



CLICdp-Pub-2017-002
07 August 2017

Jet reconstruction at high-energy electron-positron colliders

M. Boronat*, J. Fuster*, I. Garcia*¹⁾, Ph. Roloff[†], R. Simoniello[†], M. Vos*

* IFIC (CSIC/UVEG), Valencia, Spain, [†] CERN, Geneva, Switzerland

Abstract

In this paper we study the performance in e^+e^- collisions of classical e^+e^- jet reconstruction algorithms, longitudinally invariant algorithms and the recently-proposed Valencia algorithm. The study includes a comparison of perturbative and non-perturbative jet energy corrections and the response under realistic background conditions. Several algorithms are benchmarked with a detailed detector simulation at $\sqrt{s} = 3$ TeV. We find that the classical e^+e^- algorithms, with or without beam jets, have the best response, but are inadequate in environments with non-negligible background. The Valencia algorithm and longitudinally invariant k_t algorithms have a much more robust performance, with a slight advantage for the former.

This work was carried out in the framework of the CLICdp collaboration

¹⁾Corresponding author: ignacio.garcia@ific.uv.es

1 Introduction

The next large collider facility could be a high-energy electron-positron collider. Linear e^+e^- colliders are the best tool to explore the energy range from several 100 GeV to a few TeV. The Technical Design Report of the International Linear Collider (ILC [1, 2]) project envisages a programme of precision Higgs and top physics at centre-of-mass energies $\sqrt{s} = 250$ GeV, 500 GeV and, after an energy upgrade, 1 TeV. The Compact Linear Collider (CLIC [3]) scheme has been shown to reach accelerating gradients that extend the e^+e^- programme into the multi-TeV regime. CLIC envisages a first phase at $\sqrt{s} = 380$ GeV, followed by its high-energy programme of multi-TeV operation, with stages at 1.5 TeV and 3 TeV [4]. A large circular machine, as envisaged by the FCCee [5] and CEPC [6] projects, could provide high luminosity at 250 GeV. FCCee may reach the top quark pair production threshold [5]. On a longer time scale a muon collider [7] also has the potential to reach the multi-TeV regime.

Measurements of hadronic final states are a key ingredient of the programme of any next-generation lepton collider. Excellent jet reconstruction is essential to characterize the couplings of the Higgs boson and top quark at the sub-percent level. To distinguish hadronic W and Z boson decays a jet energy resolution of approximately 3% is required. The linear collider detector concepts [8, 9] achieve this performance with highly granular calorimeters [10, 11] and particle-flow algorithms [12]. Excellent jet clustering is needed to benefit fully from their potential.

The increase in energy comes with a number of challenges for jet reconstruction. Compared to LEP and SLC, high-energy machines produce an abundance of multi-jet final states, final states with multiple energy scales (in associated production), forward-peaked processes and highly boosted objects. Backgrounds such as $\gamma\gamma \rightarrow \text{hadrons}$ production are increasingly important at high energy [13]. The classical e^+e^- algorithms cannot cope with this environment, in particular with the beam-induced background [3, 14, 15]. A critical evaluation of jet reconstruction at lepton colliders is therefore mandatory.

In this paper we study the performance of jet reconstruction in multi-TeV e^+e^- collisions in detail. We benchmark the performance of a number of sequential recombination algorithms: the classical e^+e^- k_t algorithm [16] used by the LEP experiments and SLD, the longitudinally invariant k_t algorithm [17, 18] used at hadron colliders and a generalization of the Valencia algorithm [19], a robust e^+e^- algorithm.

Several aspects of jet reconstruction performance are studied in simulated events. In a particle-level study we estimate the size of perturbative and non-perturbative corrections to the jet energy. We establish their dependence on the process and the centre-of-mass energy, and on the parameters of the jet algorithms. These particle-level simulations are also used to study the impact of energy deposited on the signal event by background processes. We conclude by using a realistic simulation of the CLIC detector and particle flow reconstruction to study the jet reconstruction performance in top quark pair production and di-Higgs boson production.

The layout of this paper is as follows. In Section 2 we review the challenges of jet reconstruction at high-energy lepton colliders. In Section 3 the jet reconstruction algorithms are introduced. In Section 4 the perturbative and non-perturbative corrections are studied. In Section 5 we study the response of the algorithms to the signal jet and the background. Section 6 presents the results of a full-simulation study of a few benchmark processes. In Section 7 we discuss possible directions for future work and in Section 8 the most important findings of this work are summarized.

2 Challenges for jet reconstruction at high-energy e^+e^- colliders

The experimental environment at previous lepton colliders, such as LEP and SLC, was very benign compared to that at hadron colliders. While this remains true at future high-energy lepton machines, jet reconstruction faces a number of new challenges.

2.1 Multi-jet final states

Future lepton colliders offer the possibility to study $2 \rightarrow 4$, $2 \rightarrow 6$, and even $2 \rightarrow 8$ processes. The dominant branching ratios of the W , Z and Higgs bosons are from hadronic decays. Final states with four jets, most notably $e^+e^- \rightarrow Zh$, play a key role in the physics programme of any future electron-positron collider. At high energy final states with six jets (e.g. $e^+e^- \rightarrow t\bar{t}$), or even eight or ten jets (e.g. $e^+e^- \rightarrow t\bar{t}h$), become important. Imperfect clustering of final-state particles can affect the reconstruction of hadronic decays in an important way.

The impact of incorrect assignments of final-state particles to jets is illustrated with an example. We consider the Higgsstrahlung process with hadronic decays of Z - and Higgs bosons, where reliable reconstruction of Z and h -boson candidates is the key to a precise measurement of the Higgs boson couplings [20]. The hard scattering process $e^+e^- \rightarrow Zh$ is simulated with the MadGraph_aMC@NLO [21] package and the $Z \rightarrow q\bar{q}$ and $h \rightarrow b\bar{b}$ decays and subsequent hadronization with Pythia8 [22]. No beam energy spread, initial-state-radiation, or modelling of background or the detector are included. Stable particles are clustered into jets with the Durham algorithm (exclusive clustering with $N = 4$).

In Figure 1 the invariant mass distribution of the Higgs boson candidates is shown for three centre-of-mass energies. We find that the distribution has a non-zero width, even in this relatively perfect simulation. The finite resolution is purely due to imperfect clustering of final-state particles into jets. The effect of confusion in jet clustering is less pronounced at higher centre-of-mass energy, as the greater boost of the Z - and Higgs bosons leads to a cleaner separation of the jets.

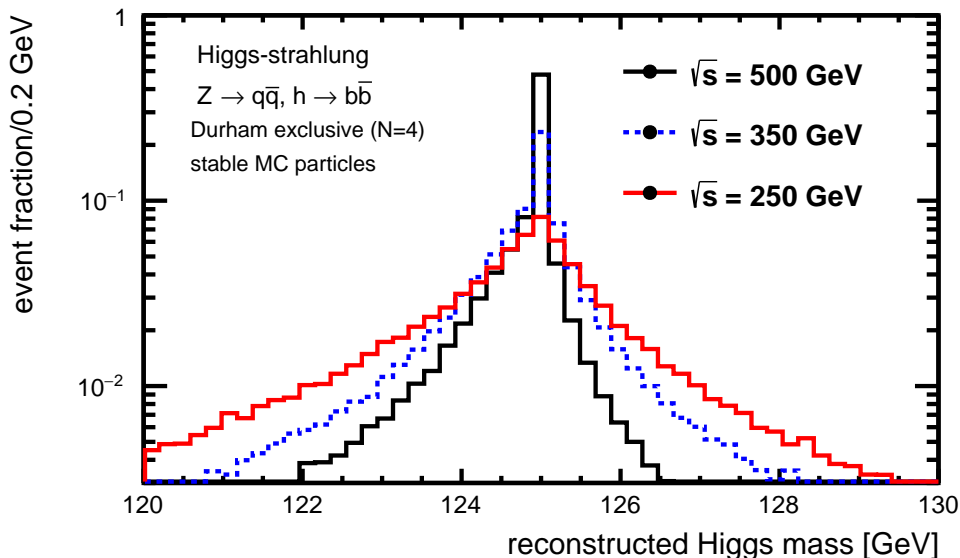


Figure 1: The reconstructed Higgs boson mass peak in $e^+e^- \rightarrow Zh$, $Z \rightarrow q\bar{q}$, $h \rightarrow b\bar{b}$ events. The three histograms correspond to centre-of-mass energies of 250 GeV (red, continuous line), 350 GeV (blue, dashed line) and 500 GeV (black, continuous line).

The most important challenge stems from the larger jet multiplicity. In events with only two jets (i.e. $e^+e^- \rightarrow Zh$ events with $Z \rightarrow \nu\bar{\nu}$ and $h \rightarrow b\bar{b}$) the clustering contribution to the mass resolution is negligible. In the ILC and CLIC analyses of di-Higgs boson [23] and $t\bar{t}h$ production [24, 25] jet clustering is found to be the limiting factor for the Higgs mass resolution.

Finally, we note that as the centre-of-mass energy increases, t -channel processes become more important. The most obvious example is vector-boson fusion production of the Higgs boson. The final-state products at high energy are strongly forward-peaked [26] and special care is needed to ensure robust jet

reconstruction performance over the full polar angle coverage of the experiment. Di-Higgs boson production through vector boson fusion ($e^+e^- \rightarrow \nu\bar{\nu}hh$) presents a double challenge of high jet multiplicity and forward jets. We therefore take this analysis at $\sqrt{s} = 3$ TeV as a benchmark.

2.2 Jet substructure

The production of very energetic gauge bosons, Higgs bosons and top quarks with hadronic decays, collectively denoted as boosted objects, poses a challenge to the experiments at a future high-energy collider. Whenever the energy of the decaying object exceeds its mass m significantly the highly collimated decay products typically cannot be resolved. In such cases, an analysis of the internal structure of jets at a scale² $R < 2m/p_T$ is performed to identify the boosted object [27, 28].

At high-energy linear colliders, the separation of boosted W - and Z -bosons and the reconstruction of boosted top quarks challenge the detector and reconstruction algorithms. The highly granular calorimeters [10, 11] of the Linear Collider detector concepts and the particle flow paradigm are eminently suited for substructure analyses. We analyze the large- R jet mass resolution in top quark pair production at $\sqrt{s} = 3$ TeV, as a first exploration of the jet substructure performance of experiments at future lepton colliders.

2.3 Beam-induced background

While the environment at lepton colliders remains much more benign than the pile-up conditions of high-energy hadron colliders, several background sources cannot be ignored in the detector design and evaluation of the performance of the linear collider experiments. The most relevant background source for jet reconstruction at linear e^+e^- colliders is $\gamma\gamma \rightarrow \text{hadrons}$ production [3]: photons emitted from the incoming electron and positron beams (bremsstrahlung and beamstrahlung) collide and produce *mini-jets* of hadrons³. In high-energy colliders the probability to produce a mini-jet event in a given bunch crossing is of the order of one.

To evaluate the detector performance, ILC and CLIC detector concepts superpose a number of $\gamma\gamma \rightarrow \text{hadrons}$ background events on the hard scattering process. The distribution of the particles formed in $\gamma\gamma$ collisions is forward-peaked, with approximately constant density per unit of rapidity over the instrumented region (a feature also present in the *pile-up* due to *minimum-bias* events in proton-proton collisions). For CLIC at 3 TeV approximately 90% of the energy is deposited in the endcap calorimeters and only 10% in the central (barrel) regions of the experiment [3].

The impact of the background on the performance depends on the bunch structure of the accelerator and the read-out speed of the detector systems. In particular for machines based on radio-frequency cavities operated at room temperature the bunch spacing can be very small; CLIC envisages a bunch spacing of 0.5 ns. Background processes deposit 19 TeV in the detectors during a complete *bunch train* of 312 consecutive bunch crossings at $\sqrt{s} = 3$ TeV [3]. A selection based on the time stamp and transverse momentum of the reconstructed particles reduces this background to approximately 100 GeV on each reconstructed physics event. The relatively large bunch spacing of approximately 500 ns at the ILC allows the detector to distinguish individual bunch crossings. In our full-simulation study simulated $\gamma\gamma \rightarrow \text{hadrons}$ events are overlaid on the signal events.

