

Resolving the Fundamental QCD Degrees of Freedom

Dilan Premgi^{1,a} and Carolina Batista^{1,b}

¹ Faculty of Sciences of the University of Lisbon, Portugal

² Instituto Superior Técnico, Lisboa, Portugal

Project supervisors: Liliana Apolinário² and Nuno Olavo Madureira²

October 16, 2025

Abstract. We present a comprehensive study on the formation time structure of QCD emissions in proton-proton collisions, with particular emphasis on charm-anticharm ($c\bar{c}$) pair production. Using Monte Carlo simulations with PYTHIA 8.3 and advanced jet reclustering algorithms, we investigate the temporal evolution of parton shower development and compare reclustering methods with Monte Carlo techniques employed at both parton and hadron levels. Our analysis employs the anti- k_r clustering algorithm for jet identification and the Cambridge/Aachen algorithm for reclustering to reconstruct formation time distributions. We illustrate how formation time calculations can be reliably reproduced using reclustering techniques, providing us a precise way to accurately understand the space-time structure of QCD jets. The study aims to address the hot QCD problem by performing a detailed analysis of proton-proton collision systems that provides the methodological foundation for future heavy-ion collision studies where jet-medium interactions in the Quark-Gluon Plasma can yield tomographic information about the extreme QCD matter.

KEYWORDS: Jet, Reclustering, Hadronization, Formation Time, Parton Shower, Monte Carlo Simulations

1 Introduction

The study of Quantum Chromodynamics (QCD) under extreme conditions, namely at high densities and temperatures, is central to understanding the strong interaction. Such extreme conditions can be met at high-energy particle collisions, where partonic matter exhibits a fluid-like behaviour at large scales, forming the quark-gluon plasma (QGP) [1]. This state of matter was first theoretically proposed in the late 1970s [2]. The QGP was experimentally confirmed in the early 2000s through high-energy heavy-ion collisions at the Super Proton Synchrotron (SPS) at CERN and subsequently studied extensively at the Relativistic Heavy-Ion Collider (RHIC) and the Large Hadron Collider (LHC). [3–5]

This strongly coupled state of quarks and gluons existed in the primordial Universe soon after the Big Bang, making its study crucial for understanding the evolution of matter in the early Universe [1]. Despite its extremely short lifetime and the extreme conditions required for its formation, such as density exceeding that of atomic nuclei and temperatures reaching several hundred MeV, the QGP provides a unique environment for testing our understanding of QCD under such extreme conditions. [1]

The properties of QGP, such as transport coefficients, thermalisation timescales, opacity to colored probes, and the interaction between perturbative and non-perturbative dynamics, are active areas of theoretical and experimental research [1, 6]. QGP also raises numerous questions, from which we can highlight a few: How does QGP form and behave? And how do we merge perturbative QCD

at small scales with the emergent collective behaviour at large scales?

One of the possibilities to address these open problems is to use jets. A jet is defined as a collimated spray of energetic hadrons, produced in ultra-relativistic particle collisions. A jet acts as a proxy for a hard-scattered quark-gluon ejected in the collision. Since our ultimate objective is to examine the QGP, we want to extract information from heavy-ion collision jets regarding the interaction between the jet and the medium. They provide crucial information about the underlying QCD processes and are the visible representation of parton showers.

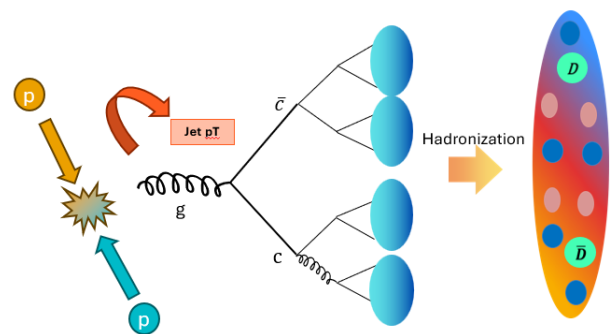


Figure 1: Schematic representation of jet formation in a high-energy proton-proton collision. A hard scattering produces a high- p_T gluon, which undergoes successive branchings, including heavy quark pair production ($g \rightarrow c\bar{c}$). The resulting parton shower eventually hadronizes into a collimated spray of particles- a jet.

Figure 1 illustrates this process schematically. In a proton-proton (pp) collisions, a high-energy gluon, for instance, is produced in the hard scattering. Subsequently,

^ae-mail: fc59903@alunos.fc.ul.pt

^be-mail: fc61888@alunos.fc.ul.pt

this gluon initiates a parton shower, undergoing multiple splittings that may include the production of heavy quark pairs such as $c\bar{c}$. At later stages, the partons transition into hadrons through the non-perturbative process of hadronization, forming the observable jet. By reconstructing jets and their internal structure, we can probe the dynamics of the parton shower and gain access to the building blocks of QCD processes.

To probe the properties of the QGP, it is essential to study its interactions with hard partons and the resulting modification. In the absence of the QCD medium (vacuum) its behaviour is well understood. Among the relevant observables, the formation time plays a particularly important role, as the QGP evolution provides a time sequence frame from which we can study the space-time substructure of the jets.

In perturbative QCD, a splitting such as $g \rightarrow c\bar{c}$ is characterised by the relevant energy scale and the kinematic configuration of the emitted partons. The associated formation time can be understood as the time interval required for the two outgoing partons to behave as independent QCD radiation sources. In the high-energy limit, this formation time, τ_{form} , can be estimated as [7]

$$\tau_{\text{form}} = \frac{1}{2zE(1-z)(1-\cos(\theta_{12}))} \quad (1)$$

where E denotes the energy of the parent parton, z the fraction of energy carried by the softer child parton, and θ_{12} the angle of emission (i.e. the angle between both outgoing partons).

