed in the next section. All particles contained in the Standard Model are summarized with some of their properties in Fig. 2.1.

We know that there must be physics beyond the Standard Model. Despite some probably taste dependent problems like the electroweak hierarchy problem, we know *e.g.* from measurements of rotation curves of galaxies that there must be non Standard Model gravitating matter (dark matter) [21] and the recent accelerated expansion of the universe raises the question for dark energy [22].

## 2.2. QCD and Monte Carlo simulations

Quantum Chromodynamics (QCD) is a non-abelian gauge theory with gauge group $SU(3)$ that describes the interactions between colored objects, *i.e.* quarks and gluons. The defining property of the generators is given by their anti-commutation relations

$$[T^a, T^b] = if^{abc}T^c \,, \tag{2.22}$$

where $f^{abc}$ are the totally antisymmetric structure constants of the gauge group. The 8 generators are given by $T^a = \frac{\lambda^a}{2}$ with the Gell-Mann matrices $\lambda^a$. While the gauge fields live in the adjoint representation, the quarks are considered to transform under the fundamental representation. The Lagrangian neglecting Fadeev-Popov ghost fields and a gauge fixing term reads

$$\mathcal{L}_{QCD} = -\frac{1}{2}\mathrm{Tr}\left(G_{\mu\nu}G^{\mu\nu}\right) - i\bar{\Psi}\gamma^\mu D_\mu \Psi \tag{2.23}$$

with the field strength tensor

$$G_{\mu\nu} = \frac{-i}{g_3}\left[D_\mu, D_\nu\right] \tag{2.24}$$

and the covariant derivative

$$D_\mu = \partial_\mu + i g_3 G_\mu \; . \tag{2.25}$$

$G_\mu = G^a_\mu T^a$ gives rise to the 8 gluon fields $G^a_\mu$. Expanding the filed strength tensor and using the properties of the generators the physics described by the Lagrangian become clearer

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_3 f^{abc} G^b_\mu G^c_\nu \; . \tag{2.26}$$

Therefore, the trace term leads beside the gluon propagator to triple and quartic self-interactions.

Under the most important quantities in particle physics are cross sections of a specific processes. In principle, given the Feynman rules one can compute the matrix elements and integrate over the phase space to receive a result for the total cross section. However, calculating a meaningful cross section that should predict the numbers experimentalists find at the LHC is far from trivial. Below we will sketch what are the complications and how one can simulate LHC processes using Monte Carlo techniques.

The physics in which one is usually interested in are encoded in the hard process. Here one considers the interaction on parton level, *i.e.* one is dealing with free incoming and outgoing particles. Using for example Feynman rules one can calculate the matrix element of the hard process. The first and obvious complication is that at hadron colliders like the LHC the incoming states are not at all free quarks which of course would violate the confinement of QCD. Instead, the incoming partons (gluons and quarks) are confined inside a proton. For simplicity, one considers the limit of zero relative transverse momentum inside the proton. The probability to find a parton $i$ with momentum fraction $x$ relative to the proton is described by the parton density functions (pdf) $f_i(x, \mu_F)$. Folding in the parton densities in the cross section calculation one receives the following master formula for inclusive cross sections at the LHC

$$\sigma^{tot}_{pp\to X} = \sum_{i,j} \int d\Phi f_i(x_1, \mu_F) f_j(x_2, \mu_F) \sigma_{ij\to X}(x_1, x_2, \mu_R) \; . \tag{2.27}$$

The total cross section for a final state $X$ including any number of collinear jets is the cross section for a certain initial state calculated at a renormalization scale $\mu_R$ multiplied with the probability of this initial state configuration in terms of parton densities, integrated over the phase space, and summed over for all possible incoming states. The discussion of the factorization scale $\mu_F$ we postpone to a later paragraph of this section.

An additional complication is that the quarks and gluons inside the proton can split. The same is true for final state gluons and quarks. For a proper description of the total process it is necessary to understand these splittings. A splitting in the final state leads from a $n$ to a $n+1$ particle final state which means that besides the splitting itself one has to worry about the changed phase space as well. The splitting of a parton $a$ into two partons $b$ and $c$ can be parametrized using the energy fraction $z = \frac{|E_b|}{|E_a|} = 1 - \frac{|E_c|}{|E_a|}$. To describe the kinematics of the splitting one can use the Sudakov decomposition [8]

$$-p_a = p_b + p_c = (-z p_a + \beta n + p_T) + (-(1-z) p_a - \beta n - p_T) \tag{2.28}$$

with an arbitrary unit vector $n$, $(p_a p_T) = 0 = (n p_T)$ and a free factor $\beta$. This decomposition allows only to set one of the outgoing partons on its mass shell. Therefore, $p_c^2 = 0 \neq p_b^2$. The $n+1$ body collinear phase space is given by

$$d\Phi_{n+1} = d\Phi_n \frac{p_{c,3} dp_T d\phi}{(2\pi)^3 2E_c} \frac{E_a}{E_b} = d\Phi_n \frac{p_{c,3} dp_T d\phi}{(2\pi)^3 2E_c z} = d\Phi_n \frac{dp_a^2 dz d\phi}{4(2\pi)^3} = d\Phi_n \frac{dp_a^2 dz}{4(2\pi)^2} \ . \tag{2.29}$$

This implies a factorization of the phase space. It can be argued that the squared matrix element factorizes as well giving rise to universal splitting kernels $\frac{2g_s}{p_a}\hat{P}(z)$ [8]. Finally, this leads for collinear splitting to

$$\sigma_{n+1} \approx \int \sigma_n \frac{dp_a^2}{p_a^2} dz \frac{\alpha_s}{2\pi} \hat{P}(z) \ . \tag{2.30}$$

The splitting kernels can be deduced to be [8]

$$\hat{P}(z)_{g\leftarrow q} = C_F \frac{1+(1-z)^2}{z} \tag{2.31}$$

$$\hat{P}(z)_{q\leftarrow g} = T_R(z^2+(1-z)^2) \tag{2.32}$$

$$\hat{P}(z)_{q\leftarrow q} = C_F \frac{1+z^2}{1-z} \tag{2.33}$$

$$\hat{P}(z)_{g\leftarrow g} = C_A \left( \frac{z}{1-z} + \frac{1-z}{z} + z(1-z) \right) \ , \tag{2.34}$$

where for $SU(3)$ $C_A = 3$, $C_F = 4/3$, and $T_R = 1/2$.

We can think of the splittings of incoming particles as t channel diagrams with $t = p_a^2 < 0$. An incoming parton of energy fraction $x_0$ and virtuality $t_0$ will perform several splittings before ending up in the hard process. Therefore, we need to describe its evolution in $x$–$t$ space. For simplicity we assume just gluon radiation and will forget about the radiated gluons. In each splitting the quark will loose energy (decreasing $x$) and will gain virtuality $t$. A possible path in x–t space is shown in Fig.2.2.

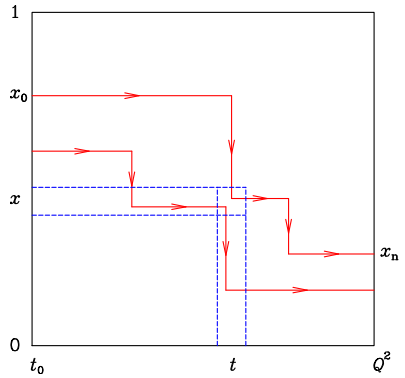


Figure 2.2.: Possible evolutions in x–t space. Taken from [11].

The change of the number of quarks sitting in a box $[x, x+\delta x][t, t+\delta t]$ is given by the leaving and entering quarks. To use paths instead of a sequence of points in the $x-t$ plane we

choose the following convention. We consider the splitting as vertical drops in $x$ at constant $t$ and choose the energy fraction $x$ to be constant despite of the splittings. Therefore, we have

$$\delta f(x,-t) = \delta f_{\text{in}} - \delta f_{\text{out}} \,. \tag{2.35}$$

The entering paths are given by the probability of a drop from some $x$ into the box convoluted with the probability of having such a quark integrated over $\delta t$

$$f_{\text{in}} = \delta t \left( \frac{\alpha_s \hat{P}(z)}{2\pi} \otimes f \right)(x,-t) = \frac{\delta t}{t} \int_0^1 \frac{dz}{z} \frac{\alpha_s}{2\pi} \hat{P}(z) f\left(\frac{x}{z}, -t\right) \,. \tag{2.36}$$

The quarks leaving the box vertically are given by

$$f_{\text{out}} = \delta t \int_0^1 dy \frac{\alpha_s}{2\pi t} \hat{P}(y) f(x,-t). \tag{2.37}$$

Using the definition of the plus subtraction scheme

$$F(z)_+ = F(z) - \delta(1-z) \int_0^1 dy F(y) \tag{2.38}$$

this can be summarized to

$$\frac{\delta f(x,-t)}{\delta(-t)} = \frac{1}{-t} \int_0^1 \frac{dz}{z} \hat{P}(z)_+ f\left(\frac{x}{z}, -t\right) \,. \tag{2.39}$$

Finally, we find the flow equation for the quark density

$$\frac{df_q(x,-t)}{d\log(-t)} = \sum_{j=g,q} \int_x^1 \frac{dz}{z} \frac{\alpha_s}{2\pi} P_{q\leftarrow j}(z) f_j\left(\frac{x}{z}, -t\right) \,, \tag{2.40}$$

where we now included the possibility of a gluon giving rise to a quark. The two regularized splitting kernels read $P_{q\leftarrow q}(z) = \hat{P}_{q\leftarrow q}(z)_+$ and $P_{q\leftarrow g}(z) = \hat{P}_{q\leftarrow g}(z)$. In complete analogy one can construct the evolution equation for the gluon density and bring it to a similar form by dumping all complications into the relevant regularized splitting kernels

$$P_{g\leftarrow q} = C_F \frac{1+(1-z^2)}{z} \tag{2.41}$$

$$P_{g\leftarrow g} = 2C_A \left( \frac{z}{(1-z)_+} + \frac{1-z}{z} + z(1-z) \right) + \frac{11}{6} C_A \delta(1-z) - \frac{2}{3} n_f T_R \delta(1-z) \,. \tag{2.42}$$

Identifying $-t$ with the factorization scale $\mu_F^2$, *i.e.* the virtuality we define the hard process to start we can summarize the gluon and quark densities in the DGLAP equation

$$\frac{df_i(x,\mu_F)}{d\log\mu_F^2} = \sum_j \int \frac{dz}{z} \frac{\alpha_s}{2\pi} P_{i\leftarrow j}(z) f_j\left(\frac{x}{z}, \mu_F\right) = \frac{\alpha_s}{2\pi} \sum_j (P_{i\leftarrow j} \otimes f_j)(x,\mu_F) \,. \tag{2.43}$$

Starting from the splitting functions we can follow a probabilistic approach of parton splitting which will lead to the concept of parton showers. The key objects in this business are the Sudakov form factors given by

$$\Delta_i(t) := \Delta_i(t, t_0) = \exp\left( - \sum_j \int_{t_0}^{t} \frac{dt'}{t'} \int_0^1 dy \frac{\alpha_s}{2\pi} \hat{P}_{j\leftarrow i}(y) \right) . \tag{2.44}$$

In the following we will discuss the evolution of a quark where we will always follow the quark leg:

$$\Delta_q(t) = \exp\left[ - \int_{t_0}^{t} \frac{dt'}{t'} \int_0^1 dy \frac{\alpha_S}{2\pi} \hat{P}_{q\leftarrow q}(y) \right] . \tag{2.45}$$

Inserting the normalized splitting kernel and using the substitution $t'' = (1-y)^2 t$ we can perform the y integration:

$$\int_0^1 dy \hat{P}_{q\leftarrow q}(y) = \frac{C_F}{2\pi} \int_0^1 dy \left[ \frac{2}{1-y} - 1 - y \right] \tag{2.46}$$

$$= \frac{C_F}{2\pi} \left[ \int_{t'}^{t} \frac{dt''}{t''} \alpha_S - \int_0^1 dy \alpha_S (1+y) \right] . \tag{2.47}$$

If we are just interested in the leading power dependence in t and y we can pull $\alpha_S$ in front of the integrals and solve them to

$$\int_0^1 dy \hat{P}_{q\leftarrow q}(y) = \frac{C_F}{2\pi} \alpha_S \left[ \log \frac{t}{t'} - \frac{3}{2} \right] =: t' \Gamma_{q\leftarrow q(t,t')} , \tag{2.48}$$

where we introduced the quark-quark splitting function $\Gamma_{q\leftarrow q}$. However, using the leading power dependence yields the problem that

$$\lim_{t' \to t} \Gamma(t, t') \neq 0 . \tag{2.49}$$

From the technical side this can be handled by requiring $\Gamma > 0$ and setting $\Gamma = 0$ otherwise.

To find a connection between the Sudakov form factor and a probabilistic description of parton splitting we compare it to a Poisson distribution describing the probability to find $n$ events if all events are independent and occur with probability $p$

$$\mathcal{P}(n; p) = \frac{p^n e^{-p}}{n!} . \tag{2.50}$$

The probability to find no event is $\mathcal{P}(0, p) = \exp(-p)$. Since we found

$$\Delta_i = \exp\left[ - \int_{t_0}^{t} dt' \Gamma(t, t') \right] , \tag{2.51}$$

where the exponent is the negative integrated splitting probability we can interpret the Sudakov form factor as the non-splitting probability. In the formula above we introduce a lower cut-off to avoid divergences. However, this cut–off drops out once we calculate the probability for no splitting between the virtualities $t_1$ and $t_2$ fulfilling $t_1 > t_2 > t_0 > 0$

$$\frac{\Delta_i(t_1)}{\Delta_i(t_2)} < 1 . \tag{2.52}$$

What we need is a way to simulate the evolution in virtuality–energy fraction space. Starting at a point $(t_1, x_1)$ we can get the next virtuality $t_2$ by generating a flat random number $r_t \in [0,1]$ via

$$\frac{\Delta(t_1)}{\Delta(t_2)} = r_t \ . \tag{2.53}$$

The next value for the energy fraction $x_2 < x_1$ we can get by generating a second flat random number $r_x \in [0,1]$ and solving

$$\frac{\int_0^{x_1/x_2} dy \frac{\alpha_S}{2\pi} \hat{P}(y)}{\int_0^1 dy \frac{\alpha_S}{2\pi} \hat{P}(y)} = r_x \ . \tag{2.54}$$

To deal with soft divergences ($y \to 0$ and $y \to 1$) of the unregularized splitting kernels one can introduce appropriate cut-offs. So far we can simulate the evolution in t-x space by generating two flat random numbers. However, what we need is the full four momentum vector. At this stage we have just two conditions; a third one follows from the on-shell mass. Therefore, we are left with one free parameter. But we know that QCD is insensitive to the azimuthal angle $\phi$. Consequently, we can fix it by generating a third flat random number between 0 and $2\pi$. All these steps allow us to simulate the evolution of final state partons in terms of Monte Carlo methods.

A critical disadvantage of hadron colliders like the LHC is the fact that one does not collide single partons. At the LHC the partons that enter the hard process are confined inside protons. In consequence, one will not only see the hard process one is interested in. The remnants of the colliding protons will give rise to an QCD background filling the detector. This background contribution is called underlying event. The influence of underlying event on the mass of jet constructed with a cone size $R$ is described by [23]

$$\langle \delta m_j \rangle \simeq \Lambda_{\mathrm{UE}} \, p_{T,j} \left( \frac{R^4}{4} + \frac{R^8}{4608} + \mathcal{O}(R^{12}) \right) \ , \tag{2.55}$$

where $\Lambda_{\mathrm{UE}}$ denotes the amount of transverse momentum of the underlying event radiation per unit rapidity and is for the LHC $\mathcal{O}(10\ \mathrm{GeV})$.
However, the situation at the LHC is even worse since one does not collide single protons but proton bunches. Therefore, one will not have only a single proton–proton collision but several collisions in the detector. Those secondary collisions are called pile up. Despite these two sources of QCD radiation that do not originate from the partons involved in the hard process there are two more. First, each incoming parton is a source of initial state radiation. This radiation arises before the parton enters the hard process and hence may fudge the jet clustering. In contrast final state radiation emitted from outgoing partons must be included in the clustering process to correctly reconstruct the partons. Methods that aim for a reduction of the influence of underlying event and pile up on a jet are denoted jet grooming. Two of those methods are filtering [24] and pruning [25].

There is one major problem when dealing with final state quarks and gluons. On the one hand we know that the parton shower will give good results simulating soft and collinear splittings; on the other hand one is often interested in hard jets. The reason for this is that if one studies signal processes with hadronical decays of heavy particles the decay jets have typically high transverse momenta due to the mass drop. Accordingly, we should generate hard jets as part of the matrix element. The problem is that both approaches should overlap

since they are describing both the radiation of jets. Therefore, the parton shower could *e.g.* generate an additional hard jet which would transform a $n$ jet to a $n+1$ jet event. This means that adding the samples after showering would lead to double–counting of events. In this context double–counting means any wrong treatment even missing an event. Consequently, we need a procedure which allows us to combine samples with different matrix element level multiplicities in a proper way. The two most common are the CKKW [26] and MLM [27] matching. While the MLM scheme is based on the comparison of the showered final state and the matrix level truth the CKKW algorithm vetoes specific splittings in the parton shower.

In the MLM merging we start by running a jet cone algorithm with cone size $R_{\mathrm{MLM}}$ on the final state to construct jets above a transverse energy threshold $E_{T,\mathrm{MLM}}$ which corresponds to the minimum transverse energy used for the additional hard partons during event generation. After this step we try to match the list of found jets with the parton level truth at matrix level. Starting with the hardest parton we assign it to the jet which is closest in the $\eta$-$\phi$ plane. If the distance between the two is below $R_{\mathrm{MLM}}$ they match. The matched jet is removed from the list of jets and one goes on trying to find a matching for the second hardest parton in the same way. If there is a matching for each hard parton and no jet is left after matching, the event is accepted; otherwise it is rejected. Only for the highest multiplicity sample the last condition is slightly changed. In that case one allows for left over jets after the matching is done. By using this recipe one ends up with exclusively defined event samples for the different multiplicities which now can be combined without any further restrictions.

An alternative method for merging parton shower and hard matrix element simulations of jets is the CKKW scheme which is the default in SHERPA. In contrast to the MLM we aim for a clear separation of matrix element and parton shower simulation in terms of a vetoed parton shower. We split the phase space according to a matching scale $y_{\mathrm{match}}$. The way we turn inclusive to exact $n$ particle final states is based on non-splitting probabilities. Sudakov factors allow us to construct fixed multiplicity final state without additional jets above a given resolution scale. We start by generating events and calculating the leading order cross sections $\sigma_{n,i}$ for all $n$-jet processes, where we distinguish the $i$ non-interfering configurations. To guarantee hard jets as well as finite results for the cross sections, we use a lower jet radiation cut-off $t_{\mathrm{match}}$. Each event is then assigned a weight based on the following steps:

1. Select a final state according to its relative probability $P_{n,i} = \sigma_{n,i} / \sum_{k,j} \sigma_{k,j}$.

2. Assign the momenta from a phase space generator to the hard external particles and compute the matrix element squared plus parton shower below $t_{\mathrm{match}}$.