²At hadron colliders the scale R is usually expressed in terms of the ΔR distance between two objects, defined as $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$, where ϕ and η are the azimuthal angle and pseudo-rapidity

³Other sources, in particular incoherent e^+e^- pair production due to beamstrahlung photons, have a non-negligible impact on the design of the innermost detector elements, but can be ignored in the study of the jet reconstruction performance. A detailed discussion is found in Reference [29].

2.4 Initial State Radiation

In e^+e^- annihilation processes the system formed by the final state products is, to first approximation, produced at rest in the laboratory. Initial state radiation (ISR) photons emitted by the incoming electrons and positrons changes this picture somewhat. To estimate the magnitude of the boost introduced by ISR we generate events using a parton-level calculation⁴ in MadGraph_aMC@NLO [21]. For each $2 \rightarrow 2$ process $e^+e^- \rightarrow XY$ we include also the $2 \rightarrow 3$ process $e^+e^- \rightarrow XY\gamma$. In Fig. 2 the fraction of the energy carried by the XY system is shown for several $2 \rightarrow 2$ processes and for several centre-of-mass energies. For most s -channel processes, the ISR photon energy spectrum falls off very rapidly and the visible energy distribution displays a sharp peak at 1.

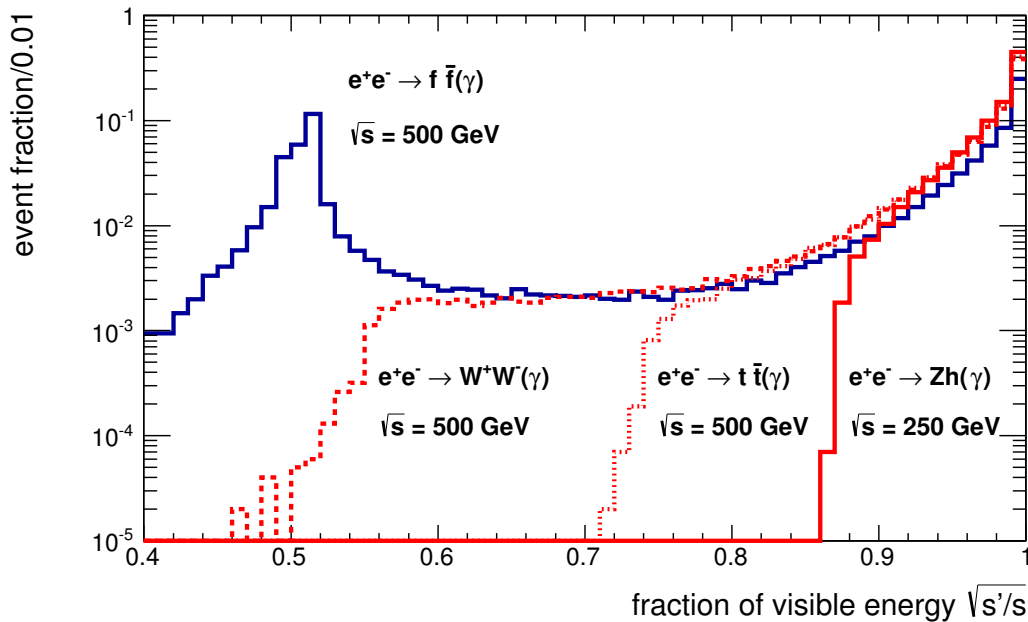


Figure 2: The fraction of the visible energy (the energy carried by all final state products except the photon) in several processes, where a photon radiated off the initial or final-state particles may accompany the final state products. The distributions correspond to pair production of light fermions in association with a photon ($e^+e^- \rightarrow f\bar{f}(\gamma)$), W -boson and top-quark pair production ($e^+e^- \rightarrow W^+W^-(\gamma)$, $e^+e^- \rightarrow t\bar{t}(\gamma)$) and the Higgsstrahlung process $e^+e^- \rightarrow Zh(\gamma)$. The centre-of-mass energy is indicated on the figure for each process.

The boost of the system along the z -axis due to ISR remains relatively small close to the production threshold: for $e^+e^- \rightarrow Zh(\gamma)$ at $\sqrt{s} = 250$ GeV and $e^+e^- \rightarrow t\bar{t}(\gamma)$ at 500 GeV $\beta_z = v_z/c$ is smaller than 0.1 in over 95% and 90% of the events, respectively⁵. For processes with a cross section that grows with \sqrt{s} the peak is even narrower. The role of ISR is only significant for radiative return to the Z in the process $e^+e^- \rightarrow f\bar{f}(\gamma)$, where f is any fermion with mass less than half that of the Z boson.

At linear colliders with very narrow beams the luminosity spectrum displays a sizeable tail towards lower centre-of-mass energy [30]. Beam energy spread and beamstrahlung may therefore lead to a pronounced boost of the visible final state objects in a small fraction of events. This effect is included in the full simulation study of Section 6.

⁴The effects of beam energy spread and beamstrahlung are not included.

⁵Compared to hadron colliders this boost is very small indeed: for di-jet production at the LHC β_z of the di-jet system is very close to 1 and even a massive system such as a top quark pair acquires a typical $\beta_z = 0.5$.

3 Jet reconstruction algorithms

In this section the jet algorithms considered in this paper are introduced. We discuss the most important differences and their implications for the performance. Three classes of sequential recombination algorithms are considered here:

- The classical e^+e^- algorithms [16] and their generalization with beam jets [31]
- The longitudinally invariant algorithms developed for hadron colliders [17, 18]
- The Valencia algorithm proposed in a previous publication [19], which is further generalized here.

3.1 The VLC algorithm

In Ref. [19] a robust jet reconstruction algorithm was proposed, that maintains a Durham-like distance criterion based on energy and polar angle. It achieves a background resilience that can compete with the longitudinally invariant k_t algorithm. Here, we further generalize the definition of the algorithm.

The VLC algorithm has the following inter-particle distance:

$$d_{ij} = 2 \min(E_i^{2\beta}, E_j^{2\beta})(1 - \cos \theta_{ij})/R^2, \quad (1)$$

where R is the radius or resolution parameter. For $\beta = 1$ the distance is given by the transverse momentum squared of the softer of the two particles relative to the harder one, as in the Durham algorithm.

The beam distance of the algorithm is:

$$d_{iB} = E_i^{2\beta} \sin^{2\gamma} \theta_{iB}, \quad (2)$$

where θ_{iB} is the angle with respect to the beam axis, i.e. the polar angle.

The two parameters β and γ allow independent control of the clustering order and the background resilience⁶. The β and γ parameters are real numbers that can take any value. For $\beta = \gamma = 1$ the expression for the beam distance simplifies to $d_{iB} = E^2 \sin^2 \theta_{iB} = p_{Ti}^2$. We discuss the impact of different choices in Sec 3.3.

This new version of the algorithm fulfils the standard IR-safety tests of the FastJet team. The VLC algorithm is available as a plug-in for the FastJet [31, 32] package. The code can be obtained from the ‘‘contrib’’ area [33].

3.2 Comparison of the distance criteria

The generalized distance criteria for three families of algorithms are summarized in Tab. 1.

For all algorithms the clustering order can be modified by an appropriate choice of the n in the exponent of the energy (or p_T) in the inter-particle distance (β in the VLC algorithm). The Durham (or k_t) algorithm, with $n = 1$, clusters pairs of particles starting with those that are soft and collinear (i.e. the inverse of the virtuality-ordered emission during the parton shower). Choosing $n = 0$ yields the Cambridge/Aachen algorithm, that has a purely angular distance criterion. The anti- k_t algorithm has $n = -1$.

In the generalized algorithms the area of jets is limited. Any particle with a *beam distance* [35] smaller than the distance to any other particle is associated with the beam jet, and therefore not considered part of the visible final state. The radius parameter R governs the relative size of the interparticle and

⁶The first version of the algorithm [19] had a single parameter β . Equation 1 furthermore differs by a factor two from the inter-particle distance of Ref. [19]. To distinguish the two algorithms we refer to the more general expression as the VLC algorithm, while the name Valencia is reserved for the setting $\beta = \gamma$ that recovers the first proposal (adjusting R by factor $\sqrt{2}$).

Table 1: Summary of the distance criteria used in sequential recombination algorithms. Generalized inter-particle and beam distances are given for three main classes of algorithms: the classical e^+e^- algorithms (comprising a version with beam jets of the Cambridge [34] and Durham algorithms), the longitudinally invariant algorithms used at hadron colliders, which comprise longitudinally invariant k_t , Cambridge-Aachen and anti- k_t and the robust e^+e^- algorithms introduced in Section 3.1.

algorithm	generalized e^+e^-	longitudinally invariant	robust e^+e^-
distance d_{ij}	$2 \min(E_i^{2n}, E_j^{2n}) \frac{1 - \cos \theta_{ij}}{1 - \cos R}$	$\min(p_{T,i}^{2n}, p_{T,j}^{2n}) \frac{\Delta R_{ij}^2}{R^2}$	$2 \min(E_i^{2\beta}, E_j^{2\beta}) \frac{1 - \cos \theta_{ij}}{R^2}$
beam distance d_{iB}	E_i^{2n}	$p_{T,i}^{2n}$	$E_i^{2\beta} \sin^{2\gamma} \theta_{iB}$

beam distance and thus determines the size of the jet. This modification renders jet reconstruction very resilient to backgrounds.

The generalized e^+e^- algorithm and the VLC algorithm have virtually the same inter-particle distance. However, the radius parameter R is redefined: the inter-particle distance d_{ij} denominator is R^2 instead of $1 - \cos R$. The hadron collider algorithms replace the particle energy E_i and angle θ_{ij} with quantities that are invariant under boosts along the beam axis, the transverse momentum $p_{T,i}$ and the distance $\Delta R_{ij} = \sqrt{(\Delta\phi)^2 + (\Delta y)^2}$, where ϕ is the azimuthal angle in the usual cylindrical coordinates and y denotes the rapidity.

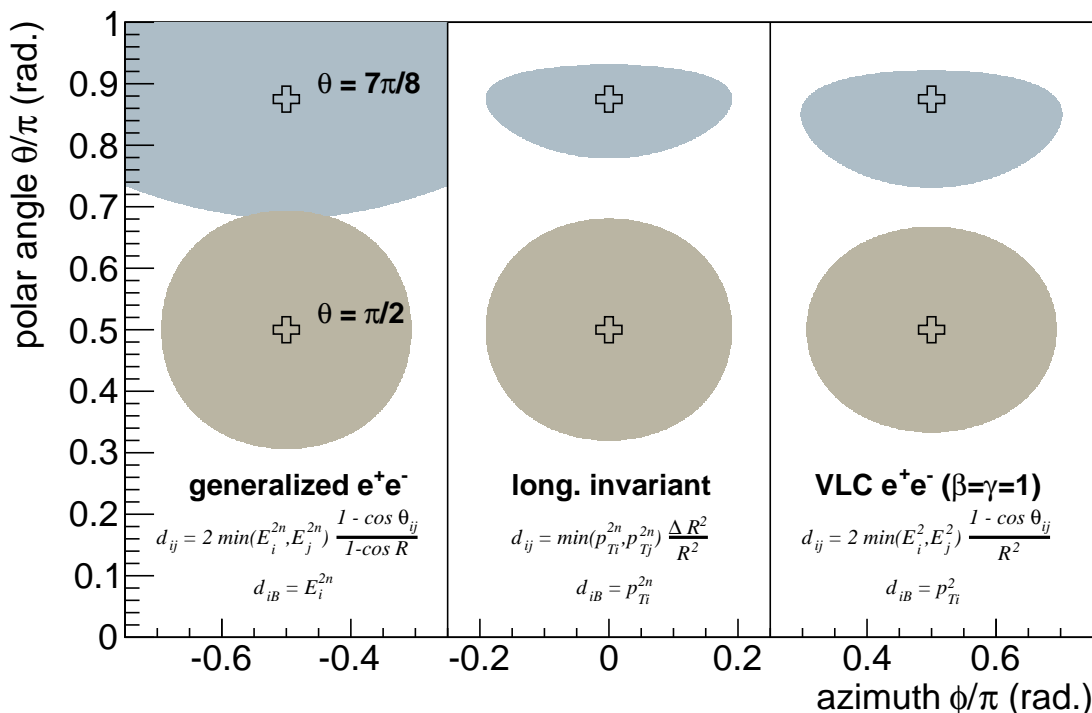


Figure 3: The area or *footprint* of jets reconstructed with $R = 0.5$ with the three major families of sequential recombination algorithms. The two shaded areas in each column correspond to a jet in the central detector ($\theta = \pi/2$) and to a forward jet ($\theta = 7\pi/8$). The jet axis is indicated with a cross.