From Eq. 1, it follows that the higher the energy, the wider the angle between the two emitted partons, and the more balanced the division of energy between them, the shorter the formation time and vice-versa. Consequently, modifications in formation time's distributions can provide valuable information on how the medium influences certain partons and affects these variables. So, the question arises of which partons to follow.

Experimentally, since only the final products of hadronization are accessible, heavy quarks are of particular interest, due to their higher mass, making them easier to trace. In this work, we focus on $c\bar{c}$ pairs reconstructed from $D^0\bar{D}^0$ mesons. Because of the higher energies required to form these quarks we are able to track them more accurately.

The mass of the heavy quark suppresses collinear radiation (the dead cone effect) and alters fragmentation patterns [6]. For $g \rightarrow c\bar{c}$ splittings, the mass terms in Eq. (1) are generally non-negligible. Mass effects also influence the energy sharing z and the angular distributions; heavier quarks tend to populate different regions of (z, θ) phase space compared to light quarks and gluons.

An open issue is then the time ordering of branchings in the parton shower, and how early splittings are affected by the medium. In particular: at what time scales do splittings like $g \rightarrow c\bar{c}$ occur, and how likely are they to occur inside the medium.

This motivates the use of clustering and reclustering techniques, which allow us to trace back the shower history from the hadronic final state, and to connect reconstructed observables with theoretical predictions. For that, we use a jet finding algorithm called anti- k_r [8] to group final-state particles into jets, and then the Cambridge/Aachen (C/A) method to reconstruct them using an angularly ordered algorithm. By examining early branchings in the reclustered tree and computing splitting kinematics, we obtain reconstructed proxies for τ_{form} . This method was developed and tested in recent literature [7, 9] and is the central phenomenological technique of this work.

Both the clustering and reclustering methods are based on iterative algorithms performing pairwise groupings of particles. The generic sequential recombination jet algorithm defines a distance measure between two particles i and j (d_{ij}) as:

$$d_{ij} = \min(p_{Ti}^{2p}, p_{Tj}^{2p}) \frac{\Delta R_{ij}^2}{R^2} \quad (2)$$

where p_{Ti} represents the transverse momentum of a particle i and $\frac{\Delta R_{ij}^2}{R^2}$ the angular separation with $\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, where y_i is the rapidity and ϕ_i is the azimuthal angle of a particle i , weighted by the jet radius R . Each value is attributed to the parameter p , which produces an individual jet algorithm from this family, with its very own ordering variable.

For the anti- k_T method, the parameter is chosen to be $p = -1$, making it ideal to find jets (i.e. initial jet identification). The method preferentially clusters hard particles first, resulting in approximately circular jet areas that are robust against soft radiation and pile-up effects [10]. The algorithm works iteratively, selecting the pair of partons with the smallest d_{ij} and merging them into a pseudo-particle by combining their four-momenta.

This procedure is shown in Figure 2, where the algorithm iteratively organises the individual final-state particles (Step 1) into bigger clusters (Steps 2–4); by the end, the hardest component dominates the emerging jet [1]. Due to this characteristic, the anti- k_r algorithm is the most reliable in establishing the general jet finding.

Following the initial identification of the jet, we use a reclustering method to reconstruct the angular order branching history using the Cambridge/Aachen (C/A) algorithm [8]. The C/A algorithm reclusters the jet particles using a purely geometric distance implementation with a parameter $p = 0$ in Eq. (2). This algorithm is widely used for jet substructure as it reconstructs the branching his-

Clustering Method with Anti- k_t ($p = -1$)

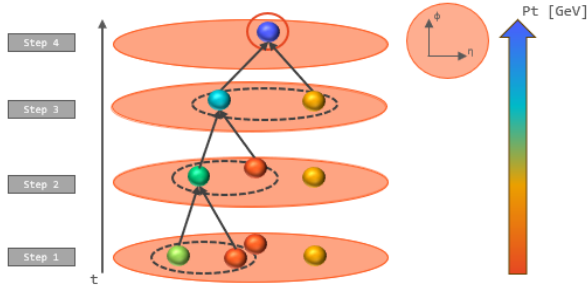


Figure 2: Illustration of the anti- k_t clustering algorithm ($p = -1$). Particles are successively combined according to a distance measure that favours high- p_T members, leading to jets with nearly conical shapes. The vertical axis represents the clustering steps in time, and the colours encode the intensity of the transverse momentum (p_T).

tory via decreasing angular order, a well-known property of QCD radiation. [11]

Reclustering Method with C/A ($p = 0$)

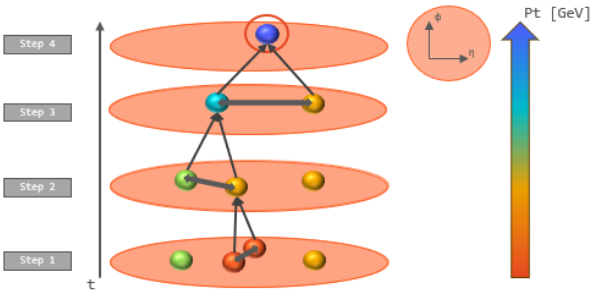


Figure 3: Reclustering sequence with the Cambridge/Aachen (C/A) algorithm ($p = 0$). Constituents are combined purely according to their angular separation in (y, ϕ) space, independent of their transverse momentum. This approach allows one to reconstruct the jet's branching history in a way that is directly connected to the space-time development of the parton shower.