3. Using a jet algorithm construct the shower history and check the splittings to correspond to possible Feynman diagrams and not to spoil any symmetries.

4. Use Sudakov factors to calculate the non-splitting probability of each external and internal line until $t_{\mathrm{match}}$ is reached.

5. Use the information ($k_T$ scales) of the shower history to adjust the values used for the strong coupling constant.

6. Build the final weight from the matrix element, the Sudakovs and the adjustment of $\alpha_S$.

This algorithm entirely relies on the matrix element simulation for the hard patrons above the scale $t_{\text{match}}$, while below this scale the evolution is generated using the parton shower. Unlike in the MLM matching, $t_{\text{match}}$ clearly separates the regions of the two approaches.

Up to this point we are dealing with quarks as fundamental objects. But we know that this picture is not true in the low energy limit. The objects we measure in the detector are hadrons. The models that are used to obtain hadrons from the quarks and gluons are non-perturbative and fitted to experimental observations. For more details on Monte Carlo generation see *e.g.* [28].

Although the particle detectors ATLAS [29] and CMS [30] perform incredibly well, the experimental data of a proton–proton collision will always be subjected to limited resolution, misidentifications and efficiencies below 100%. To account for such intrinsic experimental limitations one can use fast detector simulations like DELPHES [31] to mimic these results. If one needs an even more accurate simulation of the actual detector response there are programs aiming for a full detector simulation like GEANT [32] including the detector design which can for example give rise to dead regions. Such simulations are beyond our scope of interest. In our studies we restrict ourselves to a simple smearing in the $\eta - \phi$ plane with a resolution of $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ and a minimal transverse momentum for the detection of hadronic activity of 0.5 GeV after reclustering according to the given resolution.

## 2.3. Collider phenomenology

On the level of partons it is quite natural to describe the kinematics in the center of mass frame. The same is true for an electron–positron collider. For hadronic beams like at the LHC we have to deal with a more complicated situation. Even in the limit of vanishing transverse momenta of the partons relative to the beam, which is justified for high energies, the center of mass of the interacting partons is not an accessible reference frame. This fact translates into the usual choice of kinematic variables. While the transverse momentum relative to the beam axis is still a good variable one needs a second one that should transform under longitudinal boosts in a simple way. The quantity of choice is the rapidity

$$y = \frac{1}{2} \log \frac{E - p_z}{E + p_z} \,, \tag{2.56}$$

which is additive under longitudinal boosts. For massless particles it is identical to the pseudorapidity $\eta$

$$y := \frac{1}{2} \log \frac{E + p_z}{E - p_z} \approx \frac{1}{2} \log \frac{1 + \cos\theta}{1 - \cos\theta} =: \eta \,. \tag{2.57}$$

Together with the azimuthal angle in the transverse plane $\phi$ and the transverse momentum $p_T$ this allows us to describe the kinematics of the particles. Another quantity that is quite often used is the angular distance between two objects

$$\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} \,. \tag{2.58}$$

This parametrizes the distance of two objects in the $\eta$-$\phi$ plane and plays the dominant role in jet clustering algorithm outlined below. A typical value for the definition of jets is $R = 0.4$ being the ATLAS default for jet reconstruction. Later on we will have to deal with much larger cone sizes of $R = 1.5$ and $R = 1.8$.

What experimentalists see as outcome of an event is a bunch of color neutral hadrons or more precisely lots of colorimeter towers. From these objects one has to reconstruct the outgoing partons of the hard process. Algorithms that try to trace the hadrons back to their QCD correspondence in the hard process are referred to as jet algorithms. They aim to undo the splittings of a quark or gluon, which are most often soft and collinear. Starting from the calorimeter cells one needs a measure $d_{ij}$ to decide whether two objects should be joint or not. The distance to the beam axis is denoted as $d_{iB}$. Today there are three commonly used measures:

- $k_T$ algorithm [33]: $d_{ij} = \min(p_{T,i}, p_{T,j}) \frac{\Delta R_{ij}}{R} \quad d_{iB} = p_{T,i}$
- Cambridge/Aachen algorithm [34]: $d_{ij} = \frac{\Delta R_{ij}}{R} \quad d_{iB} = 1$
- anti-$k_T$ algorithm [35]: $d_{ij} = \min(\frac{1}{p_{T,i}}, \frac{1}{p_{T,j}}) \frac{\Delta R_{ij}}{R} \quad d_{iB} = \frac{1}{p_{T,i}}$

The jet clustering consists of the following steps:

1. Find the minimal distance of all subjets $d_{\min} = \min(d_{ij}, d_{iB})$
2. If $d_{\min} < d_{\text{cut}}$, join the two subjets. For the case that the minimal distance is of the kind $d_{iB}$ remove subjet $i$ (beam radiation)

3. Iterate until there is only one jet left or $d_{\min} > d_{\text{cut}}$.

The resulting objects of this procedure are labeled jets. An alternative to this exclusive version there is the inclusive algorithm that does not need a given cutoff value $d_{\text{cut}}$. In this version all subjets are clustered as given above until $d_{\min} = d_{iB}$. In that case the object $i$ is considered as a jet and removed from the list of substructures. This procedure is iterated until there are no subjets above a certain threshold (usually $p_T$) are left. All these algorithms are provided in FastJet [36] that we will use for jet clustering throughout this thesis.

Of course, the different algorithms will lead to different results and different jet shapes in the $\eta$–$\phi$plane. A well defined area in the plane would be nice to estimate the influence of the underlying event. From a QCD point of view one should expect the $k_T$ algorithm to give the best results, since it clusters pairs with one soft object first. However, C/A seems to perform better for substructure analyses. While the physical motivation for anti-$k_T$ is obscure, it produces almost circular jet shapes and is therefore used most often in experimental analyses. A comparison of the resultant jet shapes is given in Fig. 2.3. These figures were produced by measuring the jet area by adding essentially mass less objects (ghosts) to the event and checking if they end up in one of the jets and in which one in particular. While the shapes of the $k_T$ and Cambridge/Aachen jets are quite diffuse, the anti-$k_t$ algorithm results in almost circular shapes that allow an easy estimation of the influence of underlying event.

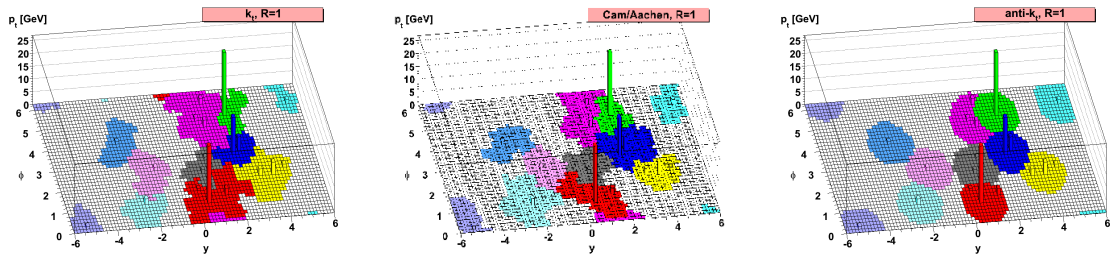


Figure 2.3.: Comparison of different jet algorithms. Taken from [35].

## 2.4. Higgs production associated with a top pair

Among the Yukawa couplings of the Higgs the top Yukawa coupling is one of the most interesting. Its special role is already given by the mass of the top quark which implies a coupling $y_t \sim 1$. Measuring its value is a crucial precision test of the mechanism of electroweak symmetry breaking, *i.e.* is the particle observed at the LHC really the Standard Model Higgs boson? The most prominent occurrence of a Higgs coupling to a top quark is for sure the top loop giving rise to the effective couplings of the Higgs to massless particles, *i.e.* gluons and photons $g_{ggH}$ and $g_{\gamma\gamma H}$. The first one results in the dominant Higgs production mechanism at the LHC; the second defines one of the discovery channels. From those effective couplings one can in principle deduce the top Yukawa coupling. However, such an extraction will always rely on the SM as underlying theory. In theories beyond the Standard Model such as supersymmetry there might be new massive particles that will contribute to the effective couplings if they have the corresponding charges. To avoid such model dependencies one needs a direct, *i.e.* tree level, coupling.

The simplest process fulfilling this condition is the associated production of a Higgs together with a top pair $pp \to t\bar{t}H$. Since the process $t\bar{t}H$ already has a small cross section of 611 fb [3] for a Higgs mass of $m_H = 125$ GeV at $\sqrt{s} = 14$ TeV, one is interested in the dominant decay channel, *i.e.* $H \to b\bar{b}$ with a branching ratio of 57.7% [3, 37] and hadronic top decays. A possible Feynman graph for the fully hadronic decay channel is given in Fig. 2.4. In principle, this decay allows for a complete reconstruction of the hard process. However, this comes with the price of 4 $b$ quarks in the final state which can be combined in various ways to form a Higgs candidate. In this way even the signal events contribute to the background via combinations of $b$ quarks not originating from a Higgs decay. This problem is known as combinatorial background and is one of the reasons why the process $t\bar{t}H$ was removed from the list of possible Higgs discovery channels. A possible way around this problem is to reconstruct two top quarks and then just take the $b$ quarks that are left after the top reconstruction. In Section 4.2 we propose such an analysis using the bucket algorithm [5]. We show that this should indeed allow a search for the Higgs in the fully hadronic channel of $t\bar{t}H$ at the LHC. Another tool to reconstruct top quarks is the HEPTopTagger [4] which will be studied in Section 2.6 where we try to optimize this algorithm with the long–term goal of using it for $t\bar{t}H$.

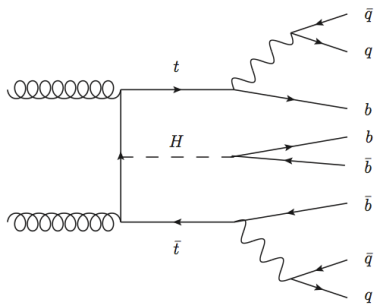


Figure 2.4.: Example Feynman graph for associated Higgs production with a top pair with fully hadronical final state.

## 2.5. Some top physics

This section summarizes some general properties of the top quark. A discussion of how to identify hadronically decaying top quarks (top tagging) at the LHC is subject of the next section.

The top quark was postulated [38] in 1973 by Kobayashi and Maskawa as part of a third quark generation that would explain the observed CP asymmetry in kaon decays. While the lighter, down-like $b$ quark was already found in 1977 at the Tevatron, it lasted until 1995 until the up-like top quark was found by CDF [39] and D0 [40] in proton–antiproton collisions.

With its mass of $m_t = 173$ GeV [18] it is by far the heaviest fundamental particle in the Standard Model. Another distinct property of the top quark is that it decays before it can hadronize [41]. From its mass being of order of the weak scale follows an expected Yukawa coupling $y_t \sim 1$ which makes the top quark well-suited to study electroweak symmetry breaking. While at the Tevatron the dominant production channel was quark–antiquark annihilation, at the LHC the top pairs are produced most often via gluon fusion. Some leading order Feynman graphs for top pair production are illustrated in Fig. 2.5.

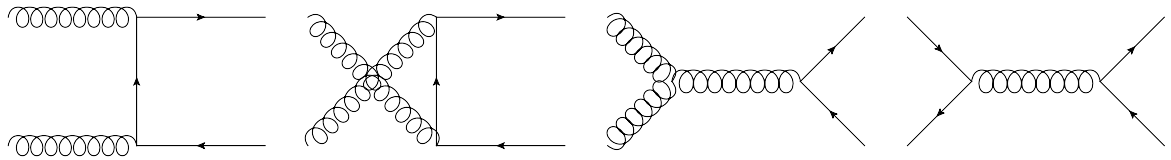

Figure 2.5.: The dominant Feynman graphs for top pair production.

Almost all top quarks decay to a $W$ boson and a $b$ quark. The PDG branching ratio for this decay is 99%. Hence, the final states of the top decay are dictated by the $W$. The $W$ boson itself can decay to lepton plus neutrino or quark antiquark. Assuming lepton universality the branching ratios are $\mathrm{Br}(W^+ \to l_i^+ \nu_i) = 1/9$. The decay to quarks has a branching ratio of 6/9. The Feynman graph of a hadronical top decay is shown in Fig. 2.6. To avoid the possibility of hadronical decaying $\tau$ leptons we will limit ourselves to electrons and muons in leptonical decays for the rest of this thesis.

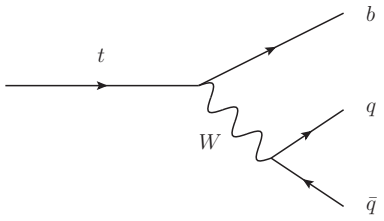


Figure 2.6.: Feynman graph for a hadronical decaying top quark.

## 2.6. Finding top quarks

The main problem in top analyses is to find and reconstruct the top quarks from the calorimeter towers resulting of an event. Most problematic in this business is the QCD background distributed over the entire detector. A top at rest decays in a typical Mercedes star shape. If we have to deal with a QCD busy environment like at the LHC, there is no chance to find all decay products since they are spread in all directions. However, the situation changes if we consider top quarks that are not at rest. In that case the decay products will be collimated due to the boost and will end up in a jet of relatively large cone size, a so called fat jet. Fig. 2.7 shows the angular separation of the top decay products $R_{bjj}$ corresponding to the cone size inside which one should be able to find them in dependency of transverse momentum of the decaying top quark.

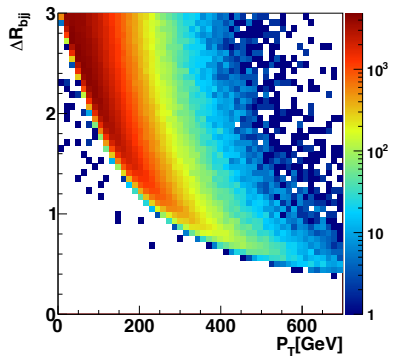


Figure 2.7.: Cone size $\Delta R_{bjj}$ in which all decay products of a top quark can be found in dependence of the transverse momentum $p_T$ of the decaying top quark for top pair production at a center of mass energy of $\sqrt{s} = 14$ TeV. Taken from [4].

For large transverse momentum the decay products are well collimated. But the relative number of tops lying in this kinematic regime is small. Therefore, one has to find a compromise between the fraction of top quarks available for analysis and $R_{bjj}$, which leads to moderately boosted top quarks. A tool that tackles the reconstruction of those tops is the HEPTopTagger [4]. It aims for tops with $p_T > 200$ GeV giving rise to fat jets with cone sizes below $R_{\text{fat}} = 1.5$. A nice feature of this algorithm is that it does not include a b-tag. The tagging procedure consists of the following steps:

1. **Construct fat jets**: First, we construct jets following the Cambridge/Aachen algorithm with a large cone size $R_{\text{fat}}$ (default $R_{\text{fat}} = 1.5$) such that all decay products of the boosted top quark should be included in a fat jet. The constructed fat jet has to fulfill $p_{T,\text{fat}} > 200$ GeV.

2. **Search for hard substructure**: Next, we analyze the substructure of each fat jet. Therefore, we undo the last clustering step and receive two subjets $j \to j_1 j_2$ which we define such that $m_{j_1} > m_{j_2}$. If $j_2$ is a typically light QCD jet from initial state radiation, underlying event or pile up we want to drop it. This is achieved using a mass drop criterion: If $m_{j_1} > f_{\text{drop}}\, m_j$, we only keep $j_1$ (default: $f_{\text{drop}} = 0.8$). Otherwise we keep both. We decompose the kept subjets further with the same procedure until all

subjets are lighter than a certain cutoff $m_{\text{sub}}$ (default: $m_{\text{sub}} = 30$ GeV). All remaining subjets are further considered as hard substructures.

3. **filtering**: At this stage we have a certain number of hard substructures. Three of them should correspond to top decay products. Therefore, we pick one possible triple and try to reconstruct a top from it. To do so, we first filter the triple. Filtering means that we start again from the constituents these jets are built from and cluster them again but now with a reduced radius $R_{\text{filter}} = \min(R_{\min}, \Delta R_{ij}/2)$, where $\Delta R_{ij}$ is the angular separation of two substructures in the picked triple. The new clustering yields in general more jets than we started from. From these we keep just the $N_{\text{filter}}$ hardest and cluster them back to 3 jets labeled $j_1$, $j_2$, and $j_3$ ordered by their transverse momenta. The effect of this filtering procedure is that the area in the $\eta$-$\phi$ plane that is used to build the 3 jets is reduced and hence we are less sensitive to underlying event and pile up. The filtered triple is now considered as a top candidate. Since every triple leads to a different top candidate, one has to choose which one is the best and will be analyzed further. For this purpose we use the difference between the candidates mass and the top mass $|m_{123} - m_t|$. Only the filtered triple that minimizes this measure is taken for the further steps.