Detailed studies [3, 14] show that the longitudinally invariant k_t algorithm is much more resilient

to the $\gamma\gamma \rightarrow \text{hadrons}$ background than the classical and generalized e^+e^- algorithms.

For each of the algorithms the catchment areas of a single central and forward jet with $n=1$ and $R=0.5$ are indicated in Figure 3. The footprint of the central jet (at $\theta = \pi/2$) is approximately circular for all algorithms. The area of the jet in the forward detector (at $\theta = 7\pi/8$) shrinks considerably for the longitudinally invariant algorithms and the VLC algorithm. The reduced exposure in this region, where backgrounds are most pronounced, is the crucial feature for the enhanced resilience of these algorithms.

An analytical understanding of this property can be obtained by considering two test particles with energies E_i and E_j and separated by a fixed angle Ω_{ij} . For the generalized e^+e^- algorithms, both the distance d_{ij} between the two particles and the ratio d_{ij}/d_{iB} of the inter-particle distance and the beam distance are independent of polar and azimuthal angle. For the longitudinally invariant algorithms the ratio d_{ij}/d_{iB} increases (while the inter-particle distance decreases) as the two-particle system is rotated into the forward region. Finally, for the VLC algorithm the ratio d_{ij}/d_{iB} increases as $1/\sin^{2\gamma}\theta$ in the forward region, with a slope that is similar to that of the longitudinally invariant algorithms for $\gamma = 1$. The distance d_{ij} is constant, as in classical e^+e^- algorithms.

A closer comparison of the shape of the footprint of the longitudinally invariant algorithms and the VLC algorithm show that, given identical jet axes, the former extend further into the forward region. This causes a slight difference in background resilience of both classes of algorithms.

3.3 Interpolation between algorithms

The two parameters β and γ of the VLC algorithm allow one to tailor the algorithm to a specific application. As these parameters are real numbers, one can interpolate smoothly between different clustering schemes.

The β -parameter that exponentiates the energy in inter-particle and beam distance governs the clustering order (similar to the exponent n in the generalized k_t algorithm). For $\beta = 1$ clustering starts with soft, collinear radiation. Choosing $\beta = 0$ yields purely angular clustering, while $\beta = -1$ corresponds to clustering starting from hard, collinear radiation. These integer choices of β correspond to k_t , Cambridge/Aachen and anti- k_t clustering, respectively. Non-integer values of β interpolate smoothly between these three schemes.

The parameter γ in the exponent of the beam distance of the VLC algorithm provides a handle to control the shrinking of the jet catchment area in the forward regions of the experiment. After setting the R -parameter to the optimal value for central jets, the area of forward jets can be *tuned* by the choice of γ to ensure the required background resilience.

We have seen that $\gamma = 1$ yields forward jets with a similar size of those of the longitudinally invariant algorithms for hadron colliders⁷. Values of γ greater than 1 further enhance the rise of the $\frac{d_{ij}}{d_{iB}}$ ratio in the forward region, causing the jet footprint to shrink faster. Values between 0 and 1 yield a slower decrease of the area when the polar angle goes to 0 or π .

For $\gamma = 0$, $d_{iB} = E_i^{2\beta}$ and we retrieve the generalized e^+e^- algorithms with constant angular opening: the generalized Cambridge algorithm [34] for $\beta = 0$ and generalized k_t or Durham [16] for $\beta = 1$. Choosing $\beta = -1$ yields an e^+e^- variant of the anti- k_t algorithm [36]. A schematic overview of the algorithms in (β, γ) space is given in Figure 4.

⁷In both algorithms the ratio d_{ij}/d_{iB} for two test particles separated by a constant angle depends on the polar angle of the system. In the VLC algorithm the ratio is proportional to $\sin^{-2\gamma}\theta$. In longitudinally invariant algorithms the ratio follows the evolution of the pseudo-rapidity and grows approximately as $\eta^2 = \log^2 \tan \theta/2$ in the forward region. Both gives rise to qualitative the same behaviour.

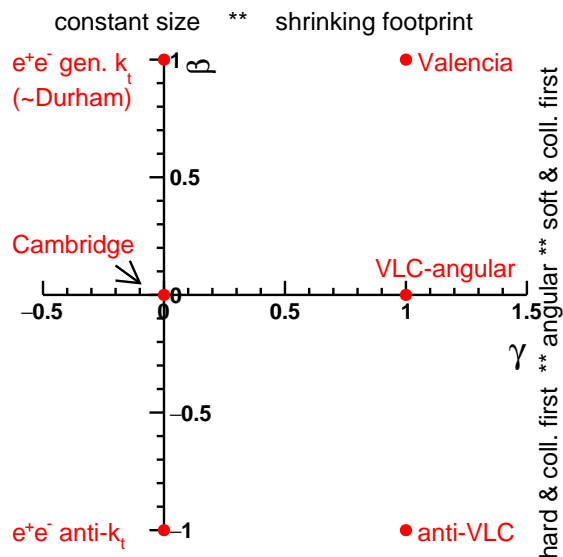


Figure 4: Diagram of the parameter space spanned by exponents β and γ of the VLC algorithm. On the y-axis generalizations with beam jets of the LEP/SLD algorithms are found, with the Cambridge algorithm with angular ordering at the origin and the Durham or k_t algorithm at $\beta = 1$. Choosing $\beta = -1$ yields reverses the clustering order (like in anti- k_t algorithm [36]). Choosing non-zero and positive values for γ yields robust algorithms with a shrinking jet area in the forward region.

4 Jet energy corrections

Before we turn to a detailed simulation including overlaid backgrounds and a model for the detector response, we study the perturbative and non-perturbative jet energy corrections of the algorithms. Both types of corrections are closely connected to the jet area [37]. In this Section we quantify their impact, following the analysis of Ref. [37]. This first exploration of the stability of the algorithms should be extended in future work to quantify the impact of next-to-leading correction, as performed for instance in Ref. [38]. Also the robustness of the conclusions for a variety of different sets of parameters (*tunes*) of the Monte Carlo simulation merits further study.

4.1 Monte Carlo setup

The Monte Carlo simulation chain uses the MadGraph5_aMC@NLO package [21] to generate the matrix elements of the hard scattering $2 \rightarrow 2$ event. Several processes are studied, but results in this Section focus on $e^+e^- \rightarrow q\bar{q}$ at $\sqrt{s} = 250$ GeV and $e^+e^- \rightarrow t\bar{t}$ with fully hadronic top decays at $\sqrt{s} = 3$ TeV. The four-vectors of the outgoing quarks are fed into Pythia 8.180 [22], with the default tune to LEP data, that performs the simulation of top quark and W boson decays, the parton shower and hadronization. No detector simulation is performed and initial state radiation and beam energy spread are not included in the simulation. Particles or partons from the Pythia event record are clustered using FastJet 3.0.6 [31] exclusive clustering with $N = 2$. The default (“E-scheme”) recombination algorithm is used to merge (pseudo-) jets.

4.2 Definition of response and resolution

The jet energy and mass distributions often display substantial non-Gaussian tails and the choice of robust estimators has non-trivial implications. To estimate the centre of the distribution (i.e. the response) the mean and median are used. In some cases we present both, to give an indication for the skewness of the

distribution. The width of the distribution (resolution) is estimated using the inter-quantile range IQR_{34} , that measures half the width of the interval centered on the median that contains 68% of all jets. We also use RMS_{90} , the root-mean-square of the values after discarding 5% outliers in both the low and high tails of the distribution.

4.3 Perturbative corrections

Following Reference [37] we estimate the total energy correction by comparing the parton from the hard scatter to the jet of stable particles. For jets of finite size this correction is dominated by energy that *leaks* out of the jet. We indeed find that the distribution of the difference of parton and jet energy is asymmetric, with a long tail towards negative corrections, where the parton energy is larger than the energy captured in the jet. This energy leakage is most pronounced for jets with a small radius parameter, as expected.

In Figure 5 the average (dashed line) and median (continuous line) relative energy correction are presented. The left plot corresponds to $e^+e^- \rightarrow q\bar{q}$ collisions at relatively low energy ($\sqrt{s} = 250$ GeV), while the right plot corresponds to $e^+e^- \rightarrow t\bar{t}$ at $\sqrt{s} = 3$ TeV. At a quantitative level the results show some dependence on the process, centre-of-mass energy and the generator tune for which they are obtained, but qualitatively the same pattern emerges in all cases. The energy correction decreases as the catchment area of the jet increases.

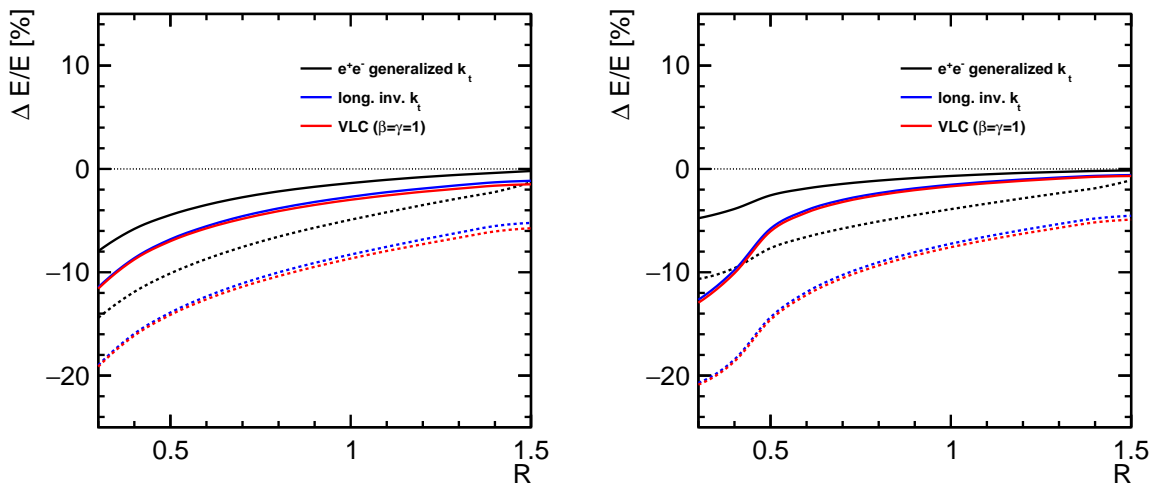


Figure 5: Jet energy correction as a function of the jet radius parameter R in $e^+e^- \rightarrow q\bar{q}$ production at $\sqrt{s} = 250$ GeV(left panel) and $e^+e^- \rightarrow t\bar{t}$ production at $\sqrt{s} = 3$ TeV(right panel). The continuous line corresponds to the median relative correction, the dashed line to the mean. Results are shown for three algorithms: the generalized e^+e^- algorithm, the longitudinally invariant k_t algorithm and the VLC algorithm with $\beta = 1$. The statistical uncertainties of the results are negligible and are not indicated.