The reclustering process using the Cambridge/Aachen technique is depicted in Figure 3. Unlike the anti- k_t , which favours high- p_T components, C/A just groups particles according to their angular proximity. The C/A algorithm is especially well-suited because of the angular geometry of the parton shower, which is reflected in the recombination sequence. The splitting history can be traced back step by step by undoing this sequence and identifying a distinctive angle to each splitting.

Previous work, for example, the "Time reclustering for jet quenching studies" by Apolinário, Cordeiro & Zapp [9], has explored reclustering techniques - using

the reclustering techniques based on a formation time ordering variable. Moreover Jasmine Brewer, et al [12] also explored similar possibilities for the particular case of heavy quarks.

In this work, we build on these methodologies with a detailed Monte Carlo study in proton-proton (pp) collisions, which serves as our controlled baseline. We restrict our study to jets with identifiable $g \rightarrow c\bar{c}$ (gluon splitting to charm quark pairs), investigating how well reclustering methods (in particular, Cambridge-Aachen reclustering) can reproduce Monte Carlo's formation time distributions at both parton-level and hadron-level. We also examine the effects of initial-state radiation (ISR) and hadronization on reconstructive efficacy.

Additionally, our implementation uses both ROOT-based analysis and FastJet clustering tools, enabling detailed jet by jet matching of Monte Carlo versus reconstructed splittings. We perform a statistical comparison analysis, including quantile analyses, correlation studies, difference distributions, and robustness tests to quantify uncertainties and deviations. The ultimate goal is to assess whether the τ_{form} , in a heavy-quark system, can be a reliable observable in more realistic pp environments, thereby contributing to time-resolved tomography of the QGP.

The remainder of the essay is organised as follows. In Section 2 we describe the Monte Carlo samples, the jet clustering and reclustering pipeline (including ROOT analysis frameworks), selection criteria, important techniques and observable definitions. Section 3 presents comparisons of formation time (τ_{form}) distributions at different levels, quartiles, correlations, and systematic tests. We will also interpret these results, especially when using heavy-quark formation time as a medium probe, performing some statistical studies, and lastly presenting the limitations. Finally, Section 4 summarises the findings and discusses future work, including heavy-ion application and experimental mapping.

2 Simulation Setup

Our analysis is based on Monte Carlo simulations using PYTHIA 8.3 [13], a comprehensive event generator for high-energy particle physics. PYTHIA provides detailed modelling of: Hard scattering processes, Parton shower evolution, and Hadronization through the Lund string model as well as ISR and MultiParton Interactions (MPIs).

For this study, we used PYTHIA 8.306 proton-proton collision simulations with a centre-of-mass energy of $\sqrt{s} = 5.02$ TeV, allowing gluons to branch into quark-antiquark pairs with masses up to the charm threshold. Two sets of simulations were performed, either with the ISR and hadronisation processes turned on, allowing access to clean, yet overly simplified, parton-level events, or with both processes turned on, allowing both parton-level

and hadron-level analysis of the collisions. Although a fully rigorous analysis would require the inclusion of MPIs and detector effects, this simulation setup should already be reasonably indicative of the reconstruction power of heavy flavour splittings in vacuum systems [7]

Since our focus lies especially on tagging heavy quarks, the simulations were configured to only select jets with reconstructed transverse momentum larger than 200 GeV ($p_T > 200$ GeV) to increase the probability for heavy quark production, and with pseudorapidity $|\eta| < 1.9$ to restrict our study to the regions where detectors from actual particle colliders would be at.

As a way to reduce possible background contamination (i.e. influence of initial state radiation) in the reconstruction of jets while still not losing significant jet energy, a jet radius of 0.4 is employed ($R = 0.4$). Building on the former line of thought, a SoftDrop algorithm is also employed, as it grooms away soft wide-angle radiation, isolating the hard structure of jets. The SoftDrop algorithm works as follows:

$$\frac{\min(p_{Ti}, p_{Tj})}{p_{Ti} + p_{Tj}} > z_{cut} \left(\frac{\Delta R_{ij}}{R} \right)^\beta \quad (3)$$

where p_{Ti} and p_{Tj} are the transverse momentas of the two partons considered to form a splitting, $\frac{\Delta R_{ij}}{R}$ is their angular separation relative to the original jet radius and z_{cut} and β are adjustable grooming parameters [14].

For our study, we use $z_{cut} = 0.2$ and $\beta = 0$, meaning that in the process of reconstructing the substructure of the jet, we discard possible splittings where the softer parton has a fraction of energy from the parent parton smaller than 20%. The $g \rightarrow q\bar{q}$ splitting is usually a relatively even process in this case because the gluon tends to share its energy between the quark and antiquark about symmetrically, giving us confidence that the vast majority of $g \rightarrow c\bar{c}$ splittings will not be affected by the SoftDrop algorithm.

3 Results and Analysis

In this section, we will present and interpret the results obtained from the PYTHIA 8.3 simulations and compare truth-level and reconstructed-level observables in pp collisions. First, we will analyse the influence of the initial-state radiation (ISR) at the Monte Carlo level. For that, we will present the transverse momentum (p_T) of the jet initiating parton (gluon). Afterwards, we will analyse the formation time distributions reconstructed at both parton and hadron levels and present our statistical analysis of the reclustering method compared to the Monte Carlo simulations.