4. **mass plane cuts**: In the high energy limit we can use $p_i^2 = 0$ to write $m_{123}^2 = m_{12}^2 + m_{13}^2 + m_{23}^2$. For $m_{123} = m_t$ this defines the surface of a sphere. We can parametrize this surface by two angles $\frac{m_{23}}{m_{123}}$ and $\arctan\left(\frac{m_{23}}{m_{12}}\right)$. For a real top quark we should be able to identify two of the jets as a $W$ boson. Hence, we can utilize the condition that one dijet mass should be close to $m_W$. Consequently, we ask for $(1 - f_W)\frac{m_W}{m_t} < \frac{m_{ij}}{m_{123}} < (1 + f_W)\frac{m_W}{m_t}$, which translates to certain restrictions in the mass plane (default $f_W = 0.15$). A study of signal and background processes leads to two additional cuts which allow to reduce QCD background that typically clusters at low $m_{23}$ and low $\arctan\left(\frac{m_{23}}{m_{12}}\right)$. For $m_{23} \approx m_W$ we require $0.2 < \arctan\left(\frac{m_{23}}{m_{12}}\right) < 1.3$. In all other cases we request $\frac{m_{23}}{m_{123}} > 0.35$. Assuming again mass less subjets we can give the corresponding mass plane cuts with $R_{\text{max,min}} = (1 \pm f_W)\, m_t/m_W$

$$R_{\min} < \frac{m_{23}}{m_{123}} < R_{\max} \quad \text{and} \quad 0.2 < \arctan\left(\frac{m_{13}}{m_{12}}\right) < 1.3 \tag{2.59}$$

$$1 - R_{\min}^2\left(1 + \left(\frac{m_{13}}{m_{12}}\right)^2\right) < \left(\frac{m_{23}}{m_{123}}\right)^2 < 1 - R_{\max}^2\left(1 + \left(\frac{m_{13}}{m_{12}}\right)^2\right) \quad \text{and} \quad \frac{m_{23}}{m_{123}} > 0.35 \tag{2.60}$$

$$1 - R_{\min}^2\left(1 + \left(\frac{m_{12}}{m_{13}}\right)^2\right) < \left(\frac{m_{23}}{m_{123}}\right)^2 < 1 - R_{\max}^2\left(1 + \left(\frac{m_{12}}{m_{13}}\right)^2\right) \quad \text{and} \quad \frac{m_{23}}{m_{123}} > 0.35\,. \tag{2.61}$$

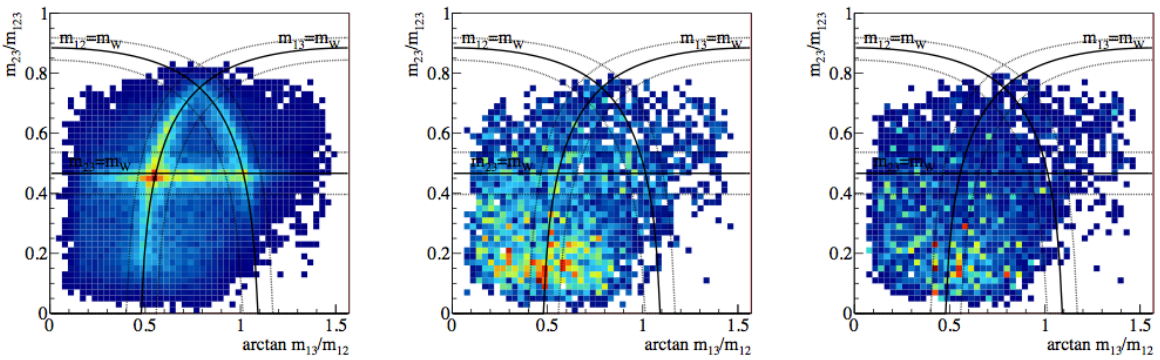


Figure 2.8.: Mass planes before mass plane cuts for $t\bar{t}$ (left), $W$ + jets (center) and pure QCD (right) at $\sqrt{s} = 14$ TeV. Taken from [4].

5. $p_T$**-cut**: Finally, the top candidate constructed from the three jets must have $p_T > 200$ GeV. We point out that this condition will be removed once we are looking for softer fat jets and tops.

The HEPTopTagger relying on jet substructure analyses has been proven to be a useful tool *e.g.* for searches for heavy resonances decaying to top quarks [42]. For more information on top tagging and different taggers see [43].

Another method to identify hadronically decaying top quarks is the bucket technique [5]. It is based on a simple sorting technique and targets mainly the low transverse momentum regime between 150 and 300 GeV. For this discussion we consider a hadronically decaying top pair as signal process. The algorithm starts by distributing the jets of an event to three buckets. Two of them will finally correspond to a top quark ($B_1$, $B_2$); the third one contains initial state radiation $B_{\text{ISR}}$. We seed the two top buckets with $b$ tagged jets and distribute the left (un-tagged) jets to the three buckets. To find the correct assignment we use the difference between the top mass and the combined mass of all jets in each top bucket

$$\Delta_{B_i} = |m_{B_i} - m_t| \qquad \text{with} \quad m_{B_i}^2 = \left( \sum_{k \in B_i} p_k \right)^2 . \tag{2.62}$$

For each possible assignment we calculate

$$\Delta^2 = \omega \Delta_{B_1}^2 + \Delta_{B_2}^2 , \tag{2.63}$$

where $\omega > 1$. For high values of $\omega$ the bucket $B_1$ decouples and will contain the best top candidate. The assignment that minimizes (2.63) is considered further. We allow top buckets with two, three or more jets including the seeded b jet. Buckets containing more than three jets are reduced to the expected three jet topology by reclustering using the Cambridge/Aachen algorithm. Buckets with three jets after the possible reclustering can be subjected to cuts on the ratio of the two non-$b$ jet invariant mass $m_{kl}$ and the bucket mass similar to the width of the A lines in the HEPTopTagger mass plane cuts

$$\left| \frac{m_{kl}}{m_{B_i}} - \frac{m_W}{m_t} \right| < 0.15 . \tag{2.64}$$

Buckets containing just one additional jet ($b/j$ buckets) most likely miss the softer of the two $W$ decay jets. For signal events this gives on parton level rise to a mass distribution with sharp endpoint at $m_{b/j} = \sqrt{m_t^2 - m_W^2}$ as shown in Fig. 2.9. As supposed in the original bucket paper one can use the $b/j$ buckets for top identification by modifying their distance measure

$$\Delta_B^{bj} = \begin{cases} |m_B - 145 \text{ GeV}| & \text{if } m_B \leq 155 \text{ GeV} \\ \infty & \text{else} \end{cases} . \tag{2.65}$$

Obviously, the quality of tops reconstructed from such b/j buckets is limited. However, if one is primary interested in tagging a top, they improve the efficiency. After finding the optimal assignment we can classify events in three categories:

- ($\mathfrak{t}_w$,$\mathfrak{t}_w$): both top buckets have $W$ candidates fulfilling Eq.(4.7),
- ($\mathfrak{t}_w$,$\mathfrak{t}_-$) or ($\mathfrak{t}_-$,$\mathfrak{t}_w$): only one top bucket has a $W$ candidate,
- ($\mathfrak{t}_-$,$\mathfrak{t}_-$): neither top bucket has a $W$ candidate.

We note that the classification of a top bucket as $\mathfrak{t}_w$ is only possible if there are at least two non-$b$ jets in this bucket. For finally claiming a bucket to correspond to a top quark it has to pass the following the mass window cuts

$$\mathfrak{t}_- : \quad 75 \text{ GeV} < m_{bj} < 155 \text{ GeV} \tag{2.66}$$

$$\mathfrak{t}_w : \quad 155 \text{ GeV} < m_B < 200 \text{ GeV} . \tag{2.67}$$

In Section 4.2 we will use the bucket algorithm in a modified form to analyze the fully hadronic decay channel of $t\bar{t}H$.

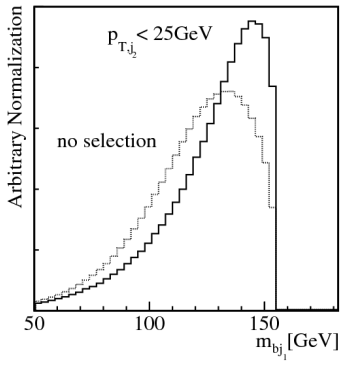


Figure 2.9.: Distribution of invariant masses obtained by adding the $b$ quark and the harder $W$ decay jet. The shown curves correspond to the parton level truth with (solid) and without (dotted) requiring the non $b$ jet to fulfill $p_{T,j_2} > 25$ GeV. Taken from [5].

## 2.7. Fox-Wolfram moments

A tool that will allow us to analyze correlations of subjets in a systematic way are Fox-Wolfram moments (FWM). They have been suggested first for studies of QCD final states in electro–positron collisions [44] in the late 1970s. Nowadays they have been proven to be useful in Higgs studies [45]. Fox-Wolfram moments are superpositions of two spherical harmonics $Y_l^m(\theta, \phi)$, where $\theta$ and $\phi$ are the usual angles in spherical coordinates. The Fox-Wolfram moments built from $N$ objects were originally defined as

$$H_l = \frac{4\pi}{2l+1} \sum_{m=-l}^{l} \left| \sum_{i=1}^{N} \frac{|\vec{p}_i|}{\sqrt{s}} Y_l^m(\theta_i, \phi_i) \right|^2 . \tag{2.68}$$

Using the properties of spherical harmonics, this can be simplified to

$$H_l = \sum_{i,j=1}^{N} \frac{|\vec{p}_i|}{\sqrt{s}} \frac{|\vec{p}_j|}{\sqrt{s}} \frac{4\pi}{2l+1} \tag{2.69}$$

$$= \frac{|\vec{p}_i||\vec{p}_j|}{s} P_l(\cos\Omega_{ij}) \tag{2.70}$$

with the angular distance measure $\cos(\Omega_{ij}) = \cos\theta_i \cos\theta_j + \sin\theta_i \sin\theta_j \cos(\phi_i - \phi_j)$ and the Legendre polynomial of order $l$ $P_l(\cos\Omega_{ij})$. The Legendre polynomials form a complete system of functions over the interval $[0, 1]$ and are solutions to the Legendre differential equation

$$(1-x^2)f''(x) - 2xf'(x) + l(l+1)f(x) = 0 . \tag{2.71}$$

For numerical implementation one can utilize the recursion relation for $l \geq 2$

$$lP_l(x) = (2l-1)xP_{l-1}(x) - (l-1)P_{l-2}(x) \tag{2.72}$$

starting from $P_0(x) = 0$ and $P_1(x) = x$. For our simulations we use the Legendre polynomial subroutines provide in GSL [46], which are based exactly on this approach. In Fig. 2.10 we show the first Legendre polynomials. One can generalize the Fox-Wolfram moments to other weight factors $W_{ij}^x = W_i^x W_j^x$ such that

$$H_l^x = \sum_{i,j=1}^{N} W_{ij}^x P_l(\cos\Omega_{ij}) . \tag{2.73}$$

Possible weight factors are for example

$$W_{ij}^T = \frac{p_{Ti} p_{Tj}}{\left(\sum_k p_{Tk}\right)^2} , \qquad W_{ij}^p = \frac{|\vec{p}_i||\vec{p}_j|}{\left(\sum_k |\vec{p}_k|\right)^2} , \qquad W_{ij}^u = \frac{1}{N^2} . \tag{2.74}$$

In order to understand the meaning of Fox-Wolfram moments it is instructive to study the case of two jets. In this case we can rewrite the Fox-Wolfram moments as follows:

$$H_l = \sum_{i,j=1}^{2} W_i W_j P_l(\cos\Omega_{ij}) \tag{2.75}$$

$$= W_1^2 + W_2^2 + 2W_1 W_2 P_l(\cos\Omega_{ij}) , \tag{2.76}$$

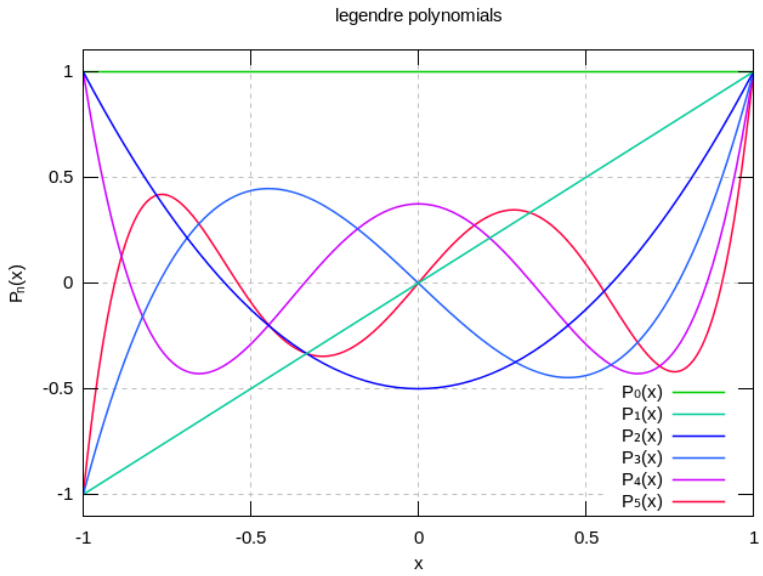


Figure 2.10.: The first Legendre polynomials. Taken from [47].

where we used $P_l(1) = 1$. We assume $W_1 \geq W_2$ and introduce $r = \frac{W_2}{W_1}$ yielding

$$H_l(r, \Omega_{ij}) = \frac{1 + 2rP_l(\cos\Omega_{ij}) + r^2}{1 + 2r + r^2} \,. \tag{2.77}$$

In Fig. 2.11 we present the dependence of the first six nontrivial Fox-Wolfram moments on $r$ and $\Omega_{ij}$. Obviously, the shape of those moments is determined by the involved Legendre polynomials. For even $l$ the maximal value at fixed $r$ is reached if the two input objects are parallel or anti-parallel. In the case of odd $l$ the antiparallel configuration results in the global minimum. Furthermore, we notice that the heights of the inner extrema are at least for small $l$ similar to each other and much less than the amplitude at the boundary. This will in general translate into 3 regions. In case of odd $l$ high values of $H_l$ correspond to low $\Omega_{ij}$ and $H_l$ will not give any values between 0 and a certain threshold; while moments with even $l$ are symmetric under $\Omega_{ij} \to \pi - \Omega_{ij}$. The most interesting part however is the region in between where we will get values based on the inner oscillations mapping out structures of the angular distribution. We notice that $H_l \geq 0$ which can be directly seen for the definition in terms of spherical harmonics in (2.68). Further it holds

$$H = \sum_{i,j=1}^{N} W_i W_j P_l(\cos(\Omega_{ij})) \leq \sum_{i,j=1}^{N} P_l \cos(\Omega_{ij}) \leq N^2. \tag{2.78}$$

In particular, for the normalized unit weight we find $0 \leq H_l \leq 1$.

In thesis we will often study Fox Wolfram moments constructed from three objects lying in a plane. Since this implies just two independent angles we can plot them as done in Fig. 2.12.

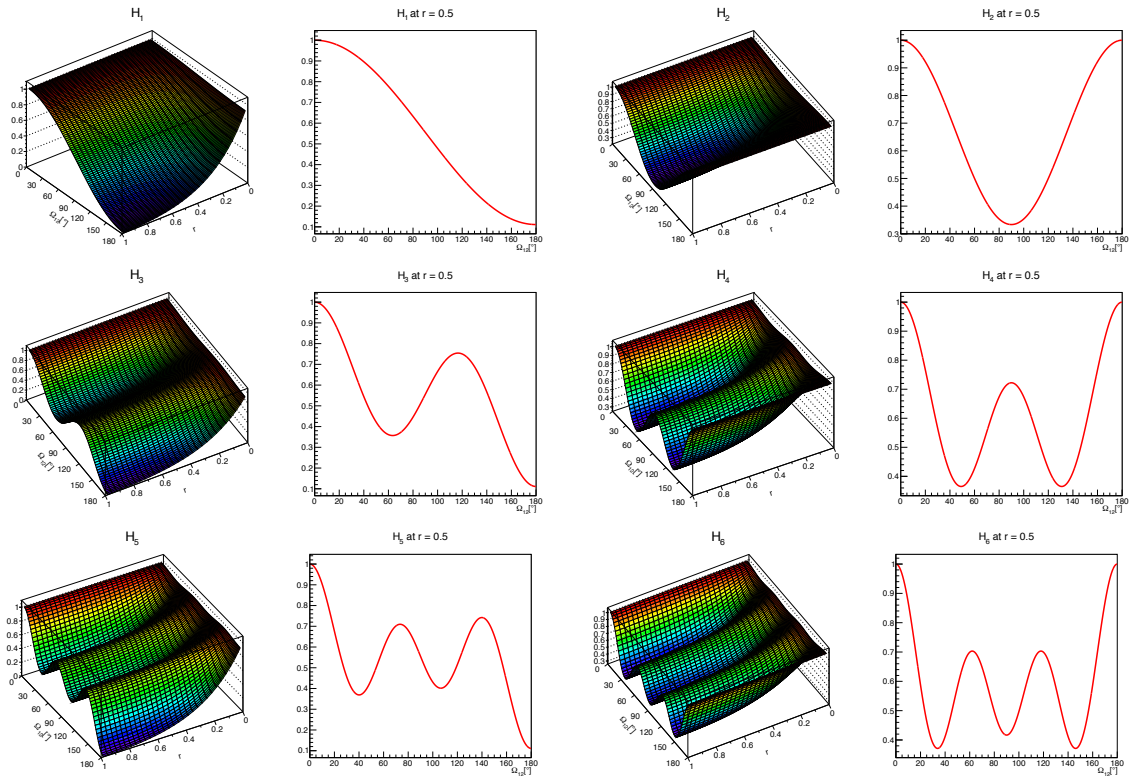


Figure 2.11.: First 6 nontrivial disubjet Fox-Wolfram moments. Taken from [45].

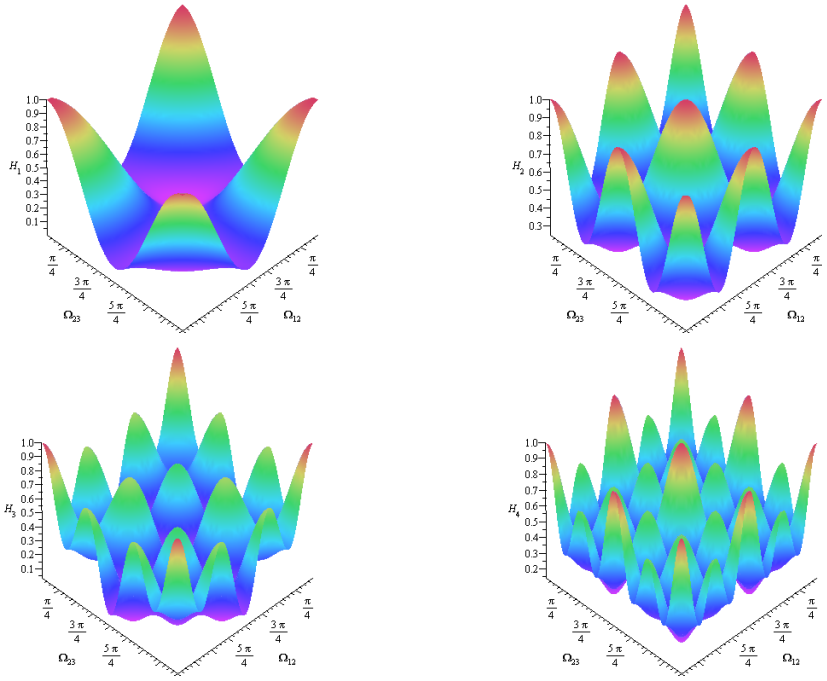


Figure 2.12.: First 4 nontrivial Fox-Wolfram moments constructed from three objects lying in a plane.

## 2.8. Simulated annealing

One of the main issues in particle physics is to distinguish a signal from the background. Of course there are clear signatures like for example a muon pair combining to a resonance, but in general this task is more difficult. For each event one has a list of (often correlated) variables from which one has to conclude whether a single event belongs to the signal or not. Such classification tasks can be treated using computer algorithms analyzing a set of variables simultaneously (multivariate analyses). In this and the following section we will introduce some basic ideas of two of those analyses. On the one hand we will use a cut-based approach namely simulated annealing; on the other hand we will present a classification in the form of a sequence of decisions giving rise to a decision tree.