The energy corrections for the generalized e^+e^- algorithm vanish relatively rapidly, with the median correction reaching sub-% level for $R \sim 1$. The VLC and longitudinally invariant k_t algorithm show much slower convergence towards zero correction. This is entirely due to jets close to the beam axis. For central jets the three classes of algorithms yield identical results (within the statistical accuracy). The VLC and k_t algorithms have similar footprints and, indeed, very similar energy corrections.

The clustering order (as controlled by n in the generalized algorithm and by β in the VLC algorithm) has a minor impact on the energy corrections. The (inclusive) Cambridge/Aachen algorithm and anti- k_t algorithm give similar results to the k_t variants of the same algorithm shown here.

4.4 Non-perturbative corrections

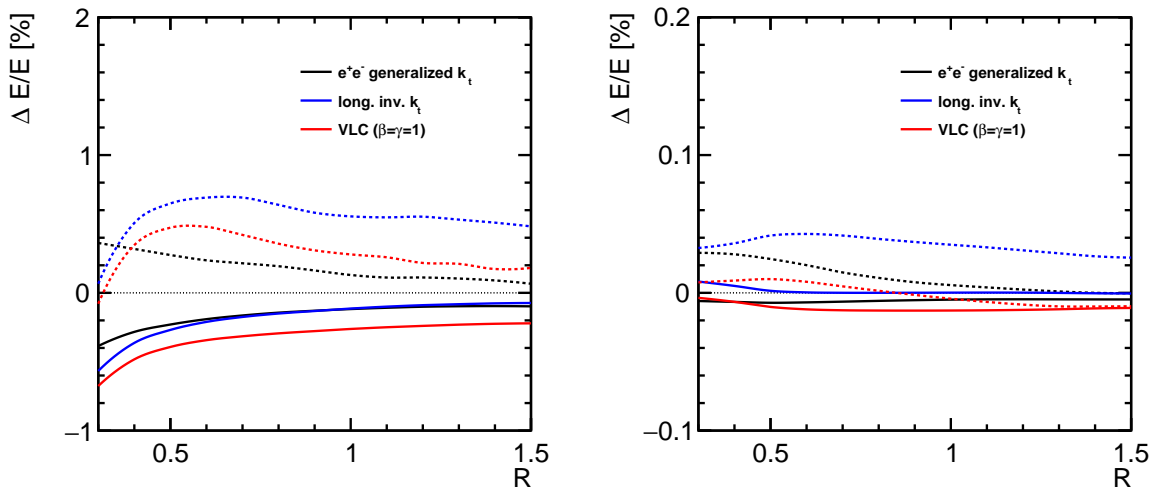


Figure 6: Non-perturbative jet energy corrections to the jet energy as a function of the jet radius parameter R in $e^+e^- \rightarrow q\bar{q}$ production at $\sqrt{s} = 250$ GeV (left panel) and $e^+e^- \rightarrow t\bar{t}$ production at $\sqrt{s} = 3$ TeV (right panel). The continuous line corresponds to the median relative correction, the dashed line to the mean. Results are shown for three algorithms: the generalized e^+e^- algorithm, the longitudinally invariant k_t algorithm and the VLC algorithm with $\beta = 1$. The statistical uncertainties of the results are negligible and are not indicated.

The largest part of the jet energy correction due to the finite size is amenable to perturbative calculations. A small residual correction is related to the hadronization and must be extracted from (or tuned to) data. The non-perturbative energy correction is estimated as the difference between the energy of the parton-level jet, clustering all partons before hadronization, and the jet reconstructed from stable final-state particles. The difference in energy between the parton-level and particle-level jet is typically small, but the distribution is offset from 0 and has a long asymmetric tail. Mean and median are again different and even have opposite signs.

The dependence of this correction on R is shown in Figure 6. The non-perturbative part is very small compared to the total correction. It is well below 1% at $\sqrt{s} = 250$ GeV, for any value of R studied here. For high-energy collisions the correction is well below the per mille level. The generalized e^+e^- algorithm again has the best convergence, while for both VLC and longitudinally invariant k_t the median or mean remain sizeable even for $R = 1.5$.

4.5 Jet mass corrections

The previous discussion has focussed on the jet energy response. Corrections to other jet properties may also be important. Here, we study the corrections to the jet mass⁸, which can be taken as a proxy for the substructure of the jet. The non-perturbative jet mass correction is defined (analogously to the non-perturbative energy correction) as the difference between the masses of the parton-level and particle-level jet.

The dependence on the radius parameter R is shown in Figure 7. The non-perturbative contribution to the jet mass is quite large. The relative correction can be several tens of % at low energy up to $R =$

⁸The jet mass is defined as the invariant mass formed by the vector-sum of the momenta of the (massless) jet constituents.

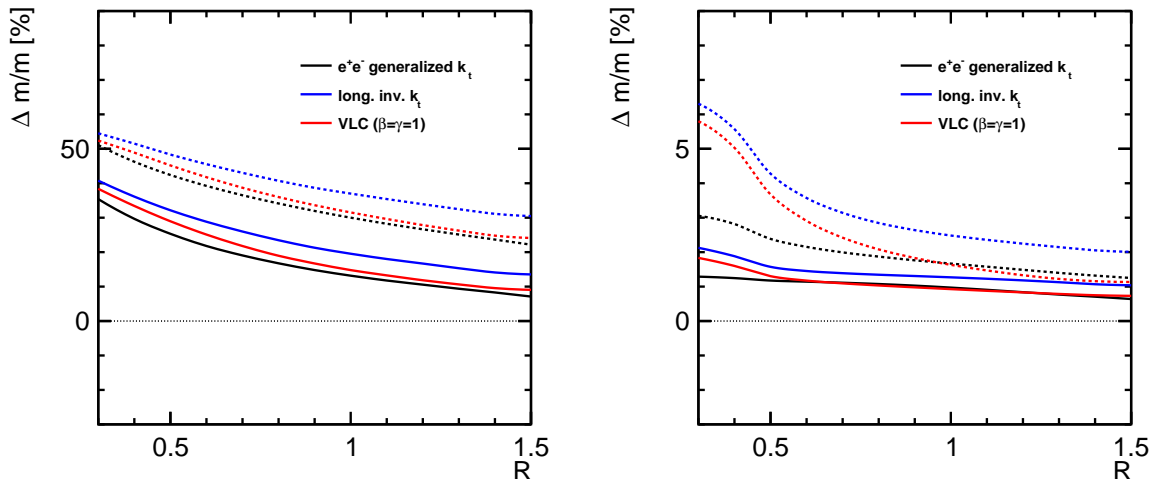


Figure 7: Non-perturbative corrections to the jet mass as a function of the jet radius parameter R in $e^+e^- \rightarrow q\bar{q}$ production at $\sqrt{s} = 250$ GeV (left panel) and $e^+e^- \rightarrow t\bar{t}$ production at $\sqrt{s} = 3$ TeV (right panel). The continuous line corresponds to the median relative correction, the dashed line to the mean. Results are shown for three algorithms: the generalized e^+e^- algorithm, the longitudinally invariant k_t algorithm and the VLC algorithm with $\beta = 1$. The statistical uncertainties of the results are negligible and are not indicated.

1. It drops to a few % for $\sqrt{s} = 3$ TeV. In this case, the algorithms with the e^+e^- inter-particle distance (generalized Durham and VLC) converge slightly faster than the longitudinally invariant algorithms.

5 Particle-level results

In this Section the response of several algorithms is studied on simulated $e^+e^- \rightarrow t\bar{t}$ events at $\sqrt{s} = 3$ TeV. Clustering is exclusive, with $N = 2$. Both highly boosted top quarks are reconstructed as a single, large- R jet. We gain insight in the impact of the background by superposing randomly distributed background on the signal events. To this end the Monte Carlo setup described in Section 4.1 is extended with a simple mechanism to superpose a random energy flow on the signal event.

5.1 Jet energy response without background

Before we study the impact of the background, the response of the jet algorithms to the signal event is estimated. The energy response is determined as the median reconstructed energy. The mass distribution has a sharp peak at the top quark mass and a long tail towards larger masses. The mass response is therefore estimated as the mean reconstructed jet mass. In both cases the response of the Durham algorithm – which clusters all final state particles into the jets – is taken as a reference. The reconstructed energy is divided by 1.5 TeV, the reconstructed jet mass by the average jet mass of ~ 370 GeV.

In Figure 8 the energy and mass response is shown as a function of polar angle for three algorithms: the generalized $e^+e^- k_t$ algorithm (black), the longitudinally invariant k_t algorithm (blue dashed) and VLC with $\beta = \gamma = 1$ (red). The R -parameter is set to 1.5 for all three algorithms. The generalized $e^+e^- k_t$ algorithm recovers over 99.9% of the top quark energy for $R = 1.5$, independent of the jet polar angle. The shrinking jet areas in the forward region of the longitudinally invariant k_t and VLC algorithms lead to a slightly smaller response for $|\cos\theta| > 0.6$. The polar angle dependence of longitudinally invariant

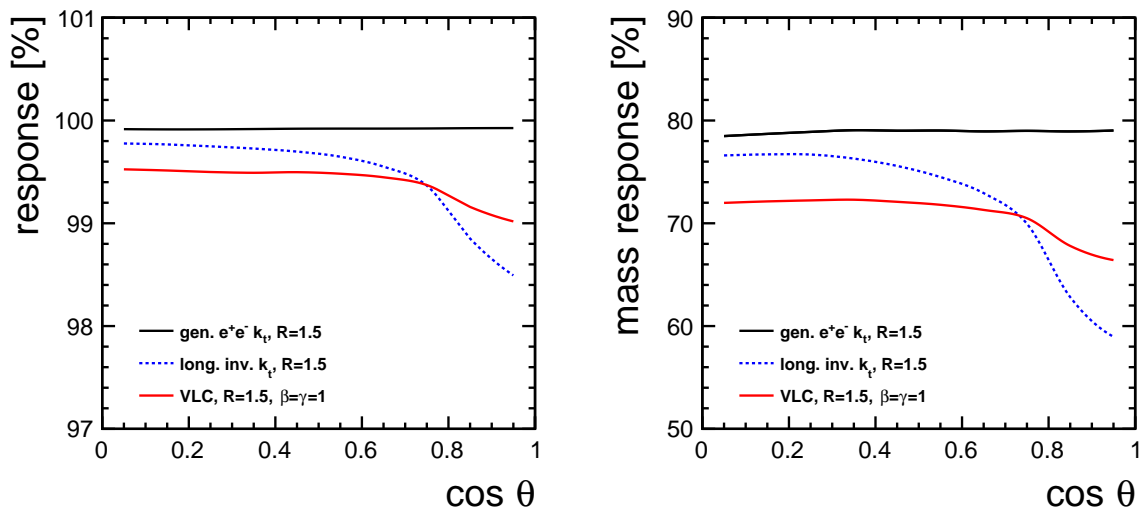


Figure 8: The response to 1.5 TeV top jets as a function of polar angle for three jet algorithms, all with radius parameter $R = 1.5$. The left plot shows the median reconstructed jet energy, the right plot the mean jet mass. Both quantities are normalized to the response of the Durham algorithm: $E = 1.5$ TeV, $m \sim 370$ GeV.

k_t is more pronounced.

The mass response of all three algorithms is substantially lower than for the Durham algorithm. The generalized $e^+e^- k_t$ algorithm has a flat response at nearly 80%. The VLC and longitudinally invariant k_t algorithms display the same pattern as for the energy response: VLC starts off with a lower response in the central region, but the response is much flatter versus polar angle.

5.2 Jet energy response with background

To gain insight in the performance in a more realistic environment with background, we overlay two hundred 1 GeV particles on each signal event. The background distribution is strongly peaked in the forward direction following an exponential distribution peaked at $\theta = 0$, an approximation to the $\gamma\gamma \rightarrow \text{hadrons}$ background in energy-frontier electron-positron colliders (a more realistic simulation of this background follows in Section 6.1).