3.1 Initial state radiation

Although pp collisions do not produce a quark-gluon plasma (QGP) (i.e. no background medium is formed as a consequence of the collision), it is nevertheless necessary

to account for radiation emitted prior to the hard scattering. Such radiation, originating outside the jet cone, can contaminate the jet shower and consequently affect the measured observables. To investigate the influence of this effect on our chosen observable, we first analyze its impact on p_T of the jet-initiating parton, which in our case study is a gluon produced in the hard scattering and angularly aligned with a reconstructed jet.

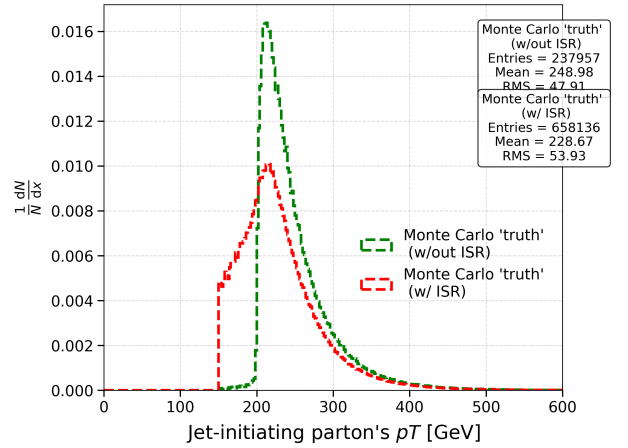


Figure 4: Comparison of the normalised jet-initiating parton's p_T distributions for Monte Carlo. The green dashed line represents the case without ISR, and the red dashed line represents the case including ISR. Statistical boxes with entries, mean, and RMS are displayed for both cases.

Figure 4 shows the distribution of the jet-initiating parton's p_T obtained from Monte Carlo simulations with Initial State Radiation (ISR) in red (w/ISR) and in green (w/out ISR). Both distributions exhibit a peak around $p_T \approx 200$ GeV, corresponding to the lower p_T threshold imposed on the jets. When ISR is enabled, the distribution becomes broader and develops a pronounced tail for $p_T < 200$ GeV. The tail is cut-off at 150 GeV, as our simulations imposed a lower cut of 150 GeV on the hard scattering scale to enhance the production of jets with $p_T > 200$ GeV without introducing distortions to the jet spectrum.

When ISR is active, contamination from partons originating outside the jet cone can occur at the stage where the anti- k_T algorithm identifies the jets. The low- p_T tail observed in Fig. 4 reflects this contamination, indicating additional energy is being reconstructed inside the jets which does not come from the jet-initiating parton, but instead from ISR emissions contaminating our jets.

3.2 Gluon p_T

To evaluate the performance of the reclustering method C/A in the presence of ISR, both before and after hadronization (i.e., at parton- and hadron-level, respectively), we analyze the transverse momentum distribution of the gluon that decays into a $c\bar{c}$ pair. We refer to this

as the gluon's p_T , and compare its behavior to that of the jet-initiating parton's p_T distribution.

For this purpose, ISR and hadronization were enabled, and a comparison plot was produced showing three distinct curves for the gluon p_T distribution: (i) the true Monte Carlo history of the jets, corresponding to the genuine gluon p_T distribution (in green); (ii) the reconstructed jets using the final-state partons before hadronization (in blue); and (iii) the reconstructed jets using the final-state hadrons after hadronization (in red), as shown in Fig. 5.

It is important to note that the parton-level case is not physically realistic, since experimental data only provide access to hadronization products. Nevertheless, this serves as a useful first approach for comparing MC simulations with the reclustering method.

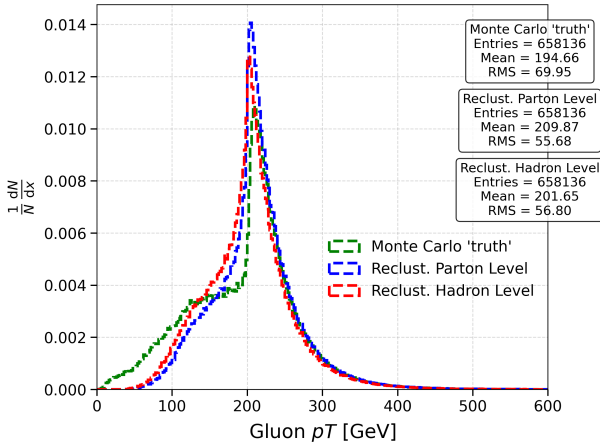


Figure 5: Distributions for gluon p_T in MC and at parton-level and hadron-level. The green dashed line represents MC simulations, while the blue one represents the Reclustering (C/A) method for parton-level, and the red one for hadron-level. Statistical boxes with entries, mean, and RMS are displayed for each curve.

All three distributions exhibit a peak at $p_T \approx 200$ GeV, consistent with the jet selection threshold. The Monte Carlo truth distribution displays a larger area for $p_T < 200$ GeV, while for $p_T > 200$ GeV, the overall shape resembles that observed in Fig. 4.

The peak near $p_T \approx 200$ GeV and the similarity with the distribution in Fig. 4 indicate that the majority of gluons producing $c\bar{c}$ pairs are indeed the jet-initiating gluons. This is expected, as the production of a $c\bar{c}$ pair requires a relatively high-energy gluon.

