České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská

Katedra fyziky
Studijní program: Jaderná a částicová fyzika

Identifikace c-jetů v p+p a A+A srážkách pomocí strojového učení

Identification of c-jets in p+p and A+A collisions with Machine Learning

DIPLOMOVÁ PRÁCE

Vypracoval: Bc. Jitka Mrázková
Vedoucí práce: doc. RNDr. Jana Bielčíková, Ph.D.
Rok: 2023

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA JADERNÁ A FYZIKÁLNĚ INŽENÝRSKÁ
PRAHA 1 - STARÉ MĚSTO, BŘEHOVÁ 7 - PSČ 115 19

Katedra: fyziky *Akademický rok:* 2021/2022

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: Bc. Jitka Mrázková

Studijní program: Jaderná a částicová fyzika

Název práce: (česky) Identifikace c-jetů v p+p a A+A srážkách pomocí strojového učení

Název práce: (anglicky) Identification of c-jets in p+p and A+A collisions with machine learning

Pokyny pro vypracování:

1) Rešerše experimentálních výsledků produkce jetů obsahujících těžké kvarky v p+p a A+A srážkách na urychlovačích RHIC a LHC.

2) Přehled metod strojového učení se zaměřením na lokálně agregované deskriptory a JetVLAD model.

3) Studium efektivity výběru jetů obsahujících těžký kvark (c-jetů) v JetVLAD modelu na simulovaných datech z p+p srážek při energiích urychlovače RHIC a porovnání s energiemi dosažitelnými na LHC.

4) Studium efektivity výběru c-jetů na simulovaných datech A+A srážek při energii urychlovače RHIC pomocí JETSCAPE.

5) Diskuse získaných výsledků.

Doporučená literatura:

[1] A. Zhang, Z. C. Lipton, M. Li and A. Smola: Dive into Deep Learning, 2020, interaktivní open source učebnice: https://d2l.ai

[2] T. Sjöstrand, S. Mrenna and P. Skands: PYTHIA 6.4 Physics and Manual JHEP05 (2006) 026 and Comput. Phys. Comm. 178 (2008) 852

[3] J. H. Putschke, et al.: The JETSCAPE Framework, arXiv: 1903.07706

[4] J. Bielčíková, et al.: Identifying Heavy-Flavor Jets Using Vectors of Locally Aggregated Descriptors, JINST 16 (2021) 03, P03017

Jméno a pracoviště vedoucího diplomové práce:

RNDr. Jana Bielčíková, Ph.D.,
Katedra fyziky, Fakulta jaderná a fyzikálně inženýrská ČVUT v Praze a Ústav jaderné fyziky AV ČR, v.v.i.

Datum zadání diplomové práce: 20.10.2021

Termín odevzdání diplomové práce: 02.05.2022

Doba platnosti zadání je dva roky od data zadání.

.....
garant studijního programu

.....
vedoucí katedry

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE Fakulta jaderná a fyzikálně inženýrská -1-

.....
děkan

V Praze dne 20.10.2021

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta jaderná a fyzikálně inženýrská

Břehová 7
115 19 Praha 1

P R O H L Á Š E N Í

Já, níže podepsaný

Jméno a příjmení studenta:	Jitka Mrázková
Osobní číslo:	473934
Název studijního programu (oboru):	Jaderná a částicová fyzika

prohlašuji, že jsem diplomovou práci s názvem:

Identifikace c-jetů v p+p a A+A srážkách pomocí strojového učení

vypracoval(a) samostatně a uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne 3.5.2023 .....

podpis

Acknowledgment

I would like to thank my supervisor doc. RNDr. Jana Bielčíková, Ph.D. for her patient guidance of this diploma thesis and for her valuable advice. Also many thanks to Raghav Kunnawalkam Elayavalli, Ph.D for valuable consultations in the area of particle physics, to Ing. Georgij Ponimatkin for consultations in the area of machine learning and help with software setup and to Ing. Jan Vaněk for help with implementation of EvtGen software.