The two event displays in Figure 9 provide a zoom image of the $\theta - \phi$ plane for a single event. The location of the jet axis is indicated as a red circle. The approximate catchment area of both jets is shown in grey. The green squares represent particles from the top decay that are associated with each jet, the blue squares to background particles clustered into the jet. Both algorithms find a very similar jet axis, centered on the high-energy core of the jet. However, the algorithms have quite distinctive footprints. The longitudinally invariant algorithms expose a larger area in the forward region, which renders it more vulnerable to background in this region.

A quantitative view is obtained by comparing the energy and mass of jets obtained when clustering the same events with and without background particles. The bias (the average difference) in the jet energy and jet mass is shown in Figure 10. The background leads to a significant bias for forward jets reconstructed with the longitudinally invariant k_t algorithm. The VLC algorithm, on the other hand, is only affected in the very forward region and the bias is much less pronounced.

The jet mass is known to be quite sensitive to soft and diffuse radiation, with the contribution scaling as the third power of the jet area [39]. We indeed find that the mass is strongly affected. A comparison of the jet reconstruction performance of the same process in a fully realistic environment is presented in

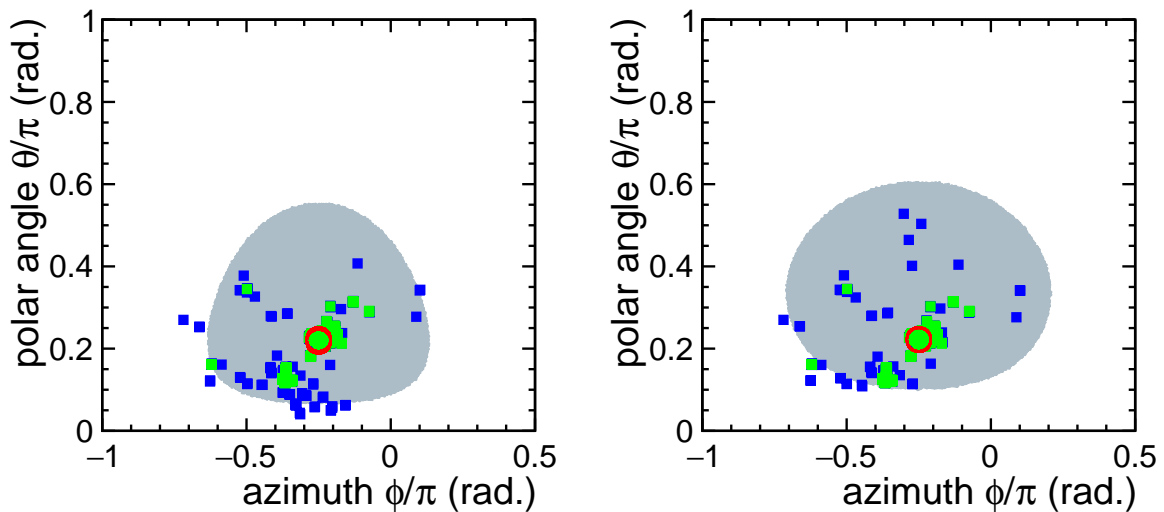


Figure 9: Event display for a $e^+e^- \rightarrow t\bar{t}$ event at $\sqrt{s} = 3$ TeV. The left panel shows the result of clustering with the longitudinally invariant k_t algorithm with $R = 1.2$, the right panel the corresponding VLC jet with the same radius parameter. The image zooms in on the $\theta - \phi$ area around one of the top jets. The location of the jet axis is indicated as a red circle. The area where the distance to the jet axis is smaller than the radius parameter (ΔR_{iC} for longitudinally invariant k_t , $1 - \cos \theta_{iC}$ for VLC) is indicated by the shaded region. The green squares represent particles from the top decay that are associated with each jet, the blue squares to background particles clustered into the jet.

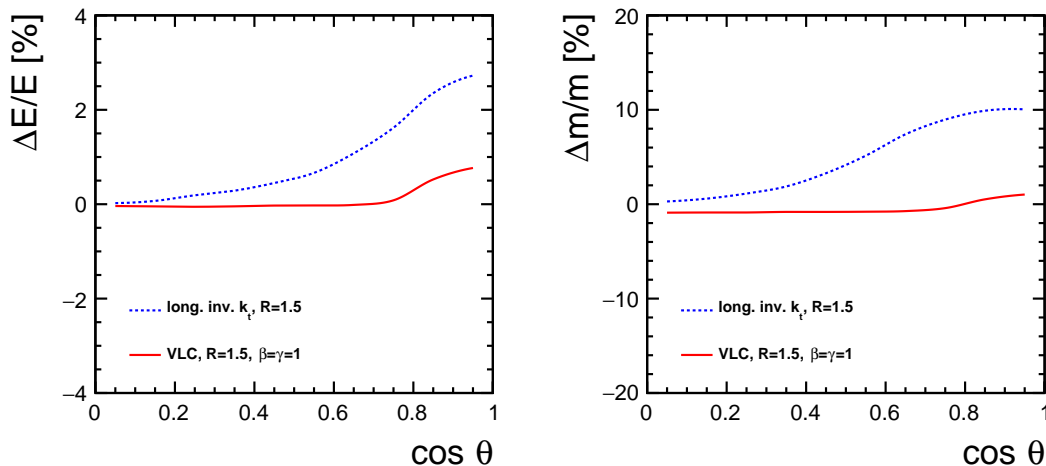


Figure 10: The average contribution to the jet energy (left plot) and jet mass (right plot) of 200 GeV of forward-peaked background.

Section 6.3.

6 Results from full simulation

The performance of the different algorithms is compared in full-simulation samples. We choose two benchmark scenarios with fully hadronic final states that challenge jet reconstruction: di-Higgs boson

production (with $h \rightarrow b\bar{b}$, i.e. a final state with four b-jets) and $t\bar{t}$ production. Both analyses are performed at $\sqrt{s} = 3$ TeV, with the CLIC_ILD detector and a realistic background of $\gamma\gamma \rightarrow \text{hadrons}$.

6.1 Monte Carlo setup

The studies in this Section are performed on CLIC 3 TeV Monte Carlo samples. Events are generated with WHIZARD [40] (version 1.95). The response of the CLIC_ILD detector [8] is simulated with GEANT4 [41]. Multi-peripheral $\gamma\gamma \rightarrow \text{hadrons}$ events are generated with Pythia and superposed as *pile-up* on the signal events.

At CLIC bunches are spaced by 0.5 ns and detector systems are expected to integrate the background of a number of subsequent bunch crossings. In this study, the background corresponding to 60 bunch crossings is overlaid. In the event reconstruction, the information of the tracking system and the calorimeters is combined to form particle-flow objects (PFO) with the Pandora [42] algorithm. Timing cuts on PFOs reduce the background level, with a very small impact on the signal energy flow. The *nominal* (or *default*) selection of Ref. [3, 14] reduces the 19 TeV of energy deposited in the calorimeters by the entire bunch train to approximately 200 GeV superposed on a reconstructed events. A more stringent set of cuts, referred to as the *tight* selection in Ref. [3], reduces the background energy by another factor of two. Both scenarios are studied in the following.

The event simulation and reconstruction of the large data samples used in this study was performed using the ILCDIRAC [43, 44] grid production tools.

6.2 Higgs pair production

The study of Higgs boson pair production is crucial to assess the strength of the Higgs self-coupling. The analysis is very challenging at both hadron and lepton colliders due to the very small cross section. At an e^+e^- collider, the significance of this signal is enhanced at large centre-of-mass energy, as the production rate in the vector-boson-fusion channel $e^+e^- \rightarrow \nu\bar{\nu}hh$ grows strongly with centre-of-mass energy. In this section we focus on events where both Higgs bosons decay to hadrons, through the dominant $h \rightarrow b\bar{b}$ decay of the SM Higgs boson. This final state can be isolated [45] provided the four jets are reconstructed with excellent energy resolution.

The challenge of this measurement lies in the fact that both Higgs bosons are typically emitted at small polar angle [26]. The most frequently observed topology has both Higgs bosons emitted in opposite directions: one in the forward direction and the other in the backward direction. At 3 TeV (at least) one of the Higgs bosons is emitted with $|\cos\theta| > 0.9$ in approximately 85% of events. In this area of the detector the background level due to $\gamma\gamma \rightarrow \text{hadrons}$ production is most prominent.

Despite the large centre-of-mass energy, the Higgs bosons are produced with rather moderate energy: in 3 TeV collisions the most probable energy of the Higgs bosons is approximately 200 GeV, with a long, scarcely populated tail extending to 1.5 TeV. The modest Higgs boost is sufficient for the b-quarks to continue in the same hemisphere as their parent Higgs boson, but it is rarely large enough for the Higgs boson to form a single jet.

We perform exclusive jet reconstruction with $N_{jets} = 4$. Higgs boson candidates are reconstructed by pairing two out of the four jets. The combination is retained that yields the best di-jet masses (i.e. that minimizes $\chi^2 = (m_{ij} - m_h)^2 + (m_{kl} - m_h)^2$, where m_{ij} and m_{kl} are the masses of the two di-jet systems and $m_h = 126$ GeV is the nominal Higgs boson mass used in the simulation).

The distribution of the reconstructed mass of both di-jet systems forming the Higgs boson candidates is shown in Figure 11. The results of two algorithms are shown, both with the radius parameter R set to 1.3. The red line denotes the result of the VLC algorithm with $\beta = \gamma = 1$, the black line that of the longitudinally invariant k_t algorithm. Numerical results of the centre and width of the reconstructed di-jet mass distribution are presented in Table 2. The response of both algorithms is found to agree to within 0.5%, for all methods to estimate the central value of the distribution. The Higgs mass resolution

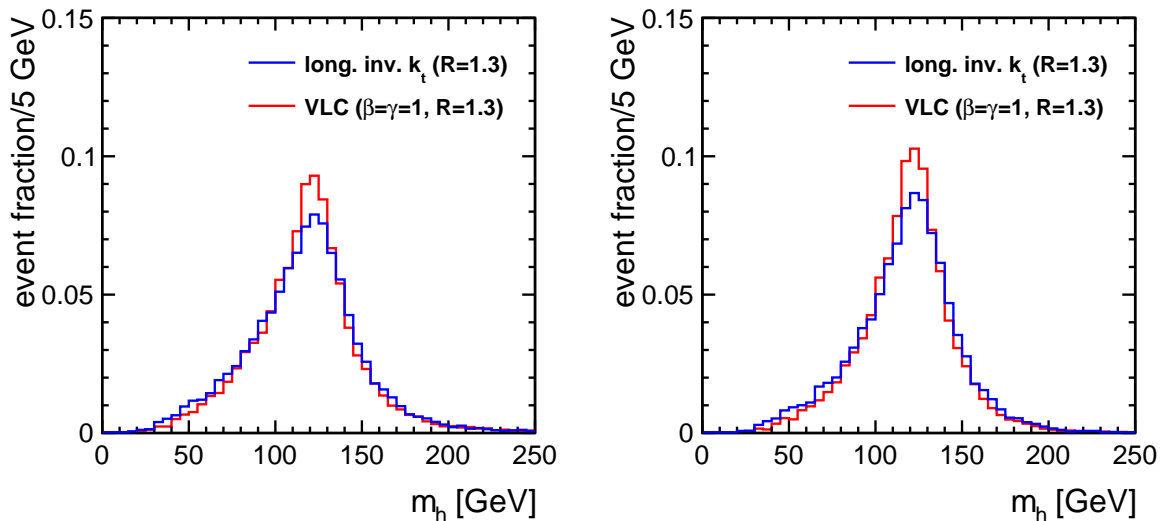


Figure 11: The reconstructed di-jet mass distribution for fully hadronic decays of $e^+e^- \rightarrow \nu\bar{\nu}hh$, $h \rightarrow b\bar{b}$ events at a 3 TeV CLIC. In the left panel all Higgs boson candidates are included, in the right panel only those that match onto exactly two b-quarks from Higgs boson decay. The nominal level of $\gamma\gamma \rightarrow \text{hadrons}$ background is overlaid on the signal. Particle flow objects are selected using the tight selection.

obtained with the VLC algorithm is better for both figures of merit. The IQR_{34} divided by the median yields 22.6% for the VLC algorithm versus 25.4% for k_t .