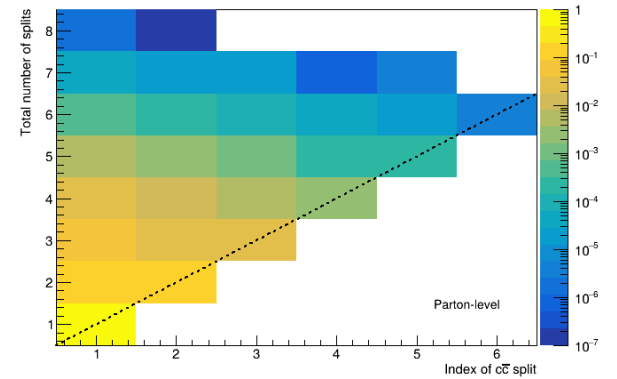
The larger area for $p_T < 200$ GeV in the Monte Carlo truth distribution, compared to both reclustering curves, can be attributed to the same effect that causes the tail observed in Fig. 4. Since reclustering is performed on the same jets found in the anti- k_T clustering process, it is also

affected by ISR contamination, which can lead to an over-estimation of the reconstructed gluon p_T . Consequently, the reclustered distributions appear narrower, as shown in Fig. 5.

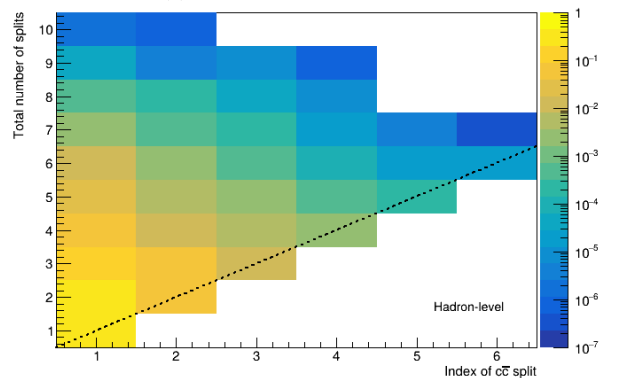
Between the two reclustering cases (parton- and hadron-level), the hadron-level curve exhibits slightly a larger area for $p_T < 200$ GeV. This suggests that hadronization induces an overall energy loss. During hadronization, partons that would contribute to the jet energy can hadronize with other partons in the event forming soft hadrons emitted slightly outside the jet radius, and thus are not included in the reconstructed jet.

Nonetheless, there is a close resemblance between the Monte Carlo truth and the reclustered distributions, both before and after hadronization, that demonstrates the robustness of the C/A reclustering method when analyzing the gluon p_T . This indicates that an angular-based algorithm can effectively be used to study heavy-quark observables such as this one.

To further compare reclustering at the hadron- and parton-levels, particularly regarding possible differences in splitting history, we produced a graphical representation of the splitting sequence for both cases, shown in Fig. 6.



(a) $c\bar{c}$ location at Parton-level



(b) $c\bar{c}$ location at Hadron-level

Figure 6: 2D distribution showcasing the location of the $c\bar{c}$ split inside the jet (according to its total number of splits), for parton-level and hadron-level.

The 2D distribution in Fig. 6 presents on the x -axis the index of the $c\bar{c}$ -pair-forming split (i.e., the splitting where the $c\bar{c}$ pair is resolved), and on the y -axis the total number of splittings along the main jet branch. A diagonal line is drawn to illustrate the upper-triangular nature of the distribution, since the $c\bar{c}$ pair formation always occurs within the jet, corresponding to a splitting at or earlier than the total number of splittings. The color scale represents the probability density, with bright yellow indicating the most likely configurations and dark blue the least likely.

Both distributions support our earlier hypothesis that the gluon forming the $c\bar{c}$ pair is predominantly the jet-initiating gluon. This is evident from the higher probability densities (warmer tones) associated with early splittings, and the decreasing densities (cooler tones) at later stages. The fluctuations observed in regions with many splittings, where this trend appears to break, are likely to contribute to the events where ISR dominates more.

The same effect (from warmer to colder tones) is observed along the y -axis: jets with fewer splittings exhibit higher probability densities. This behavior is expected since jets with a larger number of splittings are typically more energetic, and high- p_T jets are less frequently produced, as seen in Fig. 4. This comes also to explain why most of the emitting $c\bar{c}$ gluon's occur in the first splittings. There's a higher probability to form jets with lower p_T that, therefore, will have an overall lower total number of splits because of the fewer energy to distribute in each split. Since $c\bar{c}$ requires an energetic gluon, these jets can't radiate much before producing the $c\bar{c}$ pairs, which result in the high probability to radiate $c\bar{c}$ pairs in first splittings.

Comparing the two distributions, we find that the reclustering at hadron-level exhibits a higher probability density for $c\bar{c}$ formation in the first splitting of jets with two total splittings. When reclustering at hadron-level, additional splittings may appear simply because the analysis occurs at a later stage in the event's evolution. Nevertheless, both methods consistently show that the $c\bar{c}$ pair forms primarily in the first splitting, independent of the total number of splittings. These additional splittings explain the slightly higher density of jets with a higher total number of splits observed at hadron-level, particularly for jets with a relatively high total number of splits.

3.3 Formation Time

While these distributions provide valuable insight into $c\bar{c}$ formation and the general behavior of jet splittings, they do not quantitatively assess how accurately the C/A reclustering method reproduces the true formation times of gluons decaying into $c\bar{c}$ pairs. To evaluate the performance of the reclustering method in reconstructing this observable, as well as to study the effects of hadronization on the formation time, we analyzed the corresponding distributions,

shown in Fig. 7. The same conditions applied in the study of the gluon p_T were used here, with Fig. 7 plotting the Monte Carlo (MC) distribution alongside the reclustering results at both parton- and hadron-level.