Bc. Jitka Mrázková

Název práce:
Identifikace c-jetů v p+p a A+A srážkách pomocí strojového učení

Autor: Bc. Jitka Mrázková

Studijní program: Jaderná a částicová fyzika
Druh práce: Diplomová práce

Vedoucí práce: doc. RNDr. Jana Bielčíková, Ph.D.
Katedra fyziky, Fakulta jaderná a fyzikálně inženýrská ČVUT v Praze a Ústav jaderné fyziky AV ČR, v.v.i.

Abstrakt: Jety – sprškы kolimovaných částic – nacházejí široké využití v oblasti fyziky vysokých energií. Tato práce je zaměřena na identifikaci jetů pocházejících z těžkých kvarků. Za účelem identifikace jetů bude využit JetVLAD taggovací model založený na metodách strojového učení. Schopnost modelu extrahovat vzorky jetů je testována na několika odlišných souborech dat a s použitím různých taggovacích přístupů. Metoda bude nejprve demonstrována na souborech dat z PYTHIA generátoru pro proton-protonové srážky při energiích 200 GeV a 510 GeV v těžišťovém systému, s diskuzí o možném rozšíření na vyšší energie urychlovače LHC v CERN. Závěr této práce bude věnován možné aplikaci modelu JetVLAD na simulace srážek těžkých iontů z generátoru JETSCAPE.

Klíčová slova: jetová fyzika, c kvarky, kvark-gluonové plazma, strojové učení

Title:
Identification of c-jets in p+p and A+A collisions with Machine Learning

Author: Bc. Jitka Mrázková

Abstract: Jets of collimated particles are used in a wide range of analyses in high energy physics. This work is focused on identifying jets originating from heavy quarks. For this purpose, we use the recently introduced JetVLAD tagging model based on machine learning techniques. We show the performance of this model on different datasets and for different tagging approaches. The method is demonstrated on PYTHIA generated proton-proton collisions at center-of-mass energies 200 and 510 GeV, with a discussion of possible extension to higher energies available at the LHC at CERN. In the end of this work, we will discuss the possible applicability of the JetVLAD model to heavy-ion collision simulations from the JETSCAPE generator.

Key words: jet physics, charm quarks, quark-gluon plasma, machine learning

Contents

List of Figures

Introduction

Jets, collimated sprays of energetic hadrons, represent one of the key observables in high energy physics. They are used in a broad range of analyses in experimental particle physics. For example, jets can be used to test perturbative quantum-chromodynamics (QCD) predictions in proton-proton collisions or to explore various effects of the quark-gluon plasma (QGP) on particle production in heavy-ion collisions.

In this work, we are particularly interested in the identification of heavy-flavor jets that originate from c or b quarks. As a tool for tagging heavy-flavor jets, we utilize the recently introduced JetVLAD model [13] based on machine learning. This model uses the information contained within the jet constituents by taking charged-particle jet constituents as an input and aggregating them into a descriptor vector that characterizes it. The analysis part of this work will be focused on the study of the JetVLAD model performance in simulated $p+p$ collisions at lower collision energies, namely $\sqrt{s} = 200$ GeV and $\sqrt{s} = 510$ GeV, which are available at RHIC (Relativistic Heavy Ion Collider). Then, an outlook on the center-of-mass energy of $\sqrt{s} = 7$ TeV, which is more typical for the measurements at the LHC (Large Hadron Collider), will be provided. Lastly, the applicability of the JetVLAD model on JETSCAPE generated heavy-ion collisions will be discussed.

This thesis is structured as follows. The first chapter serves as an introduction to the physics of jet measurements including description of important experimental observables and measurements. The second chapter represents a brief introduction to jet reconstruction algorithms and their properties. In the third chapter, the basics of machine learning with prerequisites to the JetVLAD model are introduced. Next chapter is dedicated to the study of the JetVLAD model performance in $p+p$ collisions and discussion of the results. In the last chapter, the JETSCAPE event generator will be introduced and the applicability of the JetVLAD model on these data will be discussed.

Chapter 1

Recent results of jet measurements

Hard scattering processes in high energy collisions lead to the production of particle showers. Typical showers of energetic hadrons originating from scattered partons that have undergone fragmentation and hadronization are called jets.

Jets play an important role in many areas of particle physics. For example, in proton-proton collisions, they provide important tests for quantum-chromodynamics (QCD) calculations. In heavy-ion collisions, jets can be used in tomographic studies of the quark-gluon plasma (QGP).