The dependence of the IQR_{34} resolution on the radius parameter and the parameter γ of the VLC algorithm is shown in Fig. 12. The best mass resolution is obtained for large values of R in both algorithms. The choice of $R \sim 1.3$ is close to optimal for both algorithms. Variation of the γ parameter, that controls the evolution of the VLC jet area in the forward region, leads to a shift of the optimal value of R . With $\gamma < 1$ the jet area is reduced at a slightly slower rate and the best resolution is obtained for smaller R . Choosing $\gamma > 1$ the jet area shrinks more rapidly and a larger R is required to capture the complete energy flow.

6.3 Top quark pair production

The second benchmark we analyze is pair production of boosted top quarks in multi-TeV operation of the CLIC e^+e^- collider. At these energies the top quark decay products are so collimated that hadronic top quarks can be reconstructed as a single large- R top-jet ($R \sim 1$). Only the fully hadronic final state $e^+e^- \rightarrow t\bar{t} \rightarrow b\bar{b}q\bar{q}'q''\bar{q}'''$ is considered. Events where either the top or anti-top quark is emitted in the forward or backward direction ($|\cos\theta| > 0.7$) are discarded to avoid the incomplete acceptance in that region. To cope with the increased background at 3 TeV the *tight* PFO selection is applied. Jets are reconstructed with exclusive ($N=2$) clustering with $R=1.2$. For comparison the same algorithm is also

Table 2: Response and resolution of the di-jet mass distributions obtained with the k_t and VLC algorithms in the left panel of Figure 11. The columns list the median di-jet mass, the IQR_{34} and the RMS_{90} .

algorithm	median [GeV]	IQR_{34} [GeV]	RMS_{90} [GeV]
k_t ($R = 1.3$)	118.5	30.1	24.3
VLC ($R = 1.3, \beta = \gamma = 1$)	118.9	27.1	22.0

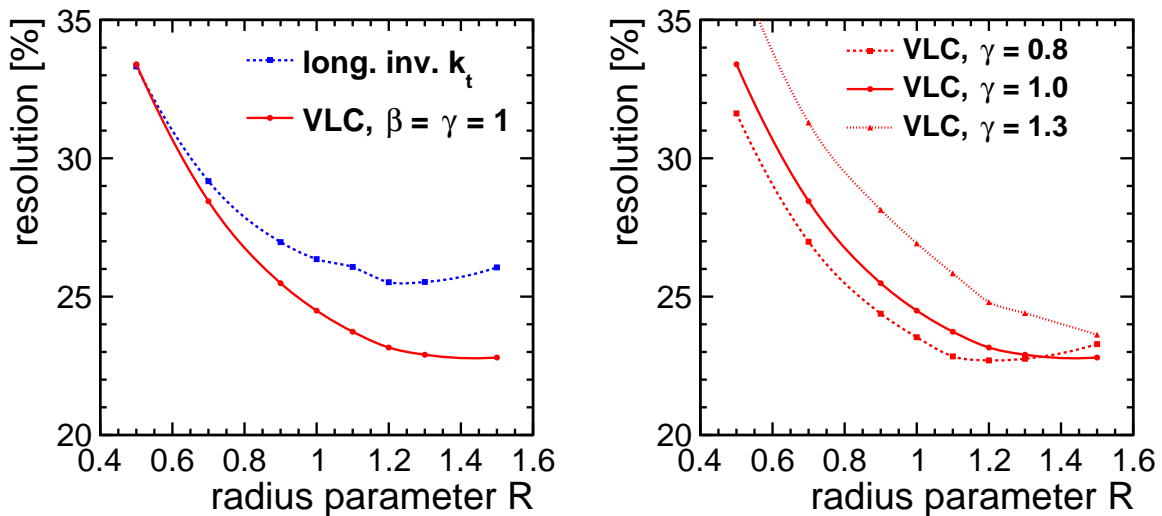


Figure 12: The reconstructed di-jet mass resolution (determined as the 34% inter-quantile range IQR_{34}) for simulated, fully hadronic decays of $e^+e^- \rightarrow \nu\bar{\nu}hh$, $h \rightarrow b\bar{b}$ events produced in 3 TeV e^+e^- collisions at CLIC. The nominal $\gamma\gamma \rightarrow \text{hadrons}$ background is overlaid on the signal event. Particle flow objects are selected using the tight selection.

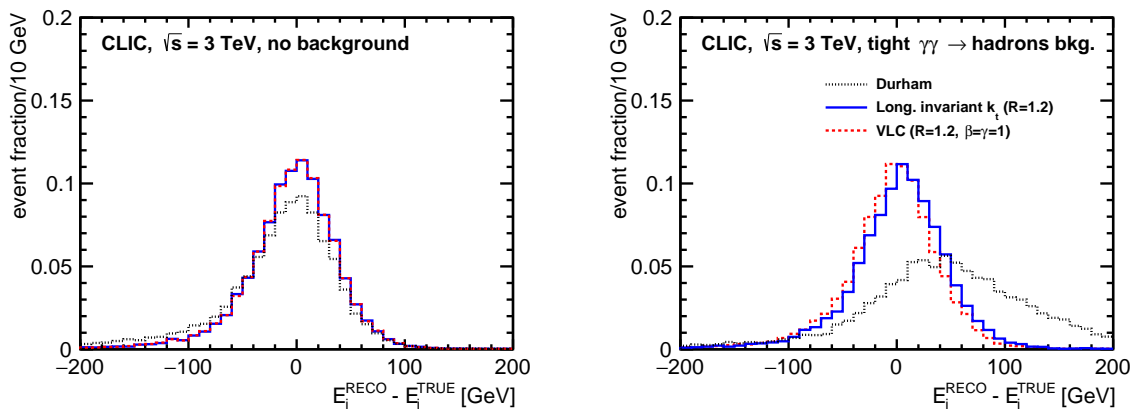


Figure 13: The jet energy residuals (reconstructed minus true energy) for fully hadronic decays of $t\bar{t}$ events at a 3 TeV CLIC. No backgrounds are added in the left plot. In the right plot 60 bunch crossings of $\gamma\gamma \rightarrow \text{hadrons}$ background are overlaid on the signal and particle flow objects are selected using the tight selection.

run on all stable Monte Carlo particles. These include neutrinos, but not the particles from $\gamma\gamma \rightarrow \text{hadrons}$ background.

The jet energy is a fairly good measure of the top quark energy. The correction to the top quark energy is typically 3.5% and the energy resolution is typically 8%. To measure the performance we compare the jet reconstructed from particle flow objects with the jet found by the same algorithm on the stable particles from the signal event (i.e. excluding the $\gamma\gamma \rightarrow \text{hadrons}$). The jet energy residual is defined as the difference of the energy of detector-level and particle-level jets. The distribution is shown in Fig. 13. The response is measured as the median of the residual distribution. The resolution is measured as the IQR_{34} .

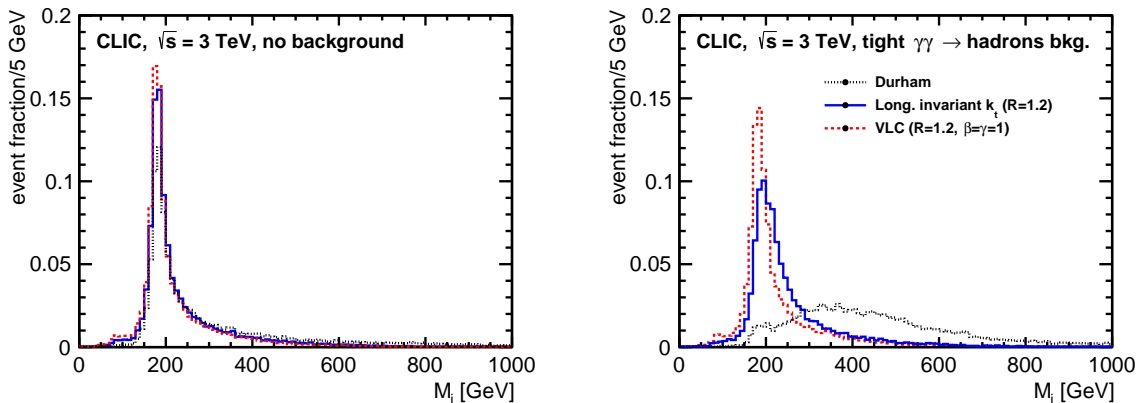


Figure 14: The reconstructed jet mass distribution for fully hadronic decays of $t\bar{t}$ events at a 3 TeV CLIC. No backgrounds are added in the left plot. In the right plot 60 bunch crossings of $\gamma\gamma \rightarrow \text{hadrons}$ background are overlaid on the signal and particle flow objects are selected using the tight selection.

Table 3: The bias and resolution of the energy and mass measurements of reconstructed top jets in top quark pair production with fully hadronic top quark decay at a centre-of-mass energy of 3 TeV. Results are presented for the median response and two estimates of the resolution: the 34% inter-quantile range (IQR_{34}) and the RMS of central 90% of jets (RMS_{90}). All results are obtained by comparing the jet energy reconstructed from particle flow objects to the jet of stable MC particles from the signal event. The performance of the classical e^+e^- algorithm is such that the figures-of-merit cannot be estimated reliably under *nominal* background conditions (indicated by “-” entries in the Table).

CLIC, $\sqrt{s} = 3$ TeV, energy resolution (no bkg./tight/nominal) [%]									
	median			IQR_{34}			RMS_{90}		
Durham	-0.9	3.1	-	4.6	6.6	-	3.7	5.7	-
generic $e^+e^- k_t$ ($R=1$)	-0.3	0.5	-	3.4	4.0	-	2.7	3.4	-
long. inv. k_t ($R=1.2$)	-0.2	0.4	1.8	3.1	3.2	3.4	2.5	2.7	2.8
VLC ($R=1.2$)	-0.2	-0.2	0.5	3.1	3.2	3.2	2.5	2.6	2.6
CLIC, $\sqrt{s} = 3$ TeV, mass resolution (no bkg./tight/nominal) [%]									
	median			IQR_{34}			RMS_{90}		
Durham	-1.0	37.7	-	14.3	-	-	11.7	33.8	-
generic $e^+e^- k_t$ ($R=1$)	0.5	4.7	-	5.1	23.2	-	4.6	17.0	-
long. inv. k_t ($R=1.2$)	1.1	8.0	21.2	4.1	12.0	20.6	3.5	9.9	16.3
VLC ($R=1.2$)	0.8	1.7	5.6	4.1	7.1	9.4	3.5	6.0	8.0

Quantitative results are presented in Table 3. The RMS_{90} is also presented to facilitate comparison to other studies. In the absence of background, all algorithms reconstruct the energy of the jet quite precisely, with a bias of less than 1% and a resolution of 2-4%. The performance of the classical e^+e^- algorithms is degraded as soon as the $\gamma\gamma \rightarrow \text{hadrons}$ background with tight PFO selection is added. The VLC and longitudinally invariant k_t algorithms show very little performance degradation even with the nominal PFO selection.