Simulated annealing [48] (SA) is aimed to find the global minimum of a function. It is mimicking the physical process of annealing of a metal as illustrated in Fig. 2.13. Starting at a high temperature a sudden cooling will result in the system ending in an unwanted local minimum. If the cooling is performed more slowly instead, there is always the possibility of reasonable thermal fluctuations perturbing the system and therefore prevent the system sticking in a local minimum. Finally, this will lead at least to a better if not the global minimum.

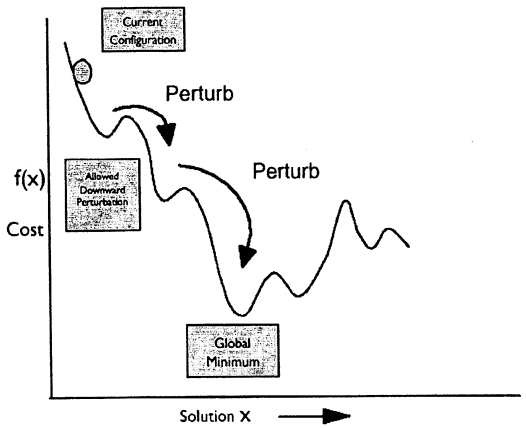


Figure 2.13.: Illustration of the principle of Simulated Annealing. Taken from [48].

The task is to find the global minimum of a function $f : D \to \mathbb{R}$. For simplicity we restrict ourselves in this section to the case of $D \subset \mathbb{R}$; the generalization to higher dimensions is obvious. The algorithm starts with a randomly chosen $x \in D$ at time step $t = 0$ and number of tries $i = 1$. Let $(T_t)_{t\in\mathbb{N}}$ be a series of temperatures and N be the maximal number of tries per time step.

- For $i \leq N$ pick a random number $y$ in the neighborhood $U(x)$ of the point $x$.
- Evaluate the function $f$ at the new point $y$. If $f(y) \leq f(x)$ jump to the point y. If $f(y) \leq f(x_{\min})$ set $x_{\min} = y$. Otherwise we jump to $y$ with a probability given by the

Boltzmann factor

$$\exp\left(-\frac{f(y)-f(x)}{T_t}\right) . \tag{2.79}$$

- If there was no jump to $y$ and $i < N$ increase $i$ and iterate. For the case that one moved from $x$ to $y$ or $i$ reached $N$ increase $t$ and reset $i = 1$.

This procedure is iterated until a stop criterion, *e.g.* $t = t_{\max}$ is fulfilled. The found minimum is $f(x_{\min})$. There are many variants of the algorithm sketched above. One of them implemented in the TMVA framework is discussed in Appendix A.

## 2.9. Boosted decision trees

Another optimization method are decision trees. A decision tree consists of a sequence of yes/no decision where the following question depends on the former decision. In Fig. 2.14 we show a simple example. Starting from the root node we go either left or right depending on the cut on a certain variable and reach the next node. Nodes without further decisions are called leaves. In general, such a decision tree is superior to rectangular cuts since it allows *e.g.* for a situation in which the signal region does not correspond to a hypercube in parameter space but is spread over several regions. Below we will sketch a quite instructive possibility how one can build a decision tree. Some more details on the actual way boosted decision trees are created in TMVA can be found in Appendix A. For this more pedagogical section we follow the conventions and arguments from [49].

Suppose we have a set $S$ of $|S|$ events belonging to certain classes $C_j$. The probability to pick a event in $S$ belonging to the class $C_j$ is given by

$$\frac{\text{freq}(C_j, S)}{|S|} . \tag{2.80}$$

The information of an event is

$$-\log_2 \frac{\text{freq}(C_j, S)}{|S|} . \tag{2.81}$$

Therefore, the expected information can be calculated to

$$\text{info}(S) = \sum_j \frac{\text{freq}(C_j, S)}{|S|} \left[-\log_2 \frac{\text{freq}(C_j, S)}{|S|}\right] . \tag{2.82}$$

For a training set $T$ info$(T)$ can be interpreted as the average amount of information that is needed to identify the class of an event in $T$. Usually, we call info$(T)$ the entropy. Let us assume now that we have a certain test $X$ which allows us to split the events in $T$ to $n$ categories $T_i$. In this case

$$\text{info}_X = \sum_j \frac{|T_j|}{|T|}\text{info}(T_j) . \tag{2.83}$$

We can quantify the gain of information using the test $X$ as

$$\text{gain}(X) = \text{info}(T) - \text{info}_X(T) , \tag{2.84}$$

which in principle would allow us to choose the most efficient test. However, this gain criterion is biased in favor of tests with many categories. If $T$ would lead to categories that all contain just one single event, $\text{info}_X(T) = 0$ and in consequence the gain is maximal. To avoid this problem we define the split information

$$\text{split info}(X) = -\sum_j \frac{|T_i|}{|T|} \log_2 \frac{|T_i|}{|T|} \tag{2.85}$$

giving the potential information by dividing $T$ into $n$ subsets. With this at hand one can define a new measure to quantify the use of a test $X$, the gain ratio

$$\text{gain ratio}(X) = \frac{\text{gain(X)}}{\text{split info}(X)} \,. \tag{2.86}$$

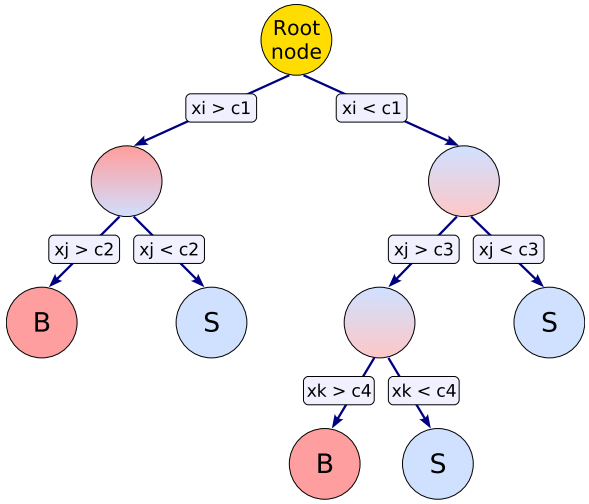


Figure 2.14.: Example of a Decision Tree. Taken from [50].

Once we have built a tree we can try to simplify it. This can be done by replacing several branches or subbranches with single leaves (pruning). In general, this will lead to a higher misclassification rate but will result in much simpler trees.

To increase the stability of a constructed decision tree with respect to fluctuations in the training sample one can build an entire forest of trees by reweighing the training sample and using a majority vote. Those procedures are called boosting. Here we will try to give a short impression of an algorithm called AdaBoost [51]. In this adaptive boost algorithm one increases the weight of events that are misclassified using one decision tree for the training of the next one. This is done by multiplying the original weights of all misclassified events by an overall boost weight

$$\alpha = \frac{1 - \text{err}}{\text{err}} \,, \tag{2.87}$$

where err is the misclassification rate of the last tree. Finally, one rescales all weights such that there is no change in the total sum of weights. If the result of an individual classifier

is defined as $h(\vec{x}) = \pm 1$, where the upper (lower) sign corresponds to a signal (background) classification and $\vec{x}$ is the vector of input parameters, we define the boosted event classification for $N$ classifiers

$$y_{\text{boost}}(x) = \frac{1}{N} \sum_{i}^{N} \log(\alpha_i) h_i(\vec{x}) \; . \tag{2.88}$$

Depending on this weighted average we classify an event either as signal or background.

The methods lined out in this and the last section will help us to look for possible improvements of the HEPTopTagger. While boosted decision trees are in principle more powerful, simulated annealing allows to give a set of cuts to classify the events. This easy way to see what is going on is lost for boosted decision trees once we have to deal with an entire forest of trees.

# 3. HEPTopTagger Analyses

The HEPTopTagger has been proven to be useful in several experimental studies at the LHC with a center of mass energy of 7 and 8 TeV. After the recent shutdown the LHC will be upgraded to operate at 13 TeV. Therefore, it is the right time to think about possible improvements and making the tagger ready for the upcoming data. In the following sections we will analyze a different sequence of tagging steps as well as illustrate the possible performance beyond the default operating point in terms of ROC curves. Furthermore, we investigate whether the tagger benefits from using angular correlations between the subjets. Finally, we test how we can extend the HEPTopTagger towards lower transverse momenta.

The analyses in this section were done in collaboration with Catherine Bernaciak and Tilman Plehn and follow [7].

## 3.1. A new default

In this section we study two possible modifications of the HEPTopTagger [4] as it is described in Section 2.6. Although both changes in the algorithm will not influence the tagging performance in a significant way, they will affect the distribution of reconstructed masses $m_{t,\text{rec}}$ from QCD multi–jet background. For this study we use a semi–leptonic $t\bar{t}$ sample including up to 2 hard jets on matrix element level generated with Alpgen [52] and Pythia [53] using MLM [27] merging. Our background sample consists of leptonic $W$ bosons with 2 to 4 additional hard jets. On generator level we require all hard jets to fulfill $p_{T,j} > 25$ GeV, $|\eta_j| < 5$, and $\Delta R_{jj} > 0.4$. The merging parameters are $E_{T,\text{clust}} = 30$ GeV, $R_{\text{clust}} = 0.4$, and $|\eta_{\text{clust}}|_{\text{max}} = 5$. To avoid complications from hadronically decaying $\tau$ leptons we limit our study to the leptonically decay $\ell = e, \mu$. Our studies start from the latest LHC center of mass energy 8 TeV and will then move to the next LHC run at $\sqrt{s} = 13$ TeV. The signal cross sections are rescaled according to the NNLO results $\sigma_{t\bar{t}}^{8\text{ TeV}} = 245$ pb [54] and $\sigma_{t\bar{t}}^{13\text{ TeV}} = 806$ pb [55]. All analyses in this and the sequent sections are based on fat jets with $R_{\text{fat}} = 1.8$ constructed using the Cambridge/Aachen algorithm [34] implemented in Fastjet [36]. Increasing the cone size to $R_{\text{fat}} = 1.8$ will increase the tagging efficiency for moderately boosted tops but might need an adjustment of filtering parameters based on experimental results. For this section we study fat jets $p_{T,\text{fat}} > 200$ GeV and $|\eta_{\text{fat}}| < 2.5$. According to the consistency cut in Section 2.6 we require the tagged objects to obey $p_{T,\text{rec}} > 200$ GeV. Likewise we use the default parameters for the HepTopTagger as summarized in Tab. 3.1: To extract the massive splitting we use the mass drop criterion $\min(m_{j_1}, m_{j_2})/m_j < f_{\text{drop}} = 0.8$. The search for hard substructures ends at subjets with $m_{\text{sub}} = 30$ GeV. The filtering step keeps the $N_{\text{filt}} = 5$ hardest objects. The reconstructed $W$ to top mass ratio is allowed to vary 15% from the true value. The additional mass plane cuts explained in Section 2.6 are $0.2 < \arctan m_{13}/m_{12} < 1.3$ and $m_{23}/m_{123} > 0.35$. Finally, the reconstructed top mass $m_{T,\text{rec}}$ has to lie between 150 and 200 GeV.

| | | | |
|---|---|---|---|
| $R_{\text{fat}}$ | 1.8 | $(m_W/m_t)^{\text{rec}}/(m_W/m_t)$ | 0.85-1.15 |
| $f_{\text{drop}}$ | 0.8 | $m_{23}/m_{123}$ and $m_{13}/m_{123}$ cuts | 0.35, 0.2, 1.3 |
| $m_{\text{sub}}$ [GeV] | 30 | $m_t^{\text{rec}}$ [GeV] | 150-200 |
| $N_{\text{filt}}$ | 5 | | |

Table 3.1.: Parameters in the HEPTopTagger algorithm, as defined in the text.

The first modification we will study in this section concerns the order of cuts in the HEPTopTagger algorithm. In the standard version introduced in Section 2.6 we pick the filtered triple of subjets that recombines best with respect to the top mass. Later on we check whether this triple passes one of the mass plane cuts of Eqs. (2.61). We now modify the order of cuts by first rejecting all triples that do not pass the mass plane cuts and picking the filtered triple with the best value of $m_{t,\text{rec}}$ among those. The benefit of this procedure is that we can still get tags if the selected triple in the old version fails the mass plane cuts. In the original tagging algorithm this configuration definitely rejected the fat jet without testing other triples. It is immediate that this new ordering will increase the number of tags both in the signal and background sample. In Fig. 3.1 we compare the distribution of reconstructed top masses $m_{t,\text{rec}}$ for the new ordering and the old standard tagger. While the shape in the signal sample is unchanged we note a less shaped background distribution. The tagging efficiency and mis-tagging rates for both approaches are given in Tab. 3.2. We define the signal efficiencies $\epsilon_S$ as the number of tagged tops divided by the number of hadronic tops in the event sample with $p_{T,t} > 200$ GeV and $|\eta_t| < 2.5$. The mis-tagging rate or background efficiency $\epsilon_B$ gives the number of tags in the leptonic $W$+jets sample relative to the number of fat jets with $p_{T,\text{fat}} > 200$ GeV and $|\eta_{\text{fat}}| < 2.5$. As expected, the new ordering increases both $\epsilon_S$ and $\epsilon_B$. Because of the increased signal efficiency and the less shaped background we will use the new ordering as new default despite the fact that there is no improvement of the signal–to–background ratio.

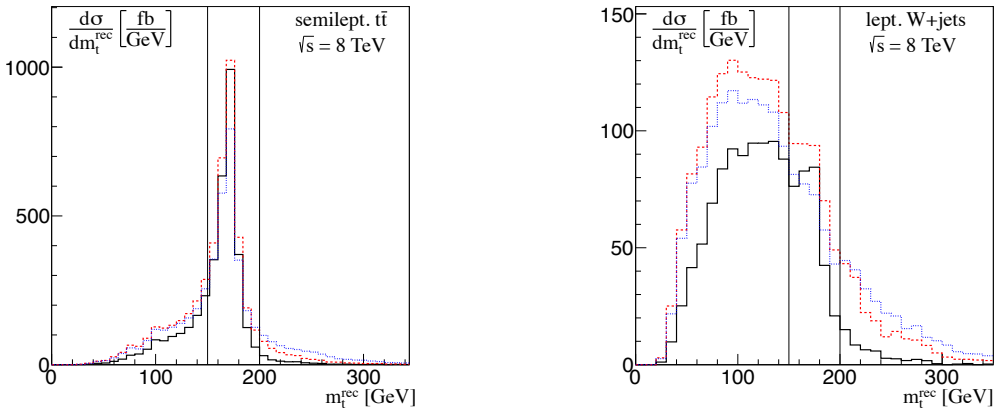


Figure 3.1.: Reconstructed top mass for the signal (left) and background (right) for 8 TeV collider energy. We show the standard order of cuts with $|m_{123} - m_t|$ selection (solid black), the inverted order of cuts with $|m_{123} - m_t|$ selection (dashed red), and the inverted order of cuts with $dj_{\text{sum}}$ selection (dotted blue).

| | | standard | inverted | |
|---|---|---|---|---|
| | | $\lvert m_{123} - m_t \rvert$ | $\lvert m_{123} - m_t \rvert$ | $dj_{\text{sum}}$ |
| 8 TeV | $\epsilon_S$ | 0.331 | 0.375 | 0.304 |
| | $\epsilon_B$ | 0.014 | 0.018 | 0.014 |
| 13 TeV | $\epsilon_S$ | 0.337 | 0.394 | 0.305 |
| | $\epsilon_B$ | 0.015 | 0.021 | 0.016 |
| | | old default | new default | |

Table 3.2.: (Mis)tagging efficiencies for standard and inverted cut order, in the latter case for the $|m_{123} - m_t|$ and $dj_{\text{sum}}$ selections. The new HEPTopTagger default setting is indicated.

In a situation with high multiplicity for example if one has to deal with a multi-jet final state and pile-up, there will be several possible triples to construct a top candidate. In such cases using the difference to the true top mass $|m_{123} - m_t|$ to choose the triple which one analyzes further will shape the backgrounds. To avoid this shaping we introduce an alternative measure in form of a modified Jade distance. While the Jade distance [56] can be approximately linked to the disubjet mass

$$m_{ij}^2 \simeq E_i E_j \Omega_{ij}^2 \simeq p_{T,i} p_{T,j} \, (\Delta R_{ij})^2 \, , \tag{3.1}$$

we increase the importance of the angular distance $\Delta R_{ij}$ by doubling its power. To guarantee a unique value of the new measure we sum over all modified jade distances in the triple of subjets

$$dj_{\text{sum}} = \sum_{(ij)} d_{ij} \quad \text{with} \quad d_{ij} = p_{T,i} p_{T,j} \, (\Delta R_{ij})^4 \, . \tag{3.2}$$

The set of three subjets that maximizes $dj_{\text{sum}}$ is our choice for the rest of the tagging procedure. In this way we prefer subjet combinations that are separated more widely. A comparison of the obtained distributions for the reconstructed masses with both the old and our new default tagger is provided in Fig. 3.1. In the signal as well as in the background sample the amount of candidates in the mass window $[150, 200]$ GeV is reduced in comparison to our new default version. Moreover, the $m_{T,\text{rec}}$ distribution from the background sample is even less shaped. The actual (mis)tagging rates are given in the last column of Tab. 3.2. We note that both signal efficiency and background rejection are worse than for the old default tagger. However, this new measure might still be useful for high multiplicity final state analyses.

To test the performance of the HEPTopTagger in a conclusive way it cannot be sufficient to study single working points as done above. One needs instead the dependence of the achievable background rejection $r_B = 1 - \epsilon_B$ as function of the signal efficiency $\epsilon_S$. Such a curve is of particular interest if one wants to study improvements possible by using additional cuts since they always will not only affect the background rejection but the signal efficiency as well. This dependence $r_B(\epsilon_S)$ is described by a receiver operating characteristics (ROC) curve. Any point on this curve corresponds to an optimized setting that can be used for an analysis. The ROC curves we show in this section are obtained by using boosted decision trees implemented in the TMVA [50] framework for ROOT [57]. To avoid rerunning the entire tagging procedure each time, we restrict ourselves to use the mass window as well as tighter restrictions on $(m_W/m_t)^{\text{rec}}/(m_W/m_t)$, $m_{23}/m_{123}$, and $\arctan(m_{13}/m_{23})$ in all our optimizations in this and all further discussions.