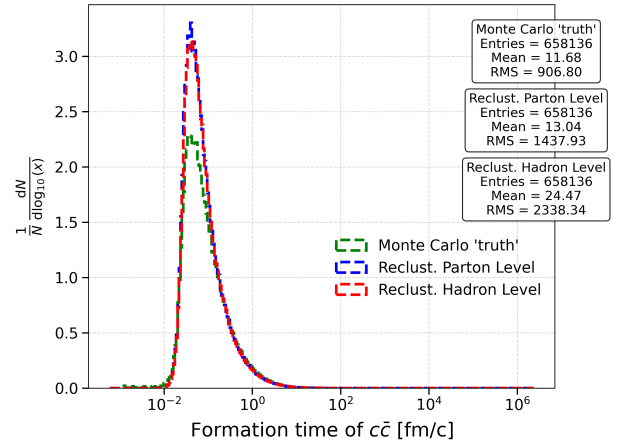


Figure 7: Comparison of formation time given by MC ($\tau_{\text{form}}^{\text{MC}}$) and the one obtained with the C/A reclustering ($\tau_{\text{form}}^{\text{Reclustering}}$) distributions at the parton- and hadron-levels. The x -axis is shown on a logarithmic scale. Statistical boxes with entries, mean, and RMS are also displayed for each curve.

The similar shapes of the three distributions suggest that the reclustering method, despite being based on an angle-ordered algorithm, performs well in reconstructing formation times. This indicates that angular ordering is a suitable proxy for estimating the formation time of heavy quarks.

As shown in Eq. 1, the formation time depends inversely on the gluon's p_T (i.e., $\tau_{\text{form}} \propto 1/p_T$). This dependence explains the presence of a peak in the formation time distribution, which corresponds to the peak observed in the gluon p_T distributions of Fig. 5. Moreover, the similarity in shape between the MC and reclustered curves at both parton- and hadron-levels implies that the reclustering method correctly reproduces the energy fraction of the softer emitted parton (c or \bar{c}) and the corresponding emission angle (the latter consistent with its angular-based nature).

Although all distributions agree on the general shape and peak position of the τ_{form} distribution, the MC distribution exhibits a lower τ_{form} probability density at the peak, indicating a broader overall form. To better analyse this difference, we produced a complementary plot with the logarithm of the probability density, shown in Fig. 8, which emphasizes differences in the lower formation time regions of the distributions.

From Fig. 8, we observe that the MC simulation includes lower formation times that are not captured by the reclustering methods and instead promoted to higher val-

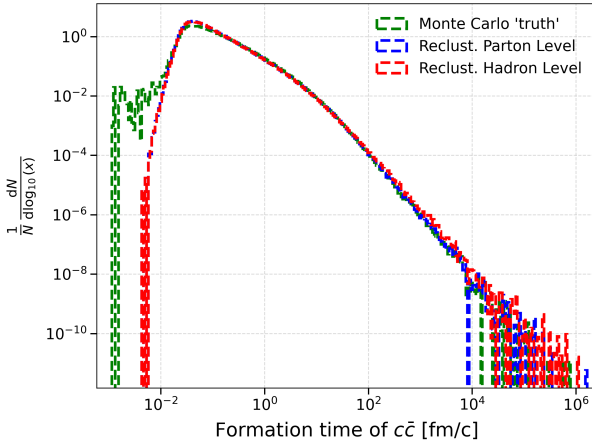


Figure 8: Comparison of $\tau_{\text{form}}^{\text{MC}}$ and $\tau_{\text{form}}^{\text{Reclustering}}$ distributions at the parton- and hadron-levels. Both axes are plotted on logarithmic scales to enhance visibility of low-probability regions.

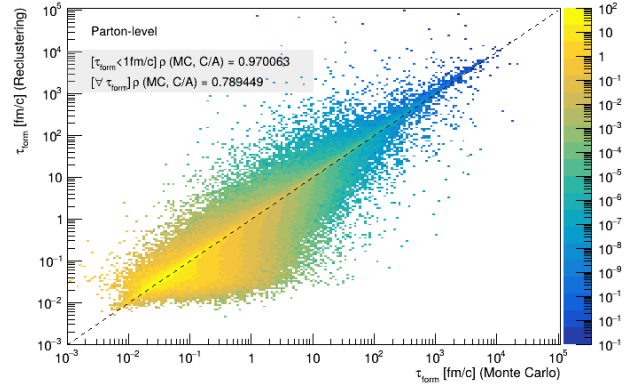
ues (along the peak). Since $\tau_{\text{form}} \propto 1/p_T$, smaller τ_{form} correspond to higher- p_T gluons.

3.3.1 Correlation in Formation Time

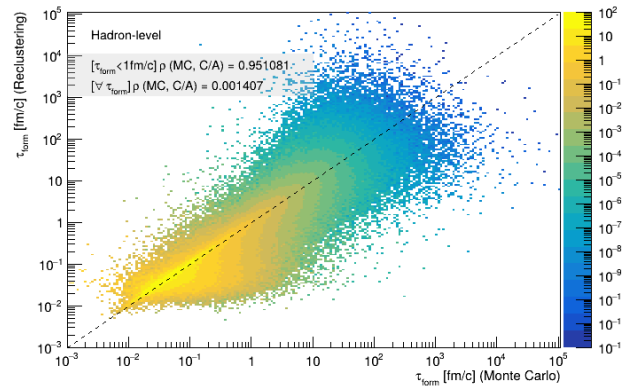
To assess the correspondence between true and reconstructed formation times more quantitatively, we constructed a 2D distribution comparing $\tau_{\text{form}}^{\text{MC}}$ and $\tau_{\text{form}}^{\text{Reclustering}}$, jet by jet, shown in Fig. 9. We analyzed correlations at both parton- and hadron-levels, shown in Figs. 9a and 9b, respectively. The color scale indicates the probability density, with warmer colors representing higher values. The x -axis corresponds to $\tau_{\text{form}}^{\text{MC}}$ and the y -axis to $\tau_{\text{form}}^{\text{Reclustering}}$. A diagonal line marks the locus of perfect correlation. For each case, two correlation coefficients are displayed: one considering $\tau_{\text{form}}^{\text{MC}} < 1 \text{ fm/c}$ and another considering the full range of $\tau_{\text{form}}^{\text{MC}}$.