The jet invariant mass is much more sensitive to soft background contamination [37, 39]. The jet mass distributions are presented in Figure 14. In the left panel, which corresponds to $t\bar{t}$ events without background overlay, all algorithms are seen to reconstruct a narrow peak close to the top quark mass. The long tail toward large mass is due to radiation off the top quark and its decay products and is also

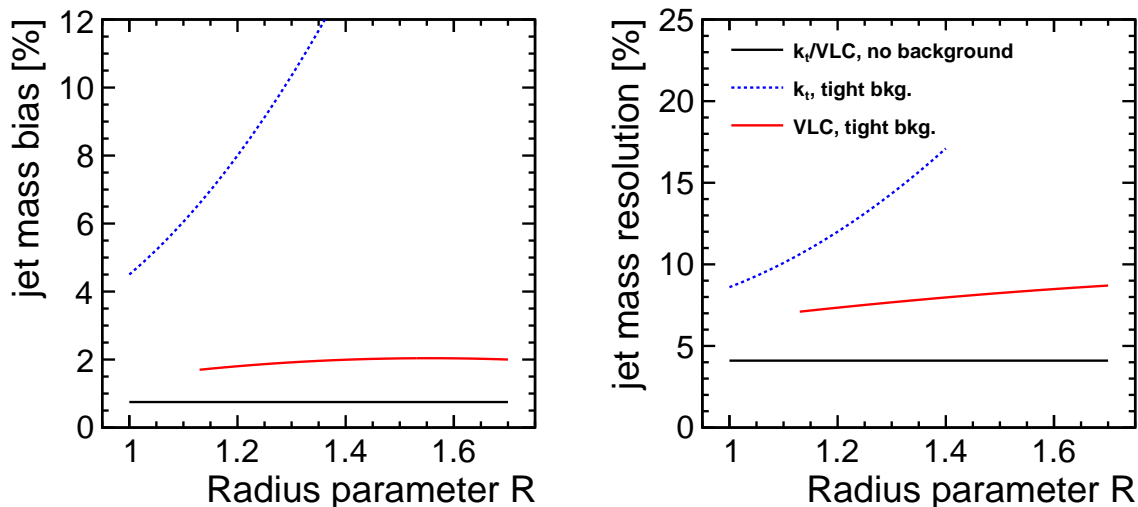


Figure 15: The bias and resolution of the reconstructed jet mass versus radius parameter R of two jet algorithms. The jets are reconstructed in fully hadronic top quark decays in $t\bar{t}$ events at a center-of-mass energy of 3 TeV. The jets reconstructed from particle-flow objects are compared to jets reconstructed from all stable particles from the signal event. The black curves correspond to the results obtained without background, the blue dashed and red curves to 60 bunch crossings of $\gamma\gamma \rightarrow \text{hadrons}$ background overlaid on the signal, with the *tight* PFO selection.

present in the jets reconstructed from stable MC particles. The plots in the right panel show a severe degradation when the $\gamma\gamma \rightarrow \text{hadrons}$ background with tight PFO selection is added, most noticeably for the Durham algorithm. The bias and resolution of the jet mass is shown as a function of radius parameter in Fig. 15.

A quantitative summary is presented in the second part of Table 3. The bias on the jet mass without background is sub-% for most algorithms. The resolution of the VLC and longitudinally invariant k_t algorithms is significantly better than that of the classical e^+e^- algorithms. The 4.1% resolution is a testimony to the potential of highly granular calorimeters and particle flow reconstruction for jet substructure measurements.

The $\gamma\gamma \rightarrow \text{hadrons}$ background has a profound effect on the performance. The performance of the classical algorithms is clearly inadequate, with a strong bias and a severe degradation, even with the tight PFO selection. The VLC and longitudinally invariant k_t algorithms are much less affected, as expected from the smaller exposed area. The VLC algorithm is found to be more resilient than the longitudinally invariant k_t , confirming the result anticipated at the particle level in Section 5.

7 Discussion

In this Section we discuss the implications for lower-energy colliders and identify several topics that merit further study.

7.1 Implications for lower-energy lepton colliders

In this paper we have focused on CLIC operation at $\sqrt{s} = 3$ TeV, arguably the most challenging environment that lepton colliders might face in the next decades. We have chosen this environment because subtle differences in jet definitions lead to significant differences in performance. This has helped us to

gain a deeper understanding of the intricacies of jet reconstruction at lepton colliders and to establish solid conclusions about the resilience of the different algorithms.

These findings are by no means limited to the CLIC environment. Subtle but significant differences in performance are expected also for the ILC at 250 GeV or 500 GeV and at circular colliders.

7.2 Further R&D on jet algorithms

The set of jet algorithms studied in this paper is by no means exhaustive. This study does not address several recent proposals, such as the XCone algorithm [46] or the *global* jet clustering proposed by Georgi [47].

A broad range of new techniques developed for the LHC have so far remained unexplored. This is particularly true for a set of tools that has proven extremely powerful in pile-up mitigation and correction in ATLAS and CMS.

Jet grooming (the collective name for (mass-drop) filtering [48], pruning [49] and trimming [50]) effectively reduces the exposed jet area to several small regions with large energy flow. This provides an effective means of capturing a large fraction of the jet energy while reducing the impact of *soft* contamination. Tests of the trimming algorithm in the CLIC environment yield very good results, improving the jet mass resolution significantly.

Techniques to correct for the effect of pile-up based on an event-by-event measurement of the pile-up activity [51] are quite successful at the LHC. Subtraction at the constituent level with dynamical thresholds [52–54] is under active development. An adaptation to the environment at lepton colliders, with a very sparse background energy flow, may prove useful.

8 Conclusions

We have studied the jet reconstruction performance of several sequential reconstruction algorithms at high-energy lepton colliders. In addition to the classical e^+e^- algorithms we include a version of the same algorithms with beam jets. We also study the performance of the longitudinally invariant k_t algorithm and a new e^+e^- algorithm, called VLC [19], which are expected to be more resilient to the impact of backgrounds.

The study is based on detailed Monte Carlo simulation. For two benchmark processes we use a full simulation of the linear collider detector concepts, including the relevant background processes.

The perturbative energy corrections of all algorithms with finite size jets are sizeable for small values of the radius parameter (10-15% for $R=0.5$) and decrease to 1-5% for $R=1.5$. This result is approximately independent of the process and centre-of-mass energy. Convergence with R is faster for the generalized e^+e^- algorithm than for the longitudinally invariant k_t algorithm and VLC, that expose a smaller area to forward jets.

The non-perturbative hadronization correction represents a very small part of the total correction. Its relevance decreases with increasing centre-of-mass energy: it is less than 1% at $\sqrt{s}=250$ GeV for all algorithms and less than a per mille at $\sqrt{s}=3$ TeV. The generalized e^+e^- algorithm again converges fastest. We have estimated the non-perturbative correction also for the jet invariant mass: these corrections are much larger than the non-perturbative energy corrections and remain of the order of a few % even for $R=1.5$ and $\sqrt{s}=3$ TeV.

The forward-peaked $\gamma\gamma \rightarrow \text{hadrons}$ background at future high-energy linear lepton colliders is one of the most important factors in the jet reconstruction performance. It has motivated ILC and CLIC to abandon classical, inclusive algorithms in favour of algorithms with a finite jet size. Algorithms that expose a reduced solid angle in the forward region of the detector, such as longitudinally invariant k_t or VLC are more robust. A particle-level study shows that these two algorithms have a different response,

with VLC showing a lower response, but one that is more stable versus polar angle. VLC is found to be less susceptible to background.

We present two studies in full simulation, namely di-Higgs production and top quark pair production at $\sqrt{s} = 3$ TeV, which present a combination of a relatively harsh background level, high jet multiplicity and forward jets. In both cases the classical e^+e^- algorithms offer an inadequate performance. The same is true for the generalized version with beam jets. VLC provides significantly better mass resolution for the Higgs study and considerably better jet mass reconstruction than the longitudinally invariant k_t algorithm.

Jet clustering is key technique for many analyses of multi-jet final states at future high-energy electron-positron colliders. This study shows that a considerable increase in performance can be obtained by a careful choice of the clustering algorithm. We recommend, therefore, that studies into the physics potential of future e^+e^- colliders carefully optimize the choice of the jet reconstruction algorithm and its parameters. We also encourage further work on robust algorithms for e^+e^- collisions.

Acknowledgement

This work has been carried out in the framework of the CLICdp collaboration. The authors acknowledge the effort of the ILC and CLIC detector & physics groups in putting together the simulation infrastructure used to benchmark the algorithms. This work benefited from services provided by the ILC Virtual Organisation, supported by the national resource providers of the EGI Federation. The authors of the VLC algorithm would like to thank Gavin Salam and Jesse Thaler for helpful suggestions and the FastJet team for guidance creating the plugin code.

References

- [1] K. Fujii et al., *Physics Case for the International Linear Collider* (2015), arXiv: 1506.05992 [hep-ex].
- [2] H. Baer et al., *The International Linear Collider Technical Design Report - Volume 2: Physics* (2013), ed. by H. Baer, arXiv: 1306.6352 [hep-ph].
- [3] L. Linssen et al., *Physics and Detectors at CLIC: CLIC Conceptual Design Report* (2012), arXiv: 1202.5940 [physics.ins-det].
- [4] M. J. Boland et al., *Updated baseline for a staged Compact Linear Collider* (2016), ed. by P. Lebrun et al., DOI: 10.5170/CERN-2016-004, arXiv: 1608.07537 [physics.acc-ph].
- [5] M. Bicer et al., *First Look at the Physics Case of TLEP*, JHEP **01** (2014) 164, DOI: 10.1007/JHEP01(2014)164, arXiv: 1308.6176 [hep-ex].
- [6] CEPC-SPPC Study Group, *CEPC-SPPC Preliminary Conceptual Design Report. 1. Physics and Detector* (2015), URL: <http://cepc.ihep.ac.cn/preCDR/volume.html>.
- [7] Y. Alexahin et al., *Muon Collider Higgs Factory for Snowmass 2013* (2013), arXiv: 1308.2143 [hep-ph].
- [8] H. Abramowicz et al., *The International Linear Collider Technical Design Report - Volume 4: Detectors* (2013), ed. by T. Behnke et al., arXiv: 1306.6329 [physics.ins-det].
- [9] N. Alipour Tehrani et al., *CLICdet: The post-CDR CLIC detector model* (2017), URL: <https://cds.cern.ch/record/2254048>.