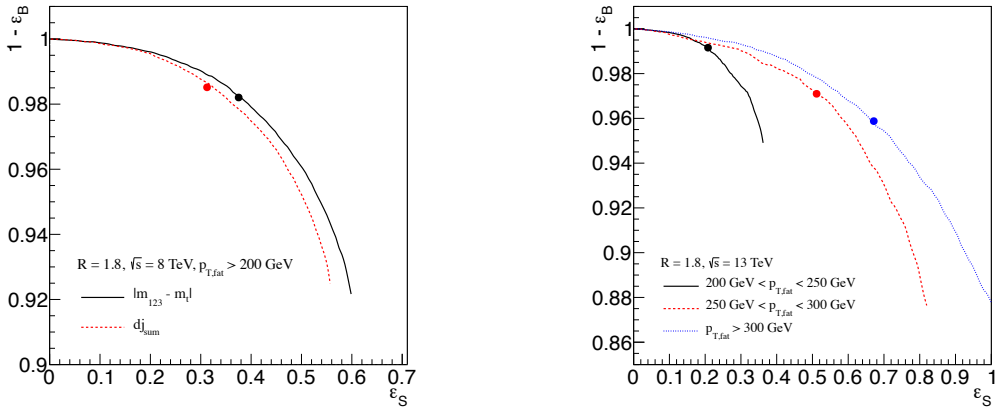


Figure 3.2.: ROC curves for the modified HEPTopTagger on semileptonic $t\bar{t}$ pairs for 8 TeV (left) and 13 TeV (right) collider energy. The standard working point from Tab. 3.1 is indicated by a dot. For 8 TeV we show the new default setup and the high-multiplicity modification. For 13 TeV we quote the performance in slices of $p_{T,\text{fat}}$.

The left panel of Fig. 3.2 shows ROC curves for the HEPTopTagger algorithm with inverted order of cuts for 8 TeV center of mass energy. We compare our new default tagger and the modified Jade distance version based on (3.2). The efficiencies at the default working points stated in Tab. 3.2 are indicated by a dot. The reason that the curves stop at a certain signal efficiency is that not all of the tops with $p_{T,t} > 200$ GeV and $|\eta_t| < 2.5$ give rise to fat jets with of $R_{\text{fat}} = 1.8$ with $p_{T,\text{fat}} > 200$ GeV and $|\eta_{\text{fat}}| < 2.5$.

In the right panel of Fig. 3.2 we study our new default tagger for a collider energy of 13 TeV. We show the ROC curves for three transverse momentum slices of the fat jets, namely $p_{T,\text{fat}} = 200 - 250$ GeV, $p_{T,\text{fat}} = 250 - 300$ GeV, and $p_{T,\text{fat}} = 300$ GeV. We do not study higher transverse momenta since these will be statistically limited in the signal sample and are not of particular interest for us. The signal efficiency in the lowest $p_T$ bin is limited because of the low boost resulting in top decay products not captured by $R_{\text{fat}}$. The performance of the new default tagger at the working point from Tab. 3.1 is indicated by a dot.

## 3.2. Angular correlations

This section aims for a systematic study of possible improvements of the HEPTopTagger by using angular correlations between the subjets. Our ansatz is to analyze these angular correlations in terms of Fox-Wolfram moments introduced in Section 2.7

$$H_l^x = \sum_{i,j=1}^{N} W_{ij}^x \, P_l(\cos\Omega_{ij}) \qquad \text{with} \quad W_{ij}^x = W_i^x W_j^x \; . \tag{3.3}$$

where we for this analysis focussing on the angular correlations use the normalized unit weight

$$W_{ij} = \frac{1}{N^2} \; . \tag{3.4}$$

To be less sensitive to the kinematics of the top quark before the decay we calculate Fox–Wolfram moments of the three subjets after boosting in the rest frame of the reconstructed

top. However, to loose not too much information we include the boost vector $a$ as fourth object.

To rank all possible Fox-Wolfram moments with respect to their power in discriminating signal and background we use their significance

$$s = \frac{1}{2}\int_{-\infty}^{+\infty} \frac{(S(x) - B(x))^2}{S(x) + B(x)} dx \to \frac{1}{2}\sum_{\text{bins}} \frac{(S(b) - B(b))^2}{S(b) + B(b)} \Delta_b \tag{3.5}$$

as provided in TMVA. For our analysis we label the two subjets which reconstruct $m_W$ best as $W_1$ and $W_2$, ordered by transverse momentum. The remaining subjet is then labeled the $b$-jet.

In Fig. 3.3 we show ROC curves for the full event sample and for slices in the transverse momentum of the fat jet. The purely QCD-inspired selection criteria of Tab. 3.1 is contrasted with the selection including angular correlations. The moment with the strongest separation of signal and background is the first moment $H_1$ built from the two $W$ decay jets. However, we see that it does not visibly improve the signal efficiency or background rejection of the tagger. In addition, we show results including the leading moments of the four most decisive combinations of input objects. They are usually low moments with $l = 1, 2, 3$. Again, no visible improvement appears. We conclude therefore that adding angular correlations does not lead to significant improvements of the tagger.

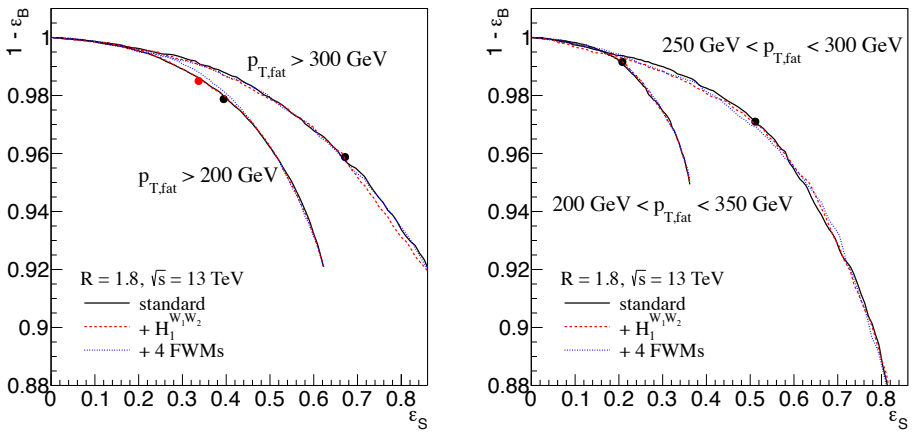


Figure 3.3.: ROC curves for the modified HEPTOPTAGGER including angular correlations via unit–weight Fox–Wolfram moments in slices of $p_{T,\text{fat}}$. The standard working point from Tab. 3.1 is indicated by a dot. We assume a collider energy of 13 TeV.

## 3.3. Moderate boost

As seen in the last section adding angular correlations does not improve our tagger significantly. However, this must not be true once we are interested in lower transverse momenta. While for $p_{T,t} > 200$ GeV in most cases all decay products end up in the fat jet the situation is different for $p_{T,t} = 150 - 200$ GeV. This $p_T$ range is interesting because of the transverse momentum distribution of top quarks. For example in $pp \to t\bar{t}H$ at 13 TeV reducing the $p_T$ bound on the top quarks from 200 to 150 GeV doubles the number of accessible events. In this regime we have many fat jets missing the softest decay jet and including a jet from initial state radiation instead. In such cases the fat jet still contains most information on the top and can be used to identify a top quark. However, the standard cuts in the tagging algorithm will not be sufficient to separate signal and background anymore. However, angular correlations of the subjets might help.

Our starting point for this analysis are again fat Cambridge–Aachen jets with $R_{\text{fat}} = 1.8$ and $|\eta_{\text{fat}}| < 2.5$. To have accesses to the lower $p_T$ regime we lower the minimal transverse momentum for the fat jets and the consistency cut to $p_{T,\text{fat}} > 150$ GeV and $p_{T,\text{rec}} > 150$ GeV. In this study we use the semi–leptonic $t\bar{t}$ sample and the leptonic $W$ + jets background described in the previous sections for a collider energy of 13 TeV. The analysis of the fat jets follows our new default version of the HepTopTagger introduced in Section 3.1. Tags passing the old consistency cut $p_{T,\text{rec}} > 200$ GeV are left untouched. We determine the quality of the tagged tops with $p_{T,t}^{\text{rec}} = 150 - 200$ GeV based on their angular separation to the parton level objets in an event. Following [58], we calculate for each possible assignment of reconstructed subjets to parton level object

$$\Delta R^2_{\text{sum}} = \sum_{i=1}^{3} \Delta R^2(p_i^{\text{rec}}, p_{j_i}^{\text{parton}}) \,. \tag{3.6}$$

The assignment that minimizes this measure defines the best mapping between the reconstructed decay products and the Monte Carlo truth. Using this assignment we distinguish three tag types:

- Type 1: All three subjets of the tagged top quark correspond to the parton level top decay products.
- Type 2: The two hardest subjets are mapped to top decay products; the softest subjet as a different origin.
- Type 3: Else.

Tab. 3.3 gives the distribution of tag types in the signal sample. This sets our starting point. As explained above we want to increase the tagging efficiency in this regime of transverse momentum using tags of type 2 that fail the normal tagging procedure. For this purpose we analyze the angular correlations of the subjets similar to Section 3.2. To receive cuts we utilize simulated annealing provided in Tmva.

In Fig. 3.4 we give the obtained ROC curve for the signal efficiency $\epsilon_{t_2,u}$ for tagging type 2 candidates that were rejected by our default algorithm vs the rejection of fat jets from the $W$+jets background that were rejected beforehand. The optimization relies on the standard HEPTopTagger variables as restricted above and the leading Fox–Wolfram moment of most

| | default | | + additional tagging step | |
|---|---|---|---|---|
| | (mis)tags [fb] | fraction | (mis)tags [fb] | fraction |
| $t\bar{t}$ | 9305 | | 11430 | |
| type-1 | 5309 | 0.57% | 5967 | 0.52% |
| type-2 | 1283 | 0.14% | 1863 | 0.16% |
| type-3 | 2712 | 0.29% | 3601 | 0.32% |
| $W$+jets | 1200 | | 1663 | |
| $\epsilon_S$ | 0.287 | | 0.353 | |
| $\epsilon_B$ | 0.007 | | 0.010 | |

Table 3.3.: Comparison of tagging results with (left columns) and without (right columns) the additional step. All tags fulfill $p_{T,\text{fat}} > 150$ GeV as well as 150 GeV $< p_{T,\text{tag}} < 200$ GeV. The chosen working for the new tagging step is given in Tab. 3.3.

decisive combination of input objects. We pick a working point indicated by the dot. The corresponding cuts are given in Tab. 3.3. The result from running the tagger with the additional cuts for originally untagged candidates between 150 and 200 GeV transverse momentum are show in the right column of Tab. 3.3. The additional tagger step increases the signal efficiency from 29% to 35%. Since we reanalyze candidates that were rejected before, an increase of the mis-tagging rate is totally expected. The amount of additional mis-tags can be limited by choosing a suitable working point. The absolute amount of type 2 tags is increased. However, the number of type 3 tags is increased as well, such that the ratio of type 2 to type 3 is approximately constant. If one wants to use the additional tagging step, one should therefore check carefully whether the quality of reconstructed top quarks is sufficient and might introduce further cuts to decrease the amount of type 3 tags.

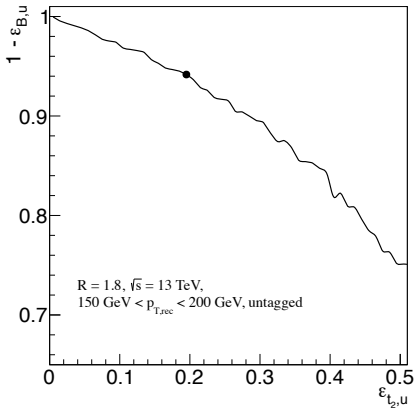


Figure 3.4.: ROC curve showing the efficiencies to identify type 2 candidates and the corresponding $W$ + jet background misidentification rate both for events that were rejected by the tagger without the additional tagging step in the transverse momentum bin $p_{T,\text{rec}} = 150 - 200$.

| variable | min | max |
|---|---|---|
| $m_t^{\mathrm{rec}}$ | 108. | 282. |
| $m_W^{rec}/m_t^{rec}$ | 0.717 | 1.556 |
| $\arctan(m_{13}/m_{12})$ | 0.441 | 0.889 |
| $m_{23}/m_{123}$ | 0.412 | 0.758 |
| $H_1^{W_1W_2}$ | 0.048 | 0.373 |
| $H_2^{aW_1W_2}$ | 0.019 | 0.524 |
| $H_2^{abW_1W_2}$ | 0.044 | 0.276 |
| $H_1^{abW_2}$ | 0.145 | 0.445 |

Table 3.4.: Cuts of the used working point as received by simulated annealing.

# 4. Buckets of Higgs and Tops

In this section we demonstrate that a study of the fully hadronic final state of $t\bar{t}H$ should be possible at the LHC. The analyses below were done in collaboration with Matthew Buckley, Tilman Plehn, and Michihisa Takeuchi and follow closely [6].

## 4.1. Multi-jet backgrounds and global cuts

In this analysis we try to tackle the fully hadronic final state of $t\bar{t}H$ where the Higgs decays to a $b\bar{b}$ pair. For a Standard Model Higgs boson with $m_H = 125$ GeV the corresponding branching ratio is given as $\text{BR}(H \to b\bar{b}) = 0.577$ [37, 3]. According to the hard process we will ask for events with four $b$-tagged jets, no leptons and at least two additional hard jets. The dominant background processes giving rise to such events are ordered by relevance $pp \to b\bar{b}b\bar{b}$, $pp \to t\bar{t}b\bar{b}$ and $t\bar{t}t\bar{t}$. If we consider a $b$ tagging efficiency of 70% with a mis-tagging rate of 1% all channels that fake multiple $b$s are suppressed. The rate that $b\bar{b}+\text{jets}$ contributes is on the level of 10% which is below the uncertainties of the considered backgrounds. The influence of pure QCD events with 4 fakes can be ignored.

The basis of our analyses are Monte Carlo event samples generated with Alpgen [52] and Madgraph [59], both with a Pythia parton shower [53, 60], as well as with Sherpa [61]. Our signal sample is generated with Sherpa and contains $t\bar{t}H$ with up to one additional hard jet merged using the implemented Ckkw scheme [26]. A comparison with a corresponding Madgraph simulation using Mlm [27] merging is given in Appendix B. The provided comparison shows that the sensitivity of our simulations with respect to QCD issues is under control. For the $t\bar{t}b\bar{b}$ background we simulate up to one additional hard jet with Sherpa and cross–check our results with Alpgen finding only negligible dependence on the simulation methods. We normalize our signal event sample to the NLO result of 504 fb [3, 62, 63]; the $t\bar{t}b\bar{b}$ background sample is normalized to 1037 fb after the generator cuts $p_{T,b} > 35$ GeV, $|\eta_b| < 2.5$, and $\Delta R_{bb} > 0.9$. Since the rate of $t\bar{t}t\bar{t}$ is small in comparison to the other background processes and has a cross section of at most 5% of the signal we do not study this background in detail.

For a study of hadronic $t\bar{t}H$ with the Higgs decaying to a $b\bar{b}$ pair the dominant background is the QCD production of two bottom anti-bottom pairs with additional jets corresponding to two $b$s + jets mimicking top decay products and the other $b\bar{b}$ pair faking the Higgs decay. This background overwhelms the signal with a cross section of 400 pb estimated with Alpgen after pre-selection cuts. Since this process is dominated by soft and untagged jets with additional enhancement from gluons splitting in $b\bar{b}$ pairs we can reduce it dramatically by asking for four hard well-separated $b$ tagged jets.

In the bucket reconstruction of two tops and a Higgs carried out below we will ask for at least two additional untagged hard jets. Therefore, our background is $b\bar{b}b\bar{b}jj$ plus parton

shower which we simulate with ALPGEN. The resulting cross section after pre-selection cuts reaches 2128 fb. To cross–check the stability of our analysis we compare our results with a $b\bar{b}b\bar{b}$ plus zero and one additional hard jet merged with SHERPA. The simulation of a merged sample with up to two additional jets is not done because of its prohibitive computational effort. Nevertheless, we can vary the used renormalization and factorization scales by a factor of 1/2 to 2 around its central value for which we use the partonic center of mass energy $\sqrt{\hat{s}}$ to check that our results are valid independent of these scale choices. A careful comparison of the different simulation methods and hard processes is provided in Appendix B. Therein we test several assumptions entering our simulations. First, we demonstrate that only background events with two or more hard additional jets have to be considered in our bucket analysis. As a conservative choice we estimate this background using ALPGEN $b\bar{b}b\bar{b}jj$. Further, we show that our analysis is not sensitive to the method by which the second jet is generated, *i.e.* as a hard jet using ALPGEN or by parton shower as done in our SHERPA simulation.

In this section we try to understand the kinematics of the four b quarks. The next one will show that several of these can be replaced by requiring the reconstruction of top quarks, which improves the purity of signal over background.

As we already stated above, the overwhelming rate of $b\bar{b}b\bar{b}$ raises the need of selection cuts before we can hope to perform any analysis. We select events with four $b$-tagged jets. These $b$-jets must be central and, to reduce the influence of gluon splittings $g \to b\bar{b}$, we require them to be well separated. The actual cuts are

$$p_{T,b} > 40 \text{ GeV}, \qquad |\eta_b| < 2.5, \qquad \Delta R_{bb} > 1.0 \qquad (4\times) \,. \tag{4.1}$$

In addition, we want at least two hard un-tagged jets fulfilling

$$p_{T,j} > 40 \text{ GeV}, \qquad |\eta_j| < 4.5, \qquad \Delta R_{jj} > 0.5 \qquad (2\times) \,. \tag{4.2}$$

These naive acceptance cuts are very inefficient, for example when compared to subjet methods. However, our aim is to show that the purely hadronic $t\bar{t}H$ process can be studied at the LHC, thus we need to ensure that the pure QCD backgrounds can be reliably removed. Moreover, four individual $b$-tags cannot be treated as statistically independent unless we at least assume very widely separated $b$-jets. This necessitates the cuts used in this proof–of–concept analysis.

The first line of Tab. 4.1 gives the cross sections for the signal and two primary backgrounds at the 13 TeV LHC, after acceptance cuts. After the selection cuts of (4.1) the $bb\bar{b}\bar{b}$+jets cross section significantly dominates the signal. Therefore, we need additional cuts to improve this. Later on we will ask for two reconstructed tops which will make those cuts unnecessary. But it is instructive to compare the later results to simple global cuts. We introduce two variables which will allow us to implement such global cuts. They both are effective masses: one calculated for the 4 $b$ tag jets and the other built from the total jet activity.

$$m_{\text{eff}} = \sum_{\text{all jets}} p_T \,, \qquad\qquad m_{\text{eff},b} = \sum_{\text{four } b\text{-jets}} p_T \,. \tag{4.3}$$

Both observables are sensitive to the kinematics of the multi-jet system. In Figure 4.1 we show the distributions for signal and backgrounds, relative to the event rates after the acceptance

| | $t\bar{t}H$ | $t\bar{t}b\bar{b}$ | $b\bar{b}b\bar{b}jj$ | $S/B$ |
|---|---|---|---|---|
| After acceptance Eqs.(4.1) and (4.2) | 1.197 | 8.363 | 54.420 | 0.019 |
| After global cuts Eq. (4.4) | 0.134 | 0.558 | 2.734 | 0.041 |
| Mass window $m_{bb} = 90 - 130$ GeV | | | | |
| closest | 0.096 | 0.299 | 1.577 | 0.051 |
| hard | 0.017 | 0.031 | 0.226 | 0.065 |
| soft | 0.060 | 0.173 | 0.893 | 0.056 |
| min | 0.071 | 0.246 | 1.143 | 0.051 |

Table 4.1.: Cross section (in fb) of signal and background events after successive selection cuts. The $b\bar{b}b\bar{b}jj$ rate is based on the ALPGEN simulation. After the full set of cuts from Eqs.(4.1), (4.2), and (4.4), we show several naive ways of selecting two $b$-jets to reconstruct the Higgs mass: the pair closest in invariant mass to the Higgs, the two hardest $b$-jets, the two softest $b$-jets, and the two $b$-jets with the minimum invariant mass. We assume a 70% $b$-tagging efficiency and neglect the small mis-tag backgrounds.

cuts of Eqs.(4.1) and (4.2). The corresponding signal–to–background ratio is around 1:50. The $m_{\text{eff}}$ distributions of the signal and the $b\bar{b}b\bar{b}$ background show a similar behavior since we require in our ALPGEN simulation two hard un-tagged jets as is discussed in Appendix B. From the right panel of Figure 4.1 we see that once we ask for $m_{\text{eff}} > 500$ GeV both $m_{\text{eff}}$ and $m_{\text{eff},b}$ are of similar shape for signal and background.