In the range $10^{-3} < \tau_{\text{form}}^{\text{MC}} < 10^{-2} \text{ fm/c}$ both distributions exhibit an overestimation of τ_{form} by the reclustering method (i.e., for these values of $\tau_{\text{form}}^{\text{MC}}$, the reclustering over-estimates as previously seen in Fig. 8).

For $10^{-2} < \tau_{\text{form}}^{\text{MC}} < 10 \text{ fm/c}$, the bulk of the events in both distributions display a bump reflecting a systematic underestimation of τ_{form} by the reclustering method. In other words, for the aforementioned interval, the reclustering algorithm tends to yield smaller formation times compared to the Monte Carlo truth. Since $\tau_{\text{form}} \propto 1/p_T$, and given the larger area observed for the Monte Carlo distribution in the gluon p_T spectrum (see Fig. 5), such a bias was expected. Specifically, for $p_T < 200 \text{ GeV}$, the Monte Carlo distribution exhibits a greater number of events, implying that, on average, the true gluon p_T in MC is lower than that reconstructed through reclustering. Consequently, by the inverse relation between τ_{form} and p_T , a significant number of events satisfy $\tau_{\text{form}}^{\text{MC}} > \tau_{\text{form}}^{\text{reclustering}}$, consistent with the bias observed in the results.



(a) Correlation at parton-level.



(b) Correlation at hadron-level.

Figure 9: Correlation of τ_{form} between MC simulation and the reclustering method (C/A) at parton- and hadron-levels. The correlation coefficients are shown for $\tau_{\text{form}}^{\text{MC}} < 1 \text{ fm/c}$ and for the full range of $\tau_{\text{form}}^{\text{MC}}$.

At $\tau_{\text{form}}^{\text{MC}} > 10 \text{ fm/c}$, the hadron-level correlation shows increased smearing (region with relatively low-probability densities), indicating that hadronization has a more pronounced effect for longer formation times, corresponding to lower- p_T gluons. Nonetheless, at these scales hadronization effects are also expected to dominate the jet substructure kinematics (these are likely to occur at a later stage of the clustering sequence).

To quantify the reclustering performance, we first computed correlation coefficients over the full τ_{form} range. The resulting values were lower than expected, particularly at hadron-level, despite the clearly visible correlation for $\tau_{\text{form}} < 1 \text{ fm/c}$, by the form of a bright yellow diagonal. Restricting the analysis to $\tau_{\text{form}} < 1 \text{ fm/c}$ produced correlation values more consistent with visual observations from Fig. 9.

The strong dependence of the correlation coefficient on the chosen $\tau_{\text{form}}^{\text{MC}}$ range suggests that this metric may not reliably capture the reclustering performance across the full distribution. This is due to the wide range of τ_{form} values (best visualized on a logarithmic scale), which

biases the correlation towards high- τ_{form} regions with little statistical significance. We also note that for QGP-like studies we are mainly interested in $\mathcal{O}(10 \text{ fm}/c)$ To mitigate this effect, we instead examined the quartiles of the distribution of differences between MC and reclustered formation times.

The difference in formation time, $\Delta\tau_{\text{form}}$, is defined as:

$$\Delta\tau_{\text{form}} = \tau_{\text{form}}^{\text{MC}} - \tau_{\text{form}}^{\text{Reclustering}}, \quad (4)$$

where $\tau_{\text{form}}^{\text{MC}}$ corresponds to the formation time given by MC-truth and $\tau_{\text{form}}^{\text{Reclustering}}$ the one obtained with the reclustering method. This difference allows an jet by jet comparison between both methods while maintaining statistical quantification via the distribution quartiles. This is illustrated in Fig 10 for Parton level (red line) and Hadron level (blue line).

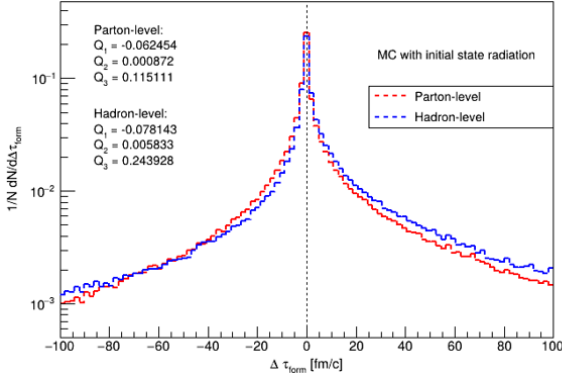


Figure 10: Distribution of the difference in τ_{form} between MC-truth and the reclustering method. The red curve corresponds to the parton-level and the blue curve to the hadron-level. Quartile values (Q_1 , Q_2 , Q_3) are shown in the legend.