- [10] C. Adloff et al., *Electromagnetic response of a highly granular hadronic calorimeter*, JINST **6** (2011) P04003, DOI: [10.1088/1748-0221/6/04/P04003](https://doi.org/10.1088/1748-0221/6/04/P04003), arXiv: [1012.4343](https://arxiv.org/abs/1012.4343) [physics.ins-det].
- [11] C. Adloff et al., *Hadronic energy resolution of a highly granular scintillator-steel hadron calorimeter using software compensation techniques*, JINST **7** (2012) P09017, DOI: [10.1088/1748-0221/7/09/P09017](https://doi.org/10.1088/1748-0221/7/09/P09017), arXiv: [1207.4210](https://arxiv.org/abs/1207.4210) [physics.ins-det].
- [12] J. S. Marshall, M. A. Thomson, *The Pandora Software Development Kit for Pattern Recognition*, Eur. Phys. J. **C75** (2015) 439, DOI: [10.1140/epjc/s10052-015-3659-3](https://doi.org/10.1140/epjc/s10052-015-3659-3), arXiv: [1506.05348](https://arxiv.org/abs/1506.05348) [physics.data-an].
- [13] P. Chen, T. L. Barklow, M. E. Peskin, *Hadron production in $\gamma\gamma$ collisions as a background for e^+e^- linear colliders*, Phys.Rev. **D49** (1994) 3209, DOI: [10.1103/PhysRevD.49.3209](https://doi.org/10.1103/PhysRevD.49.3209), arXiv: [hep-ph/9305247](https://arxiv.org/abs/hep-ph/9305247) [hep-ph].
- [14] J. Marshall, A. Muennich, M. Thomson, *Performance of Particle Flow Calorimetry at CLIC*, Nucl.Instrum.Meth. **A700** (2013) 153, DOI: [10.1016/j.nima.2012.10.038](https://doi.org/10.1016/j.nima.2012.10.038), arXiv: [1209.4039](https://arxiv.org/abs/1209.4039) [physics.ins-det].
- [15] F. Simon, L. Weuste, *Light-flavor squark reconstruction at CLIC*, Eur. Phys. J. **C75** (2015) 379, DOI: [10.1140/epjc/s10052-015-3607-2](https://doi.org/10.1140/epjc/s10052-015-3607-2), arXiv: [1505.01129](https://arxiv.org/abs/1505.01129) [hep-ex].
- [16] S. Catani et al., *New clustering algorithm for multi-jet cross-sections in e^+e^- annihilation*, Phys.Lett. **B269** (1991) 432, DOI: [10.1016/0370-2693\(91\)90196-W](https://doi.org/10.1016/0370-2693(91)90196-W).
- [17] S. Catani et al., *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, Nucl.Phys. **B406** (1993) 187, DOI: [10.1016/0550-3213\(93\)90166-M](https://doi.org/10.1016/0550-3213(93)90166-M).
- [18] S. D. Ellis, D. E. Soper, *Successive combination jet algorithm for hadron collisions*, Phys.Rev. **D48** (1993) 3160, DOI: [10.1103/PhysRevD.48.3160](https://doi.org/10.1103/PhysRevD.48.3160), arXiv: [hep-ph/9305266](https://arxiv.org/abs/hep-ph/9305266) [hep-ph].
- [19] M. Boronat et al., *A robust jet reconstruction algorithm for high-energy lepton colliders*, Phys. Lett. **B750** (2015) 95, DOI: [10.1016/j.physletb.2015.08.055](https://doi.org/10.1016/j.physletb.2015.08.055), arXiv: [1404.4294](https://arxiv.org/abs/1404.4294) [hep-ex].
- [20] M. Thomson, *Model-independent measurement of the $e^+e^- \rightarrow HZ$ cross section at a future e^+e^- linear collider using hadronic Z decays*, Eur. Phys. J. **C76** (2016) 72, DOI: [10.1140/epjc/s10052-016-3911-5](https://doi.org/10.1140/epjc/s10052-016-3911-5), arXiv: [1509.02853](https://arxiv.org/abs/1509.02853) [hep-ex].
- [21] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, DOI: [10.1007/JHEP07\(2014\)079](https://doi.org/10.1007/JHEP07(2014)079), arXiv: [1405.0301](https://arxiv.org/abs/1405.0301) [hep-ph].
- [22] T. Sjostrand, S. Mrenna, P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852, DOI: [10.1016/j.cpc.2008.01.036](https://doi.org/10.1016/j.cpc.2008.01.036), arXiv: [0710.3820](https://arxiv.org/abs/0710.3820) [hep-ph].
- [23] *Summary of Higgs coupling measurements with staged running of ILC at 250 GeV, 500 GeV and 1 TeV*, LC-REP-2013-021, 2013, URL: <https://flc.desy.de/lcnotes/notes/LC-REP-2013-021.pdf>.
- [24] T. Price et al., *Full simulation study of the top Yukawa coupling at the ILC at $\sqrt{s} = 1$ TeV*, Eur. Phys. J. **C75** (2015) 309, DOI: [10.1140/epjc/s10052-015-3532-4](https://doi.org/10.1140/epjc/s10052-015-3532-4), arXiv: [1409.7157](https://arxiv.org/abs/1409.7157) [hep-ex].

- [25] S. Redford, P. Roloff, M. Vogel, *Physics potential of the top Yukawa coupling measurement at a 1.4 TeV Compact Linear Collider using the CLIC SiD detector* (2014),
URL: <https://cds.cern.ch/record/1690648>.
- [26] J. Fuster et al., *Forward tracking at the next $e^+ e^-$ collider. Part I. The Physics case*, JINST **4** (2009) P08002, DOI: 10.1088/1748-0221/4/08/P08002, arXiv: 0905.2038 [hep-ex].
- [27] A. Abdesselam et al., *Boosted objects: A Probe of beyond the Standard Model physics*, Eur. Phys. J. **C71** (2011) 1661, DOI: 10.1140/epjc/s10052-011-1661-y, arXiv: 1012.5412 [hep-ph].
- [28] M. H. Seymour,
Searches for new particles using cone and cluster jet algorithms: A Comparative study, Z. Phys. **C62** (1994) 127, DOI: 10.1007/BF01559532.
- [29] D. Schulte, *Background at future linear colliders*,
The development of future linear electron positron colliders: For particle physics and for research using free electron lasers. Proceedings, Workshop, Lund, Sweden, September 23-25, 1999, 1999, URL: <https://cds.cern.ch/record/419592>.
- [30] S. Poss, A. Sailer, *Luminosity Spectrum Reconstruction at Linear Colliders*, Eur. Phys. J. **C74** (2014) 2833, DOI: 10.1140/epjc/s10052-014-2833-3, arXiv: 1309.0372 [physics.ins-det].
- [31] M. Cacciari, G. P. Salam, G. Soyez, *FastJet User Manual*, Eur.Phys.J. **C72** (2012) 1896, DOI: 10.1140/epjc/s10052-012-1896-2, arXiv: 1111.6097 [hep-ph].
- [32] M. Cacciari, G. P. Salam, *Dispelling the N^3 myth for the k_t jet-finder*, Phys.Lett. **B641** (2006) 57, DOI: 10.1016/j.physletb.2006.08.037, arXiv: hep-ph/0512210 [hep-ph].
- [33] *ValenciaJetAlgorithm plug-in for fastjet*, <https://fastjet.hepforge.org/contrib/>,
URL: <https://fastjet.hepforge.org/contrib/>.
- [34] Y. L. Dokshitzer et al., *Better jet clustering algorithms*, JHEP **9708** (1997) 001, DOI: 10.1088/1126-6708/1997/08/001, arXiv: hep-ph/9707323 [hep-ph].
- [35] S. Catani, Y. L. Dokshitzer, B. Webber,
The K^\perp clustering algorithm for jets in deep inelastic scattering and hadron collisions, Phys.Lett. **B285** (1992) 291, DOI: 10.1016/0370-2693(92)91467-N.
- [36] M. Cacciari, G. P. Salam, G. Soyez, *The Anti- k_t jet clustering algorithm*, JHEP **0804** (2008) 063, DOI: 10.1088/1126-6708/2008/04/063, arXiv: 0802.1189 [hep-ph].
- [37] M. Dasgupta, L. Magnea, G. P. Salam, *Non-perturbative QCD effects in jets at hadron colliders*, JHEP **0802** (2008) 055, DOI: 10.1088/1126-6708/2008/02/055, arXiv: 0712.3014 [hep-ph].
- [38] S. Bethke et al.,
New jet cluster algorithms: Next-to-leading order QCD and hadronization corrections, Nucl. Phys. **B370** (1992), [Erratum: Nucl. Phys. B523,681(1998)] 310, DOI: 10.1016/S0550-3213(98)00219-3, 10.1016/0550-3213(92)90289-N.
- [39] *Jet mass and substructure of inclusive jets in $\sqrt{s} = 7$ TeV pp collisions with the ATLAS experiment*, JHEP **1205** (2012) 128, DOI: 10.1007/JHEP05(2012)128, arXiv: 1203.4606 [hep-ex].
- [40] W. Kilian, T. Ohl, J. Reuter, *WHIZARD: Simulating Multi-Particle Processes at LHC and ILC*, Eur.Phys.J. **C71** (2011) 1742, DOI: 10.1140/epjc/s10052-011-1742-y, arXiv: 0708.4233 [hep-ph].

-
- [41] S. Agostinelli et al., *GEANT4: A Simulation toolkit*, Nucl.Instrum.Meth. **A506** (2003) 250, DOI: [10.1016/S0168-9002\(03\)01368-8](https://doi.org/10.1016/S0168-9002(03)01368-8).
- [42] J. Marshall, M. Thomson, *The Pandora software development kit for particle flow calorimetry*, J.Phys.Conf.Ser. **396** (2012) 022034, DOI: [10.1088/1742-6596/396/2/022034](https://doi.org/10.1088/1742-6596/396/2/022034).
- [43] C. Grefe et al., *ILCDIRAC, a DIRAC extension for the Linear Collider community*, J. Phys. Conf. Ser. **513** (2014) 032077, DOI: [10.1088/1742-6596/513/3/032077](https://doi.org/10.1088/1742-6596/513/3/032077).
- [44] A. Tsaregorodtsev et al., *DIRAC: A community grid solution*, J. Phys. Conf. Ser. **119** (2008) 062048, DOI: [10.1088/1742-6596/119/6/062048](https://doi.org/10.1088/1742-6596/119/6/062048).
- [45] H. Abramowicz et al., *Higgs physics at the CLIC electron-positron linear collider*, Eur. Phys. J. **C77** (2017) 475, DOI: [10.1140/epjc/s10052-017-4968-5](https://doi.org/10.1140/epjc/s10052-017-4968-5), arXiv: [1608.07538](https://arxiv.org/abs/1608.07538) [hep-ex].
- [46] I. W. Stewart et al., *XCone: N-jettiness as an Exclusive Cone Jet Algorithm*, JHEP **11** (2015) 072, DOI: [10.1007/JHEP11\(2015\)072](https://doi.org/10.1007/JHEP11(2015)072), arXiv: [1508.01516](https://arxiv.org/abs/1508.01516) [hep-ph].
- [47] H. Georgi, *A Simple Alternative to Jet-Clustering Algorithms* (2014), arXiv: [1408.1161](https://arxiv.org/abs/1408.1161) [hep-ph].
- [48] J. M. Butterworth et al., *Jet substructure as a new Higgs search channel at the LHC*, Phys. Rev. Lett. **100** (2008) 242001, DOI: [10.1103/PhysRevLett.100.242001](https://doi.org/10.1103/PhysRevLett.100.242001), arXiv: [0802.2470](https://arxiv.org/abs/0802.2470) [hep-ph].
- [49] S. D. Ellis, C. K. Vermilion, J. R. Walsh, *Techniques for improved heavy particle searches with jet substructure*, Phys. Rev. **D80** (2009) 051501, DOI: [10.1103/PhysRevD.80.051501](https://doi.org/10.1103/PhysRevD.80.051501), arXiv: [0903.5081](https://arxiv.org/abs/0903.5081) [hep-ph].
- [50] D. Krohn, J. Thaler, L.-T. Wang, *Jet Trimming*, JHEP **02** (2010) 084, DOI: [10.1007/JHEP02\(2010\)084](https://doi.org/10.1007/JHEP02(2010)084), arXiv: [0912.1342](https://arxiv.org/abs/0912.1342) [hep-ph].
- [51] M. Cacciari, G. P. Salam, *Pileup subtraction using jet areas*, Phys. Lett. **B659** (2008) 119, DOI: [10.1016/j.physletb.2007.09.077](https://doi.org/10.1016/j.physletb.2007.09.077), arXiv: [0707.1378](https://arxiv.org/abs/0707.1378) [hep-ph].
- [52] M. Cacciari, G. P. Salam, G. Soyez, *SoftKiller, a particle-level pileup removal method*, Eur. Phys. J. **C75** (2015) 59, DOI: [10.1140/epjc/s10052-015-3267-2](https://doi.org/10.1140/epjc/s10052-015-3267-2), arXiv: [1407.0408](https://arxiv.org/abs/1407.0408) [hep-ph].
- [53] D. Bertolini et al., *Pileup Per Particle Identification*, JHEP **10** (2014) 059, DOI: [10.1007/JHEP10\(2014\)059](https://doi.org/10.1007/JHEP10(2014)059), arXiv: [1407.6013](https://arxiv.org/abs/1407.6013) [hep-ph].
- [54] P. Berta et al., *Particle-level pileup subtraction for jets and jet shapes*, JHEP **06** (2014) 092, DOI: [10.1007/JHEP06\(2014\)092](https://doi.org/10.1007/JHEP06(2014)092), arXiv: [1403.3108](https://arxiv.org/abs/1403.3108) [hep-ex].