It is more instructive to compare the correlation of the two effective mass variables $m_{\text{eff}}$ and $m_{\text{eff,b}}$. In Fig. 4.2, we plot the corresponding two-dimensional plane for the signal and the ratios of signal–to–background for the primary background ($b\bar{b}b\bar{b}$+jets) for both ALPGEN and SHERPA simulations. In comparison to the ALPGEN simulation the Sherpa result has much fewer events with high $m_{\text{eff}}$. This behavior is completely expected from the acceptance cut. While we have only up to one light-flavor gluon on matrix element level in the SHERPA

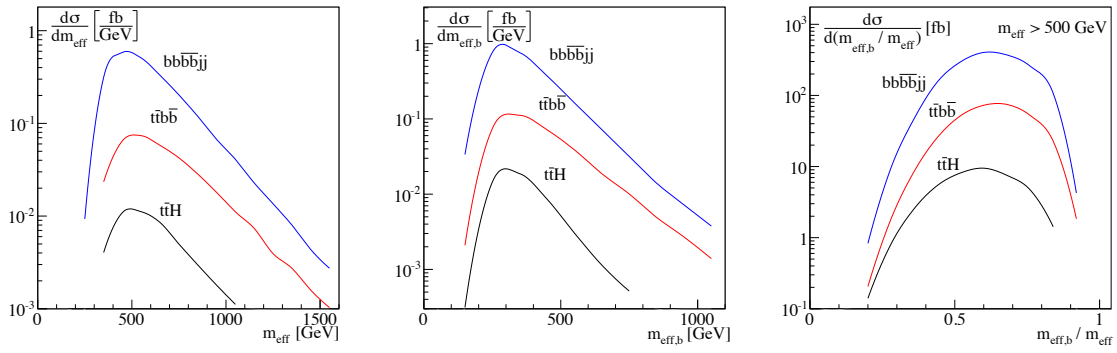


Figure 4.1.: Signal and background distributions for the effective mass of the entire jet system, the four $b$-tagged jets, and their ratio. All jets fulfill Eq.(4.1) and Eq.(4.2). We require $m_{\text{eff}} > 500$ GeV in the selection cuts of Eq.(4.4). For the $b\bar{b}b\bar{b}$ background we show the ALPGEN result with two additional hard jets plus parton shower.

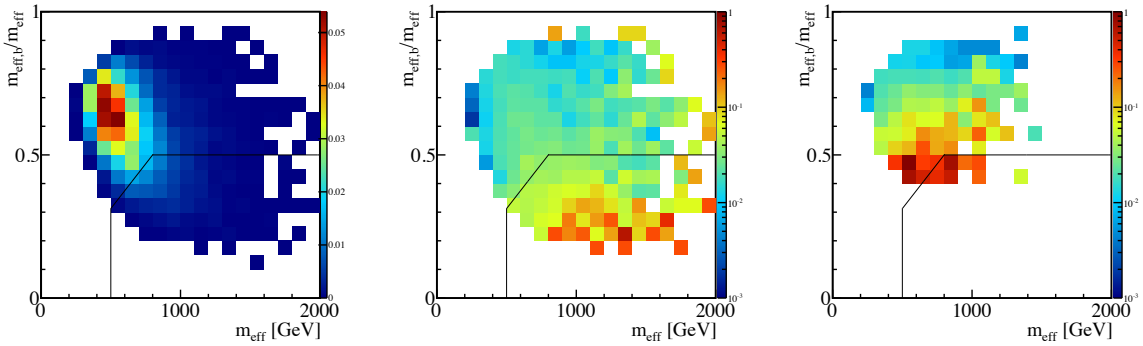


Figure 4.2.: Two-dimensional distribution of $m_{\text{eff}}$ vs. $m_{\text{eff},b}/m_{\text{eff}}$ for the $t\bar{t}H$ signal (left), and the ratio of the signal to the $bb\bar{b}\bar{b}jj$ ALPGEN background (center), and the ratio of signal to $bb\bar{b}\bar{b} + 0/1j$ SHERPA simulation (right). The lines represent the cuts of Eq.(4.4).

simulation, the acceptance cut asks for two hard un-tagged jets. As pointed out already we will rely on the more conservative ALPGEN simulation with two hard additional jets.

To reduce the background rate to a manageable level we use in addition to the acceptance cuts of Eqs. (4.1) and (4.2) the following cuts on based on our effective mass variables:

$$m_{\text{eff}} > 500 \text{ GeV} , \qquad \frac{m_{\text{eff},b}}{m_{\text{eff}}} < 0.5 , \qquad \frac{m_{\text{eff},b}}{m_{\text{eff}}} < \frac{m_{\text{eff}}}{1600 \text{ GeV}} . \tag{4.4}$$

These cuts are indicated in Fig. 4.2 by the solid black lines. The used phase space region corresponds to the lower right corner in the two dimensional plots. Using these cuts we can achieve a signal–to–background ratio of 1:25 as given in the second line of Tab. 4.1.

Starting from this stage, we want to identify an excess that contains the two $b$ jets associated to the Higgs decay. As already mentioned in the introductory chapters, this is suffering by the fact that even in the signal process we have to deal with at least four b jets that can be combined in various ways giving rise to combinatorial background. We test three naive approaches to choose a pair of b quarks. First, we pick the pair of $b$ jets that results in a dijet mass $m_{bb}$ closest to the Higgs mass $m_H = 125$ GeV. Furthermore, we try the $b$ jets with lowest or highest transverse momenta as well as the combination with minimal invariant mass. As can be seen in Tab. 4.1 all these methods fail in achieving an acceptable signal–to–background ratio.

From this result it is immediate that we need another method to reconstruct the final state and to minimize the problem of combinatorial backgrounds. Our method of choice is the bucket reconstruction technique [5] to rebuild the two top quarks in the events that pass our selection criteria of Eqs. (4.1). Once we have identified the decay products of the top quarks with sufficient accuracy, we have just two b jets left and therefore do not have to worry about combinatorics anymore. After this identification the cuts of Eq. (4.4) do not yield an improvement of the signal–to–background ratio. Consequently, we do not apply these cuts at any point below. Of course, we do not claim that this simple algorithm can replace a full experimental likelihood analysis, but it shows that a study of the fully hadronic decay channel of $t\bar{t}H$ is a realistic possibility.

## 4.2. Top buckets

As shown in the last section and Appendix B one can reduce the four $b$ quark QCD background to a manageable level. The remaining tasks are to deal with the irreducible $t\bar{t}b\bar{b}$ background and to extract the signal from the huge multi-jet background. Although the effect of Eq. (4.4) is already promising we should be able to do better by using a more specific analysis. Because of the low $t\bar{t}H$ signal rate one has to worry about the signal efficiency. Our method of choice is the bucket technique [5]. This algorithm will allow us to remove a large amount of background while not reducing the signal rate too much. Once we apply the acceptance cuts of Eq. (4.1) we are left with events containing four $b$ tagged jets and at least two hard untagged jets. Two of the four $b$ jets are presumed to originate from the decay of top quarks. The other two should then be the decay products of a Higgs boson. Without knowing which $b$ jets correspond to top decays there are various combinations of $b$ jets that can give rise to a Higgs candidate. To avoid this combinatorial background we try various criteria to define a particular pair as Higgs candidate. But as shown in the last section neither asking for the invariant mass closest to the true Higgs mass, nor the pair with minimal invariant mass, nor the pairing of the highest or lowest $p_T$ $b$ tagged jets, leads to acceptable results. In particular using the pair with minimal difference to the true Higgs mass suffers from the problem that this criterion selects both in signal and background often pairs that are close to the true Higgs mass but do not at all originate from a Higgs boson. As mentioned in the introduction the combinatorics are the reason why $t\bar{t}H$ has been removed from the list of promising Higgs discovery channels. Our approach to solve this problem is to first identify the top quarks using the bucket algorithm. Following the steps presented in Section 4.2 we distribute all jets to three buckets of which two will correspond to top quarks. The third (ISR) bucket in our case will contain two $b$ jets that build by exclusion the Higgs candidate.

We start by initializing each top bucket with a $b$ jet. Next we test all possible assignments of un-tagged jets to the three buckets where we require at least one non-$b$ jet in each of the top buckets. To find the best assignment we use the metric

$$\Delta_{B_i} = |m_{B_i} - m_t| \qquad \text{with} \quad m_{B_i}^2 = \left( \sum_{k \in B_i} p_k \right)^2 , \tag{4.5}$$

where $m_t$ is the top mass and the sum runs over all jets in the bucket $B_i$. We select the jet assignment and seeding of $b$ jets that minimizes $\Delta^2 = \omega \Delta_{B_1}^2 + \Delta_{B_2}^2$, where $\omega > 1$ is a factor chosen to stabilize the jet grouping. In this analysis, we choose $\omega = 100$. This choice essentially decouples the second bucket from the metric. Thus, bucket $B_1$ is the bucket with the invariant mass closest to the top. This procedure yields two top buckets $B_1$ and $B_2$ containing at least two jets. In rare cases we receive buckets with more than three jets. If this happens, we recluster to three jets using the Cambridge/Aachen algorithm. Applying a mass window cut for the top buckets we can reduce background events without real tops. We choose the mass window

$$155 \text{ GeV} < m_{B_{1,2}} < 200 \text{ GeV} . \tag{4.6}$$

Furthermore, we ask for a hadronically decayed $W$ boson in the buckets with two non-$b$ jets by implementing a mass ratio similar to the width of the A-shaped mass plane cuts in the HEPTopTagger

$$\left| \frac{m_{kl}}{m_{B_i}} - \frac{m_W}{m_t} \right| < 0.15 \tag{4.7}$$

for at least one combination of the non-$b$-jets (denoted $k$ and $l$) in the bucket $i$. We can classify events in one of three categories:

- ($\mathfrak{t}_w$,$\mathfrak{t}_w$): both top buckets have $W$ candidates as defined by Eq.(4.7),
- ($\mathfrak{t}_w$,$\mathfrak{t}_-$) or ($\mathfrak{t}_-$,$\mathfrak{t}_w$): only the first or second top bucket has a $W$ candidate,
- ($\mathfrak{t}_-$,$\mathfrak{t}_-$): neither top bucket has a $W$ candidate.

Buckets failing the criteria of Eq.(4.7) – this includes all top buckets containing only one non-$b$ jet – can still be used to identify top quarks using an alternative metric

$$\Delta_B^{bj} = \begin{cases} |m_B - 145 \text{ GeV}| & \text{if } m_B \leq 155 \text{ GeV} \\ \infty & \text{else} \end{cases} . \tag{4.8}$$

The physical interpretation of the $b/j$ buckets are top decays where one of the $W$ decay jets (typically the softer one) is lost in the detector. Using this metric we re-distribute the contents of all $\mathfrak{t}_-$ buckets and the ISR bucket such that $\sum_i \Delta_{B_i}^{bj}$ is minimal. We identify a resulting $b/j$ bucket as top candidate if it fulfills

$$75 \text{ GeV} < m_{bj} < 155 \text{ GeV} . \tag{4.9}$$

An overview on the rate of signal and background events passing the top reconstruction given above is provided in Tab. 4.2. Our main goal is not to provide a reliable value for the top four momenta, we are interested in the $b$ jets that do not originate from the tops. They are the objects we need to reconstruct a Higgs. For a Higgs identification we require these two $b$ tagged jets to combine to an invariant mass between 90 and 130 GeV. In the purest sample of reconstructed tops ($p_{T,t,1} > 300$ GeV), 68% of the signal events passing the top reconstruction have the two remaining $b$-jets in the ISR bucket can be associated with a parton level Higgs

| | $t\bar{t}H$ | | $t\bar{t}b\bar{b}$ | $b\bar{b}b\bar{b}jj$ | $S/B$ |
|---|---|---|---|---|---|
| After acceptance cuts Eqs.(4.1) and (4.2) | 1.197 | | 8.363 | 54.420 | 0.019 |
| 2 tops tagged | 0.894 | (0.184) | 5.882 | 29.356 | 0.025 |
| $p_{T,t,1} > 100$ GeV | 0.709 | (0.158) | 4.868 | 20.838 | 0.028 |
| $p_{T,t,1} > 200$ GeV | 0.289 | (0.080) | 2.189 | 5.194 | 0.039 |
| $p_{T,t,1} > 300$ GeV | 0.089 | (0.028) | 0.724 | 0.917 | 0.054 |
| Mass window $m_{bb} = 90 - 130$ GeV | | | | | |
| 2 tops tagged | 0.259 | (0.121) | 0.859 | 5.424 | 0.041 |
| $p_{T,t,1} > 100$ GeV | 0.208 | (0.105) | 0.688 | 3.600 | 0.048 |
| $p_{T,t,1} > 200$ GeV | 0.091 | (0.054) | 0.265 | 0.679 | 0.096 |
| $p_{T,t,1} > 300$ GeV | 0.028 | (0.019) | 0.072 | 0.082 | 0.182 |

Table 4.2.: Cross sections (in fb) of events after the acceptance cuts of Eqs.(4.1) and (4.2) and requiring two tops passing the bucket reconstruction. We also require one of the reconstructed tops to pass various $p_T$ thresholds, with and without requiring the two remaining $b$-jets to have invariant mass inside the window 90-130 GeV. Number in parenthesis correspond to the events where the reconstructed Higgs lies within $\Delta R < 0.5$ of the true Higgs.

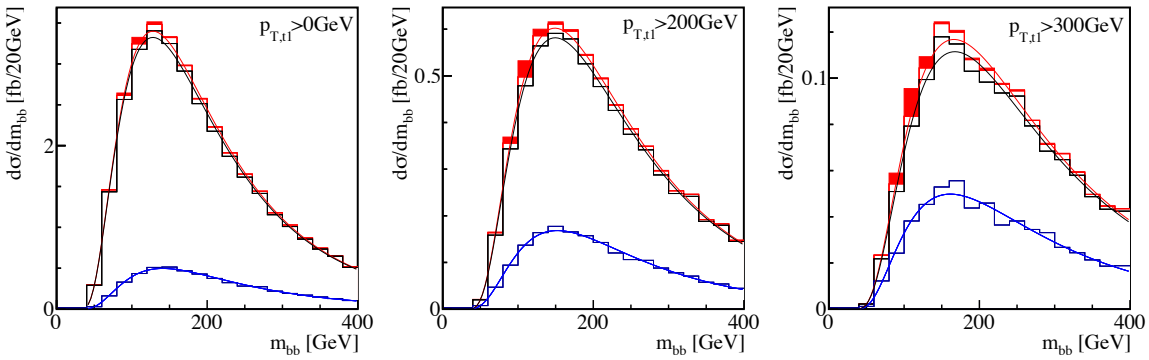


Figure 4.3.: Stacked $m_{bb}$ distribution, built from $b$-jets not used for top buckets, after top reconstruction and requiring the leading top $p_{T,t_1} > 0, 200,$ and 300 GeV (from left). The two primary backgrounds are $t\bar{t}b\bar{b}$ (blue) and $bb\bar{b}\bar{b}$ (black). The signal distribution of reconstructed tops is shown in red. The subset of signal events where the reconstructed Higgs lies within $\Delta R < 0.5$ of the true Higgs are displayed as filled-in red regions.

within $\Delta R < 0.5$. For hard top quarks we can find signal–to–background ratios around 1:10. Although this seems promising, for a determination of the top Yukawa coupling one needs easily accessible side bands.

Such side bands are provided by the invariant mass distribution of the two $b$ quarks that are left after top reconstruction. For the signal one expects a peak around the true Higgs mass of 125 GeV. The actual center of the peak may be shifted towards lower mass caused by final state radiation escaping the reconstruction. The background distribution should be well described by a log-normal distribution. The obtained mass distributions of events passing as well acceptance cuts as the top reconstruction are given in Fig. 4.3. The narrow mass peak of the signal is well separated from a broad feature of the backgrounds. The fact that not all Higgs candidates in the signal region can be associated with a parton–level Higgs boson

| | $t\bar{t}H$ | | $t\bar{t}b\bar{b}$ | $bb\bar{b}\bar{b}jj$ | $S/B$ |
|---|---|---|---|---|---|
| After acceptance cuts Eqs.(4.1) and (4.2) | 1.197 | | 8.363 | 54.420 | 0.019 |
| 2 tops tagged, $\Delta\eta$ cuts | 0.587 | (0.125) | 2.762 | 10.654 | 0.044 |
| $p_{T,t,1} > 100$ GeV | 0.485 | (0.111) | 2.392 | 8.364 | 0.045 |
| $p_{T,t,1} > 200$ GeV | 0.207 | (0.059) | 1.153 | 2.541 | 0.056 |
| $p_{T,t,1} > 300$ GeV | 0.064 | (0.021) | 0.405 | 0.507 | 0.071 |
| Mass window $m_{bb} = 90 - 130$ GeV | | | | | |
| 2 tops tagged, $\Delta\eta$ cuts | 0.170 | (0.080) | 0.376 | 1.864 | 0.076 |
| $p_{T,t,1} > 100$ GeV | 0.144 | (0.072) | 0.317 | 1.396 | 0.084 |
| $p_{T,t,1} > 200$ GeV | 0.066 | (0.039) | 0.129 | 0.338 | 0.142 |
| $p_{T,t,1} > 300$ GeV | 0.021 | (0.015) | 0.034 | 0.043 | 0.276 |

Table 4.3.: Cross sections (in fb) of events after successive selection cuts, as in Tab. 4.2 but including the $\Delta\eta$ cuts of Eq.(4.10).

is caused by $b$ combinatorics or missing final state radiation. This improves after requiring a minimal transverse momentum for the reconstructed tops as can be seen by comparing the different panels of Fig. 4.3. Using this additional $p_T$ cut we find a clean peak over the broad background distributions. The signal and background rates for events with two reconstructed tops and a left over $b\bar{b}$ of an invariant mass inside the mass window of 90-130 GeV is given in Tab 4.2. A comparison with the effective mass cuts of section 4.1 is shown in Tab. 4.3. Requiring one of the two reconstructed top quarks to have a transverse momentum above 300 GeV improves the signal–to–background ratio by a factor two.