Ideally, the $\Delta\tau_{\text{form}}$ distribution should be narrow and centered around zero and indeed this is the case. However, as anticipated from the bump seen in Fig. 9, the reclustering method exhibits a slight bias toward smaller formation times, resulting in a small positive shift (since $\Delta\tau_{\text{form}} > 0 \implies \tau_{\text{form}}^{\text{MC}} > \tau_{\text{form}}^{\text{Reclustering}}$). This shift is more pronounced at hadron-level, consistent with the broader smearing observed in Fig. 9.

From these distributions, we extracted the quartiles Q_1 , Q_2 , and Q_3 , corresponding to the 25th, 50th (median), and 75th percentiles, respectively. Their values are indicated in Fig. 10. To visualize the relative widths of both distributions, Fig. 11 displays the quartile ranges graphically with $-Q_1$ and Q_3 as asymmetric bars at Q_2 as a red marker.

Both distributions show median values (Q_2) close to zero, consistent with the correlation observed in Fig. 9. The relatively small quartile widths indicate that the reclustering method reproduces the true τ_{form} with good

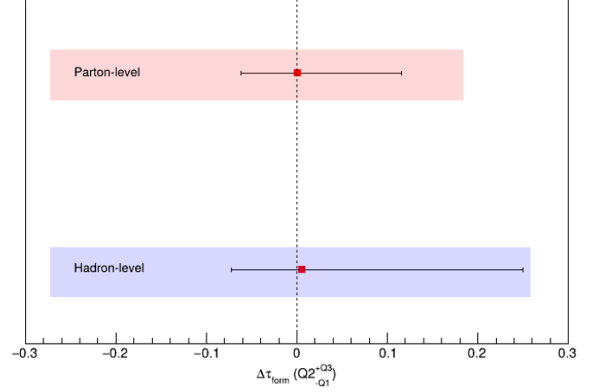


Figure 11: Visualization of the median (Q_2 , red squares) and quartile ranges (Q_1 , Q_3) for $\Delta\tau_{\text{form}}$ distributions. Red shading represents the parton-level results and blue the hadron-level results.

accuracy for most events. The asymmetry of the upper quartile ($+Q_3$) reflects the bias toward $\tau_{\text{form}}^{\text{MC}} > \tau_{\text{form}}^{\text{Reclustering}}$, confirming the slight underestimation of τ_{form} by the reclustering method.

Overall, the small quartile spreads and median values near zero demonstrate that the reclustering method provides an accurate reconstruction of the gluon formation time associated with $c\bar{c}$ emission. The remaining discrepancies between parton- and hadron-level results stem from the limitations of a realistic proton-proton environment.

4 Conclusions

To conclude, we saw that the combination of anti-kt clustering for jet identification and Cambridge/Aachen reclustering for temporal reconstruction successfully reproduces Monte Carlo formation time distributions at both parton- and hadron-levels, with reasonable accuracy.

Furthermore, Fig. 9 confirms that the reclustering method (C/A) tends towards smaller formation times w.r.t the MC-truth.

Key limitations include that our work is based on MC event simulations, so some of the elements we studied cannot be tested experimentally, since we cannot physically access them, like partonization, for instance. Nevertheless, the consistency of the results from parton to hadron level provide a good indication of the robustness of the proposed methods. We only studied one reclustering choice (C/A), which is a physics-motivated one but not unique; other algorithms may perform differently.

4.1 Future Work

The next steps in our work are:

- Extending the procedure to study the performance of these reclustering methods for heavy ion collisions, like Pb-Pb.
- Explore other reclustering techniques to see whether they perform better in hadronised environments.
- Comparison with experimental data to validate these findings and establish formation time as a measurable quantity in heavy-ion physics

5 Acknowledgments

None of this work would have been attainable without the reading material, assistance with new software, guidance in QCD and high-energy collisions and support provided by Dr Liliana Apolinário and Nuno O. Madureira, to whom we are very grateful.

References

- [1] W. Busza, K. Rajagopal, W. van der Schee, *Ann. Rev. Nucl. Part. Sci.* **68**, 339 (2018), arXiv:1802.04801
- [2] J.W. Harris, B. Müller, *Annual Review of Nuclear and Particle Science* **46**, 71 (1996), arXiv:hep-ph/9602235
- [3] J.A. et al. (STAR Collaboration), *Nuclear Physics A* **757**, 102 (2005)
- [4] K.A. et al. (PHENIX Collaboration), *Nuclear Physics A* **757**, 184 (2005)
- [5] H.B. Meyer, *Progress in Particle and Nuclear Physics* **68**, 4 (2011), arXiv:1104.3708
- [6] L. Apolinário, Y.J. Lee, M. Winn, *Prog. Part. Nucl. Phys.* **127**, 103990 (2023), arXiv:2203.16352
- [7] L. Apolinário, *Eur. Phys. J. C* **82**, 244 (2022), arXiv:2111.05794
- [8] G. Salam, *Eur. Phys. J. C* **67**, 637 (2010), arXiv:0906.1833
- [9] *Phys. Rev. D* **81**, 106 (2022), arXiv:2012.02199
- [10] M. Cacciari, G. Salam, G. Soyez, *Journal of High Energy Physics* p. 15 (2008)
- [11] N.O. Madureira, Master's thesis, Instituto Superior Técnico / LIP (2022), master's thesis, LIP / IST
- [12] J. Brewer, W. van der Schee, U. Wiedemann (2025), arXiv:2503.11764
- [13] C. Bierlich et al., *SciPost Phys. Codebases* **8** (2022), arXiv:2203.11601
- [14] G.S.J.T. Andrew J. Larkoski, Simone Marzani, *Journal of High Energy Physics* p. 46 (2014), arXiv:1402.2657