Another feature of the top reconstruction method is that we can implement further cuts on the reconstructed top quarks as well as on the Higgs. By requiring the tops and the Higgs to be not too widely separated in rapidity according to

$$\Delta\eta(t_1, t_2) < 3 \; , \quad \Delta\eta(t_i, H) < 2 \tag{4.10}$$

we can improve the S/B ratio significantly. The obtained rates are shown in Fig. 4.3. Using the most restrictive cuts, asking for a leading top with transverse momentum above 300 GeV and the rapidity cuts of Eq. (4.10), we find $S/B = 1/3.6$ with 70% correctly identified Higgs bosons. Therefore, using the top reconstruction helps to improve the signal–to–background ratio in two ways. It allows us to identify the $b$ jets corresponding to the Higgs boson by exclusion and removes in consequence the problem of combinatorial backgrounds. Moreover, we can analyze the kinematics of the reconstructed top quarks as well as the Higgs and place additional cuts to improve the signal–to–background ratio.

# 5. Conclusion & Outlook

The aim of the first part of this thesis was to improve the performance of the HEPTopTagger with an increased cone size of $R = 1.8$. We proposed a change in the ordering of cuts to reduce the background shaping. Additionally, we suggested a new distance measure during the algorithm which might be useful in high multiplicity final states. Using boosted decision trees we tried to quantify the improvements one can gain by taking into account angular correlations between the subjets. To treat the angular correlations in a systematic way, we relied on Fox–Wolfram moments. By adding the information from the moments to the boosted decision trees we could not find reasonable effects. This means that our tagger with inverted cut order does not benefit from angular correlations between the subjets. Finally, we extended the HEPTopTagger towards lower transverse momenta by focusing on top candidates that miss the softest of the decay products. Using this special candidates one can avoid limitations due to the chosen cone size which is critical in the regime of low transverse momenta. For this purpose we introduced an additional tagging step for candidates in the $p_T$ regime between 150 and 200 GeV using rectangular cuts on Fox-Wolfram moments. However, the actual use of our modifications has to be studied in future analyses.

In the second part of this thesis, we demonstrated that it should be possible to extract the signal of the fully hadronic decay channel of $t\bar{t}H$ at the LHC. The proposed method relies on the bucket technique to reconstruct the top quarks and identifies the Higgs decay jets by exclusion. In this way one avoids the challenging combinatorial backgrounds. The resultant invariant mass distributions of the left-over $b\bar{b}$ pair give access to side bands that can be used to extract background shapes and allow for a determination of the background cross section after cuts. The reconstruction of top and Higgs momenta allows additional cuts on the event kinematics that can improve the signal–to–background ratio further. Of course all this has to be understood as a prove–of–concept analysis. More detailed studies might improve the method. By using different Monte Carlo generators we could demonstrate that simulation uncertainties are under control. The theoretical uncertainties could be hardly covered by variation of the scale choices. However, these uncertainties can be controlled in experiments by a study of side bands, *e.g.* in the invariant mass distributions of the two Higgs candidate $b$ jets.

## Acknowledgements

First of all, I want to thank Tilman for the opportunity to work on this project and for his advise. I am further grateful to all other collaborators during this time: Cathy, Matt and Michi. In particular, Cathy always had time for discussions and encouraging words. Daniel, who left to Pittsburgh, I want thank for all the nice discussions we had in the Westzimmer and later on in our new office. Additionally, I want to thank the rest of the group: Nishita, Jamie, Martin, Jamil, Peter, Bea, Karin and Johann. Working with you really was fun. Furthermore, I appreciated support from the Studienstiftung des deutschen Volkes. Without it life would have been much harder. Since this is the end of my undergraduate studies, I want to mention Hannes, Jürgen and Lennart. Without you the past years would have been much less enjoyable. Should this thesis be readable at all, this is due the careful cross–reading by Cathy, Johann and Jürgen. Finally, I am grateful to my family for all their support as well as to my girlfriend Sabrina who had to suffer especially in the final phase of my studies.

# A. Implementation of Simulated Annealing and Boosted Decision Trees in TMVA

In this appendix we sketch how the multivariate analysis methods used in this thesis are implemented in the framework of TMVA [50] provided for ROOT [57].

## A.1. Simulated annealing

We sketch how the SA algorithm is implemented in the ROOT TMVA package. A general introduction to this algorithm is given in Section 2.8 where also most of the variables used in this section are defined. Furthermore, we will explain how the receiver operating characteristic (ROC) curve is obtained during this optimization procedure. The ROC curve is our main point of interest. It gives the best achieved background rejection $r_B = 1 - \epsilon_B$ for given signal efficiency $\epsilon_S$. Together with the corresponding cuts this allows to determine the cuts we should use for a chosen signal efficiency or background rejection. This is helpful if one is for example interested in using the HEPTopTagger not at its standard working point but with a background rejection that is adjusted to the particular study.

The main aim of the optimization procedure is to minimize an estimator function $f(\vec{x})$, where $\vec{x}$ is a point in parameter space. We start at $T = T_{\min}$. The temperature at a given time step $t$ is obtained via

$$T_t = T_{\min} + \log(1 + i) \ . \tag{A.1}$$

Here the integer $i$ counts the number of tries starting from the same point as it will be explained below. The next random point is determined using the temperature T and generating a random number $u \in [0, 1]$. Dependent on $u$ we pick the sign $s$ to be positive for $u \geq 0.5$ or negative otherwise. With this at hand we use the distribution

$$d(s(u), u) = sT \left(1 + \frac{1}{T}\right)^{|2u-1|-1} \tag{A.2}$$

to find a new value for the parameter $x$

$$x' = x + \frac{d}{10}(p_{\max} - p_{\min}) \ . \tag{A.3}$$

In this implementation the temperature is changed in each time step. In particular, $t$ is increased after each try and $i$ is reset only in case of a jump to a new point in parameter space. However, $i$ has impact on the next temperature. The optimization is based on the following steps:

- For the given parameters (cuts) calculate the signal and background efficiencies $\epsilon_S$ and $\epsilon_B$.

- From the parameters checked before there may be already a value for $\epsilon_B(\epsilon_S)$. This value is labeled $\epsilon_{BH}$. For the recent $\epsilon_S$ take $\epsilon_{BH}$. Additionally, take the stored values in the side bins and calculate their average $a = (\epsilon_{BH}^{\text{left}} + \epsilon_{BH}^{\text{right}})/2$. If $\epsilon_{BH} < \epsilon_B$ set $a = \epsilon_{BH}$.

- Return $f(x) = \eta$ with

$$\eta = \frac{-|\epsilon_{BH} - a| + 1 - (\epsilon_{BH} - \epsilon_B)}{1 + \epsilon_S} \,, \tag{A.4}$$

  which becomes smaller for smaller $\epsilon_B$ at fixed $\epsilon_S$ and larger $\epsilon_S$ at fixed $\epsilon_B$. To avoid problems with the $\epsilon_S = 0$ bin there is a penalty introduced for this bin.

- If there is no entry for $\epsilon_{BH}$ or $\epsilon_B < \epsilon_{BH}$ (corresponds to a better background rejection), the code replaces $\epsilon_B(\epsilon_S)$ and changes the stored cuts accordingly.

- The ROC curve being defined $\epsilon_B(\epsilon_S)$ is therefore updated with any new point in parameter space.

The procedure stops if $t$ reaches the number of maximal calls $t_{\max}$.

## A.2. Boosted decision trees

In this section we try to give an impression how boosted decision trees are implemented in TMVA. In this discussion we will follow the default algorithms. For simplicity we will assume unweighted events wherever this is possible.

First of all, we have to build a decision tree from a training sample. We assume to have $N_{\text{var}}$ available variables for our analysis. For constructing the decision tree we need to define certain nodes at which we sort incoming events based on a cut in two categories (left or right). To find the appropriate variable and the specific cut we proceed as follows:

- Divide the range of each variable $x$ into $N_{\text{cut}}$ bins. Based on the incoming events we receive possible cuts at $x_{\text{cut}} = n(x_{\max} - x_{\min})/N_{\text{bins}}$ with integer $n \in [1, N_{\text{cut}} - 1]$. The cuts coinciding with the maximal and minimal value $x_{\min}$ and $x_{\max}$ can be omitted since they do not lead to a splitting.

- Scan through all variables and possible cut values for each variable and determine the number of signal events $S_i$ and background events $B_i$ in the two resultant branches $i = L, R$.

- If the left and the right branch contain a total number of training events above a certain threshold, we consider this configuration as a possible splitting and calculate the separation gain

$$\frac{I_P - I_L - I_R}{N_P} \,, \tag{A.5}$$

  where $I_i$ are the separation indices calculated for the left ($L$), right ($R$) the parent node ($P = L + R$). The separation indices are defined as the misclassification rates multiplied by the number of events $N_i$ reaching the particular node.

- Choose the splitting with maximal separation gain.

- If the maximal splitting gain is below a cut-off $\epsilon$ we do not apply a splitting and the note becomes a leaf. The leaf is declared as signal or background leaf according the dominant kind of events ending up on it.

Starting from the first node (root node) this procedure can be iterated on every non-leaf node until there are just leaves left or the maximum number of layers is reached. In the latter case we declare nodes that are left over to leaves.

The procedure described above allows to build a single decision tree. To avoid being too sensitive to fluctuations in the used training sample one boosts the tree giving rise to a forest of several trees. The default boosting algorithm in TMVA is AdaBoost [51]. After we built a decision tree from the test sample we can calculate its misidentification rate err. Using this information we reweight all events in the training sample that have been misclassified by a factor

$$\alpha = \left( \frac{1 - \mathrm{err}}{\mathrm{err}} \right)^{\beta} . \tag{A.6}$$

The default choice for the boosting parameter is $\beta = 1$. The reweighted test sample is then used to grow a second decision tree. By iterating this procedure one builds an entire forest of $N_{\mathrm{tree}}$ trees.

Each tree in the forest can be pruned to remove internal notes. The pruning is performed using a validation sample different from the training sample.

Finally, we end up in a forest of pruned trees. The actual classification of an event is then obtained by a (possibly weighted) majority vote.

# B. Signal and Background Simulation for Buckets of Higgs and Tops

This appendix is taken from [6]. We adapted the references in the text to fit to the discussions in this thesis.

In this Appendix we will confirm that the analysis described in this paper does not critically depend on uncertainties in the way we compute our signal and backgrounds. For the signal and the $t\bar{t}b\bar{b}$ background we primarily rely on SHERPA [61] predictions with up to one additional hard jet merged using the CKKW approach [26]. For the $t\bar{t}H$ signal we test our results using MADGRAPH [59], with up to one hard jet included in the MLM scheme [27]. Both event samples are normalized to the next-to-leading order rate (extrapolated to 13 TeV) of 504 fb [62, 63], times a Higgs branching ratio of 57.7%. This corresponds to 129 fb for the purely hadronic decay channel. In Table B.1 we observe a small difference in the normalization of the two event samples. The reason is that as a cross check in the MADGRAPH simulation, we do not require hadronic top decays in the simulation. As a result, the decays to hadronic taus contribute to the signal. Our default SHERPA simulation conservatively does not include these events.

For the $t\bar{t}b\bar{b}$ background we test the SHERPA simulation with up to one hard additional jet with an ALPGEN [52] simulation without additional hard jets. Again, both samples are normalized to the next-to-leading order rate of 1037 fb [64] after the generator cuts $p_{T,b} > 35$ GeV, $|\eta_b| < 2.5$, and $\Delta R_{bb} > 0.9$. This rate is approximate because, in the absence of a next-to-leading order prediction for $\sqrt{s} = 13$ TeV, we are forced to first extract the $K$ factor for 14 TeV and the cuts of Ref. [64], including a regularizing cut on the invariant mass of the two bottom quarks. We then multiply our cross section at 13 TeV by this $K$ factor. This approach is not ideal, but better then just using the leading order prediction.

In Table B.1 we see that the transverse momentum distributions for the "reconstructed" top quarks (which, for $bb\bar{b}\bar{b}$, do not correspond to any parton-level tops) from ALPGEN are softer than for SHERPA. This effect comes from the generically harder jets of CKKW merging, compared to those from the parton shower. In order to be conservative, we use the merged SHERPA results for our analysis. On the other hand, the difference of less than 20% is well within the theory uncertainties for this background.

As argued in Section 4.1, the most dangerous background events should be correctly described by our ALPGEN simulation of the $bb\bar{b}\bar{b}jj$ background plus PYTHIA parton shower, as the required extra jets must be hard, and therefore well-modeled by the matrix-level process. With our merged SHERPA simulation of $bb\bar{b}\bar{b} + 0/1$ jet, we test several aspects of our main background simulation:

1. We check if the events with two additional hard jets are indeed the leading background after the kind of global cuts proposed in Section 4.1. This aspect is very important for the appropriate simulation of the QCD background in an actual analysis.

| | $t\bar{t}H$ | | $t\bar{t}b\bar{b}$ | |
|---|---|---|---|---|
| | MADGRAPH (merged) | SHERPA (merged) | ALPGEN (shower) | SHERPA (merged) |
| After acceptance Eqs.(4.1) and (4.2) | 1.390 | 1.197 | 7.903 | 8.363 |
| 2 tops tagged | 1.100 | 0.894 | 5.893 | 5.882 |
| $p_{T,t,1} > 100$ GeV | 0.866 | 0.709 | 4.684 | 4.868 |
| $p_{T,t,1} > 200$ GeV | 0.342 | 0.289 | 1.806 | 2.189 |
| $p_{T,t,1} > 250$ GeV | 0.180 | 0.165 | 0.978 | 1.295 |
| $p_{T,t,1} > 300$ GeV | 0.092 | 0.089 | 0.502 | 0.724 |
| | Mass window $m_{bb} = 90 - 130$ GeV | | | |
| 2 tops tagged | 0.337 | 0.259 | 1.016 | 0.859 |
| $p_{T,t,1} > 100$ GeV | 0.274 | 0.208 | 0.780 | 0.688 |
| $p_{T,t,1} > 200$ GeV | 0.112 | 0.091 | 0.260 | 0.265 |
| $p_{T,t,1} > 250$ GeV | 0.058 | 0.050 | 0.128 | 0.144 |
| $p_{T,t,1} > 300$ GeV | 0.032 | 0.028 | 0.059 | 0.072 |

Table B.1.: Signal and background cross sections (in fb) after successive selection cuts, showing the different ways of simulating the signal and the irreducible $t\bar{t}b\bar{b}$ background. All conventions correspond to the final result shown in Table 4.2. We use the SHERPA results for our main analysis.

2. We test if our analysis depends on the simulation of the second un-tagged jet either with the hard matrix element or through the parton shower. In this way we can estimate an important source of theory uncertainties.

3. As a measure of the level of agreement between the two simulations we compute the merged SHERPA event rate with a consistent variation of the renormalization and factorization scales. Ideally, the two simulations should agree within this scale variation in the signal region of the buckets analysis. Because the merged SHERPA prediction includes some leading next-to-leading order contribution such a numerical agreement also indicates that our $bb\bar{b}\bar{b}jj$ simulations should not be plagued by huge QCD corrections.

This extensive list of tests should give a clear answer to the question if purely hadronic $t\bar{t}H$ searches can be done in the presence of the large QCD backgrounds. Finally, we point out that merged SHERPA simulations can define excellent side bands in the $n_j$ distribution [65, 66, 67], which together with side bands in $m_{bb}$ should be sufficient to control the background rate in the signal region in an experimental analysis. Our results suggest that such an approach would require a merged simulation of $bb\bar{b}\bar{b}$ with up to at least two hard light-flavor or gluon jets, which is beyond our CPU capabilities.

In Figure B.1 we first show the normalized transverse momenta of the four $b$-jets. The curves are set to unit normalization, as the significantly different cross section of the $bb\bar{b}\bar{b}jj$ and the merged $bb\bar{b}\bar{b}+0/1$ jets is almost entirely due to the different number of un-tagged jets in the events. In the left panel we see that the leading $b$-jet agrees in the two approaches, while the second to fourth $b$-jets become increasingly harder in the ALPGEN $bb\bar{b}\bar{b}jj$ sample. This is because, with two additional hard jets, the available recoil momentum is slightly larger. The sensitivity to the proper simulation of the recoil is also the reason why the ALPGEN curves are not covered by the scale variation of the SHERPA simulation.

The results for the leading un-tagged jets in the right panel of Figure B.1 look much less promising. The very different integrated rates under the curves reflect the additional events with only $bb\bar{b}\bar{b}$ in the hard process plus any number of parton shower jets. This is particularly obvious for the first un-tagged jet, where the SHERPA simulation includes a majority of events

with only one additional jet while the ALPGEN sample will always include a second hard jet together with the first. For the second un-tagged jet the integrated rates in the $p_{T,j}$ distributions are similar for the two samples. The ALPGEN simulation gives a significantly harder second jet from the matrix element while the second jet in our SHERPA sample (corresponding to the first jet from the parton shower) tends to be soft. The third un-tagged jet is the second parton shower jet in our SHERPA sample, while in the ALPGEN simulation it is the first parton shower jet radiated from a harder core process. Both effects combined result in a significantly harder $p_{T,j}$ spectrum for the ALPGEN sample. These distribution suggest that if our signal region should indeed require two or even three hard un-tagged jets to mimic top decay jets, the $bb\bar{b}\bar{b}jj$ sample from ALPGEN should be the appropriate, conservative estimate.

In the upper panels of Figure B.2 we show the two relevant effective mass variables defined in Eq.(4.3). We require four $b$-jets according to Eq.(4.1) and any number of un-tagged jets after Eq.(4.2). Unlike Figure 4.1 we now only show the different results for the $bb\bar{b}\bar{b}$ background. In the upper left panel we again show the normalized observable from the multi-$b$ sector. The conclusion follows from the discussion of the transverse momenta of the four $b$-jets: the dependence of the $m_{\text{eff},b}$ distribution on the simulation is small, clearly when we look at the SHERPA scale variation, but also in terms of the difference between ALPGEN and SHERPA. The only difference is that the $bb\bar{b}\bar{b}jj$ simulation with ALPGEN predicts slightly harder multi-$b$ sectors.

In the upper central panel of Figure B.2 we see that in the background region the difference in rate between the two simulations is again dramatic, and certainly not covered by the scale variation of the $bb\bar{b}\bar{b} + 0/1$ jets simulation with SHERPA. On the other hand, this result is entirely expected, and the difference becomes increasingly smaller once our analysis of the signal region requires something like $m_{\text{eff}} > 500$ GeV. Increasing the cut to $m_{\text{eff}} > 700$ GeV brings them into agreement within scale uncertainties. In that regime the bulk of the $b\bar{b}b\bar{b}+0/1$ jet events do not contribute, so the two simulations should roughly agree within the

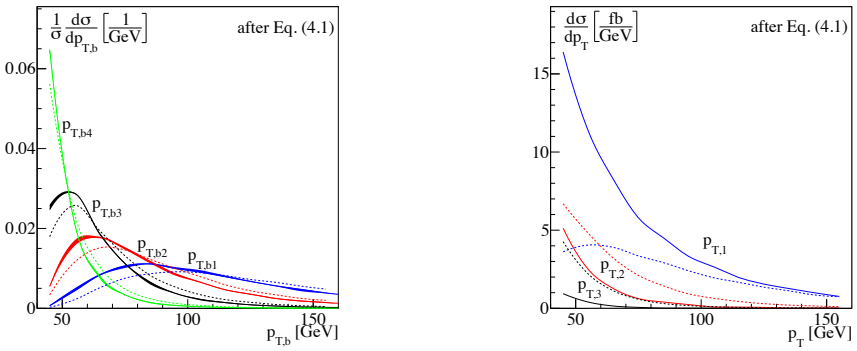


Figure B.1.: Normalized transverse momentum distributions of four $b$-tagged jets (left) and the leading three additional un-tagged jets (right). All events include four hard $b$-jets according to Eq.(4.1), but no requirement on the number of additional un-tagged jets. The solid curves correspond to the merged $bb\bar{b}\bar{b}+0/1$ jets simulation with SHERPA while the dashed curves show the $bb\bar{b}\bar{b}jj$ events from ALPGEN. The scale variation for the SHERPA result is indicated by the widths of the solid lines in the left panel.

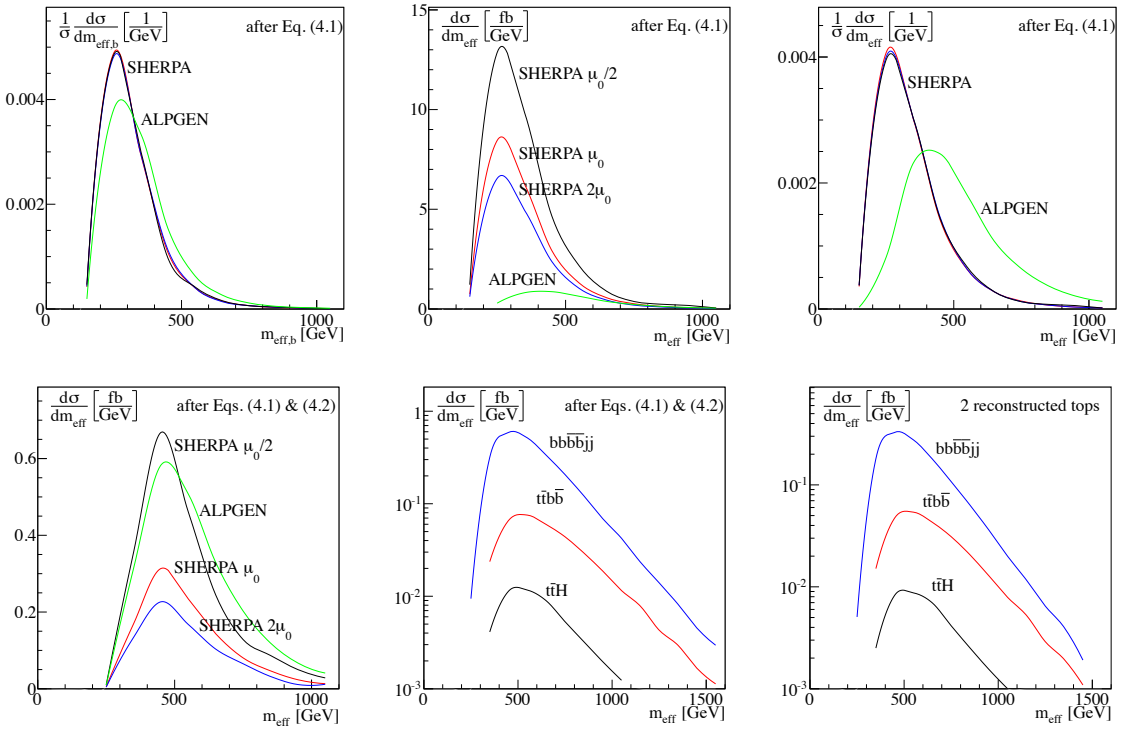


Figure B.2.: Effective mass distributions. In the upper panels we require four hard $b$-jets according to Eq.(4.1), but no cut on the number of additional un-tagged jets. We show the normalized $m_{\text{eff},b}$ distribution for four $b$-jets (left), as well as the all-jet $m_{\text{eff}}$ from SHERPA and ALPGEN normalized to the rates (center) and normalized to unity (right). In the lower panels we require two additional un-tagged jets fulfilling Eq.(4.2). Here we show the ALPGEN vs SHERPA results (left), our default signal and background samples (center), and the default signal and backgrounds for all events with two valid top buckets (right).

scale variation of the SHERPA simulation. The upper right panel confirms that in the ALPGEN simulation with the $bb\bar{b}\bar{b}jj$ hard process also gives a harder spectrum in $m_{\text{eff}}$.

In the lower panels of Figure B.2 we not only require four $b$-jets following Eq.(4.1), but also at least two un-tagged jets fulfilling Eq.(4.2). Two such additional jets are implicitly required for any event passing the bucket analysis described in Section 4.2. First, we see in the left panel that simply asking for two un-tagged jets suppresses the central prediction from the merged $bb\bar{b}\bar{b} + 0/1$ jets simulation to roughly half the $bb\bar{b}\bar{b}jj$ prediction. Both $m_{\text{eff}}$ distributions peak around 500 GeV, and the ALPGEN rate is covered by the scale variation of the SHERPA simulation. In the central lower panel we compare the distributions for our default signal and background simulations after requiring four $b$-tagged and two un-tagged jets. Finally, in the lower right panel we show the same distribution for all events passing the bucket analysis described in Section 4.2. As compared to the acceptance cuts, Eqs.(4.1) and (4.2), there is hardly any change, which means that the improvements by the bucket analysis are more promising than the global cuts proposed in Section 4.1, with the added advantage of avoiding shaping the background distributions, such as $m_{\text{eff}}$.

| | $t\bar{t}H$ SHERPA | $bb\bar{b}\bar{b}jj$ ALPGEN | $bb\bar{b}\bar{b}$+jets SHERPA | | |
|---|---|---|---|---|---|
| | | | $2\mu_0$ | $\mu_0$ | $\mu_0/2$ |
| After acceptance Eqs.(4.1) and (4.2) | 1.197 | 54.420 | 18.825 | 25.812 | 50.974 |
| 2 tops tagged | 0.894 | 29.356 | 7.507 | 10.091 | 20.473 |
| $p_{T,t,1} > 100$ GeV | 0.709 | 20.838 | 5.049 | 7.283 | 13.843 |
| $p_{T,t,1} > 200$ GeV | 0.289 | 5.194 | 1.155 | 1.419 | 3.018 |
| $p_{T,t,1} > 250$ GeV | 0.165 | 2.213 | 0.488 | 0.717 | 1.361 |
| $p_{T,t,1} > 300$ GeV | 0.089 | 0.917 | 0.218 | 0.351 | 0.645 |
| | Mass window $m_{bb} = 90 - 130$ GeV | | | | |
| 2 tops tagged | 0.259 | 5.424 | 1.143 | 1.726 | 3.354 |
| $p_{T,t,1} > 100$ GeV | 0.208 | 3.600 | 0.749 | 1.111 | 2.118 |
| $p_{T,t,1} > 200$ GeV | 0.091 | 0.679 | 0.133 | 0.161 | 0.233 |
| $p_{T,t,1} > 250$ GeV | 0.050 | 0.233 | 0.031 | 0.044 | 0.082 |
| $p_{T,t,1} > 300$ GeV | 0.028 | 0.082 | 0.020 | 0.015 | 0.041 |

Table B.2.: Cross section (in fb) for signal and background events after successive selection cuts, showing the different ways of simulating the $bb\bar{b}\bar{b}$ background. All conventions correspond to the final result shown in Table 4.2. Reconstructed tops for $bb\bar{b}\bar{b}$ do not correspond to any real parton-level object.

From the above comparison we expect that for an actual top and Higgs analysis the two background simulations should be fairly consistent once we probe sufficiently hard multi-jet configurations. The scale uncertainty of our SHERPA simulation determines the numerical level of this consistency. Moreover, in the signal phase space region the $bb\bar{b}\bar{b}jj$ background simulation with parton shower should predict larger backgrounds and give us a conservative estimate. In Table B.2 we show the different $bb\bar{b}\bar{b}$ background rates after the buckets analysis. Indeed, the simulations agree roughly within the sizable scale uncertainties. The $b\bar{b}b\bar{b}jj$ simulation with ALPGEN gives the largest rate, in particular once we require large, signal-like $m_{bb}$ values and sizable $p_T$ of the fake reconstructed top buckets. For the experimental analysis this implies firstly that the signal-to-background ratio for purely hadronic $t\bar{t}H$ events at 13 TeV can be of the order 1/3. Second, the remaining number of signal events will be an issue for a cut-and-count analysis. Lastly, the uncertainties on the background simulation will require a careful background determination from side bands and control regions. Any kind of $t\bar{t}H$ analysis which does not provide at least a slight mass peak in the $m_{bb}$ distribution around 125 GeV would have a hard time convincing the authors of this study. Our buckets analysis will carefully ensure that this simple side band is clearly visible, in spite of the fact that this requirement might lead to a slightly reduced performance of our analysis.

# Bibliography


[1] 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 (2012) [arXiv:1207.7214 [hep-ex]].

[2] 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]].

[3] S. Heinemeyer *et al.* [LHC Higgs Cross Section Working Group Collaboration], "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties," arXiv:1307.1347 [hep-ph].

[4] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, "Stop Reconstruction with Tagged Tops," JHEP **1010**, 078 (2010).

[5] M. R. Buckley, T. Plehn and M. Takeuchi, "Buckets of Tops," JHEP **1308**, 086 (2013).

[6] M. R. Buckley, T. Plehn, T. Schell and M. Takeuchi, "Buckets of Higgs and Tops," arXiv:1310.6034 [hep-ph].

[7] C. Anders, C. Bernaciak, G. Kasieczka, T. Plehn and T. Schell, "Benchmarking an even better top tagger algorithm," in preparation.

[8] T. Plehn, "Lectures on LHC Physics," Lect. Notes Phys. **844**, 1 (2012) [arXiv:0910.4182].

[9] M. Srednicki, "Quantum Field Theory," Cambridge Univ. Press, 2010.

[10] M. E. Peskin and D. V. Schroeder, "An Introduction to Quantum Field Theory," Westview Press, 2007.

[11] R. K. Ellis, J. Stirling, and B. R. Webber, "QCD and Collider Physics," Cambridge Univ. Press, 2003.

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

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

[14] 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.

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

[16] P. W. Higgs, "Broken symmetries and the masses of gauge bosons," Phys. Rev. Lett. **13**, 508 (1964).

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

[18] J. Beringer *et al.* [Particle Data Group Collaboration], "Review of Particle Physics (RPP)," Phys. Rev. D **86**, 010001 (2012).

[19] `http://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg`, November 2013.

[20] R. N. Mohapatra *et al.*, "Theory of neutrinos: A White paper," Rept. Prog. Phys. **70**, 1757 (2007) [hep-ph/0510213].

[21] G. Bertone, D. Hooper and J. Silk, "Particle dark matter: Evidence, candidates and constraints," Phys. Rept. **405**, 279 (2005) [hep-ph/0404175].

[22] J. Frieman, M. Turner and D. Huterer, "Dark Energy and the Accelerating Universe," Ann. Rev. Astron. Astrophys. **46**, 385 (2008) [arXiv:0803.0982 [astro-ph]].

[23] 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]].

[24] 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]].

[25] S. D. Ellis, C. K. Vermilion and J. R. Walsh, "Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches," Phys. Rev. D **81**, 094023 (2010) [arXiv:0912.0033 [hep-ph]].

[26] S. Catani, F. Krauss, R. Kuhn and B. R. Webber, "QCD matrix elements + parton showers," JHEP **0111**, 063 (2001) [hep-ph/0109231].

[27] M. L. Mangano, M. Moretti, F. Piccinini and M. Treccani, "Matching matrix elements and shower evolution for top-quark production in hadronic collisions," JHEP **0701** (2007) 013 [arXiv:hep-ph/0611129].

[28] 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]].

[29] G. Aad *et al.* [ATLAS Collaboration], "The ATLAS Experiment at the CERN Large Hadron Collider," JINST **3**, S08003 (2008).

[30] S. Chatrchyan *et al.* [CMS Collaboration], "The CMS experiment at the CERN LHC," JINST **3**, S08004 (2008).

[31] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens and M. Selvaggi, "DELPHES 3, A modular framework for fast simulation of a generic collider experiment," arXiv:1307.6346 [hep-ex].

[32] R. Brun, F. Carminati and S. Giani, "GEANT Detector Description and Simulation Tool," CERN-W5013.

[33] 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).

[34] Y. L. Dokshitzer, G. D. Leder, S. Moretti and B. R. Webber, "Better Jet Clustering Algorithms," JHEP **9708**, 001 (1997) [arXiv:hep-ph/9707323].

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

[36] M. Cacciari, G. P. Salam and G. Soyez, "FastJet User Manual," Eur. Phys. J. C **72**, 1896 (2012) [arXiv:1111.6097 [hep-ph]].

[37] A. Djouadi, J. Kalinowski and M. Spira, "HDECAY: A Program for Higgs boson decays in the standard model and its supersymmetric extension," Comput. Phys. Commun. **108**, 56 (1998) [hep-ph/9704448].

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

[39] 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].

[40] S. Abachi *et al.* [D0 Collaboration], "Observation of the top quark," Phys. Rev. Lett. **74**, 2632 (1995) [hep-ex/9503003].

[41] I. I. Y. Bigi, Y. L. Dokshitzer, V. A. Khoze, J. H. Kuhn and P. M. Zerwas, "Production and Decay Properties of Ultraheavy Quarks," Phys. Lett. B **181**, 157 (1986).

[42] [ATLAS Collaboration], "Search for Resonances Decaying into Top Quark Pairs Using Fully Hadronic Decays in pp Collisions with ATLAS at sqrt(s) = 7 TeV," ATLAS-CONF-2012-102.

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

[44] G. C. Fox and S. Wolfram, "Observables for the Analysis of Event Shapes in e+ e- Annihilation and Other Processes," Phys. Rev. Lett. **41**, 1581 (1978).

[45] C. Bernaciak, M. S. A. Buschmann, A. Butter and T. Plehn, "Fox-Wolfram Moments in Higgs Physics," Phys. Rev. D **87**, 073014 (2013) [arXiv:1212.4436 [hep-ph]].

[46] M. Galassi *et al.*, GNU Scientific Library Reference Manual. Network Theory, 2009.

[47] `http://en.wikipedia.org/wiki/File:Legendrepolynomials6.svg`, November 2013.

[48] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, "Optimization by Simulated Annealing," Science **220**, 671 (1983).

[49] J.R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann, 1993.

[50] P. Speckmayer, A. Hocker, J. Stelzer and H. Voss, "The toolkit for multivariate data analysis, TMVA 4," J. Phys. Conf. Ser. **219**, 032057 (2010).

[51] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," JCSS **55**, 119 (1997).

[52] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau and A. D. Polosa, "ALPGEN, a generator for hard multiparton processes in hadronic JHEP **0307**, 001 (2003) [arXiv:hep-ph/0206293].

[53] T. Sjostrand, S. Mrenna and P. Z. Skands, "PYTHIA 6.4 Physics and Manual," JHEP **0605**, 026 (2006) [arXiv:hep-ph/0603175].

[54] N. Kidonakis, "Differential and total cross sections for top pair and single top production," arXiv:1205.3453 [hep-ph].

[55] M. Czakon, P. Fiedler and A. Mitov, "The total top quark pair production cross-section at hadron colliders through $O(\alpha_S^4)$," Phys. Rev. Lett. **110**, 252004 (2013) [arXiv:1303.6254 [hep-ph]].

[56] W. Bartel *et al.* [JADE Collaboration], "Experimental Studies on Multi-Jet Production in e+ e- Annihilation at PETRA Energies," Z. Phys. C **33**, 23 (1986).

[57] B. A. Kniehl, G. Kramer and B. Potter, "Fragmentation functions for pions, kaons, and protons at next-to-leading order," Nucl. Phys. B **582**, 514 (2000) [hep-ph/0010289].

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

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

[60] T. Sjostrand, S. Mrenna and P. Z. Skands, "A Brief Introduction to PYTHIA 8.1," Comput. Phys. Commun. **178**, 852 (2008) [arXiv:0710.3820 [hep-ph]].

[61] T. Gleisberg, S. .Hoeche, F. Krauss, M. Schönherr, S. Schumann, F. Siegert and J. Winter, "Event generation with SHERPA 1.1," JHEP **0902**, 007 (2009) [arXiv:0811.4622 [hep-ph]].

[62] W. Beenakker, S. Dittmaier, M. Kramer, B. Plumper, M. Spira and P. M. Zerwas, "NLO QCD corrections to t anti-t H production in hadron collisions," Nucl. Phys. B **653**, 151 (2003); [hep-ph/0211352].

[63] S. Dawson, C. Jackson, L. H. Orr, L. Reina and D. Wackeroth, "Associated Higgs production with top quarks at the large hadron collider: NLO QCD corrections," Phys. Rev. D **68**, 034022 (2003) [hep-ph/0305087].

[64] A. Bredenstein, A. Denner, S. Dittmaier and S. Pozzorini, "NLO QCD Corrections to Top Anti-Top Bottom Anti-Bottom Production at the LHC: 2. full hadronic results," JHEP **1003**, 021 (2010) [arXiv:1001.4006 [hep-ph]].

[65] C. Englert, T. Plehn, P. Schichtel and S. Schumann, "Jets plus Missing Energy with an Autofocus," Phys. Rev. D **83**, 095009 (2011) [arXiv:1102.4615 [hep-ph]].

[66] E. Gerwick, T. Plehn, S. Schumann and P. Schichtel, "Scaling Patterns for QCD Jets," JHEP **1210**, 162 (2012) [arXiv:1208.3676 [hep-ph]].

[67] S. El Hedri, A. Hook, M. Jankowiak and J. G. Wacker, "Learning How to Count: A High Multiplicity Search for the LHC," JHEP **1308**, 136 (2013) [arXiv:1302.1870 [hep-ph]].

Erklärung:

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

Heidelberg, den 30. November 2013 .....