In this chapter, we would like to demonstrate the importance of jets by presenting some of the recent results of jet measurements in proton-proton and heavy-ion collisions. Measurements from experiments at different collision energies from both RHIC (Relativistic Heavy Ion Collider) at Brookhaven National Laboratory (BNL) and the LHC (Large Hadron Collider) at CERN will be discussed.

1.1 Jets in $p + p$ collisions

As already mentioned, jets can serve as a tool for testing the predictions of QCD calculations in proton-proton collisions. Figure 1.1 presents a measurement of double-differential inclusive jet cross-section as a function of the jet p_{T} for each jet rapidity (y) bin from the ATLAS experiment at the LHC at center-of-mass energy of $\sqrt{s} = 8$ TeV [1]. The measured cross-sections are compared to the QCD predictions calculated at next-to-leading order (NLO) in perturbation theory, which are corrected for non-perturbative and electroweak effects. Overall, the measurements show a good agreement with the predictions of perturbative quantum-chromodynamics calculations at next-to-leading order (NLO).

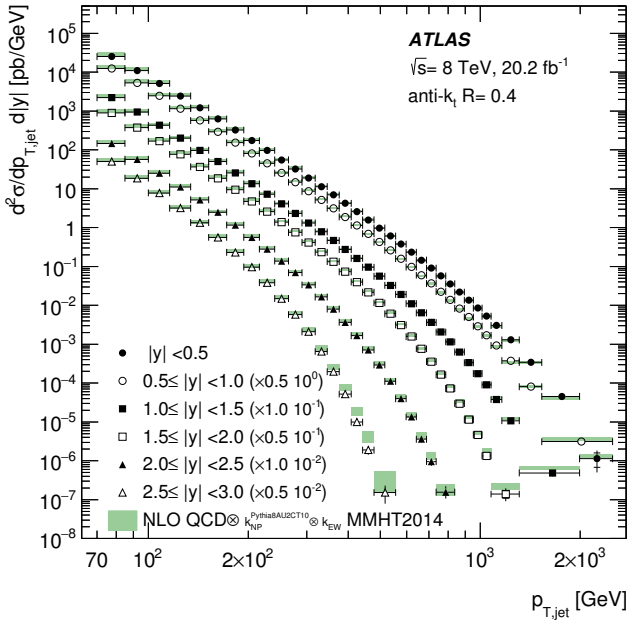

Figure 1.1: Inclusive jet cross-section as a function of jet p_{T} in different bins of jet rapidity (y) from the ATLAS experiment at the LHC at $\sqrt{s} = 8$ TeV. The cross-sections are multiplied by the factors indicated in the legend for better visibility. The data are compared to the next-to-leading-order (NLO) quantum-chromodynamics predictions. Taken from [1].

The topic of this work, especially the analysis part, is dedicated to tagging heavy-flavor jets. The term heavy flavor refers to jets originating from c or b quarks, which are produced early in the hard scattering and travel a significant distance before decaying. Therefore, we shall mention one of the effects of mass/flavor dependence in jets.

The partons produced in particle interactions with large momentum transfer undergo subsequent emissions that can lead to a cascade process known as a parton shower. The patterns of the parton showers are expected to depend on the mass of the initiating parton as described in QCD theory by the so-called dead-cone effect [18], [19]. The studies of the dead-cone effect suggest that the gluon bremsstrahlung of heavy quarks differs from that of light quarks since it is expected to be suppressed below a certain angle $\theta \sim m/E$ relative to the direction of the emitting quark with mass m and energy E. As a consequence, heavier partons (c and b quarks) are assumed to suffer lower radiative energy loss in this region than the light quarks.

Figure 1.2 demonstrates the first direct observation of the dead-cone effect in proton-proton collisions at $\sqrt{s} = 13$ TeV measured by the ALICE experiment at the LHC by comparing the angular distribution of splittings for D^0–tagged jets to that of inclusive jets [2]. The ratios of the angular distribution of splittings are shown for three radiator (c quark) energy intervals $5 < E_{\text{Radiator}} < 10$ GeV, $10 < E_{\text{Radiator}} < 20$ GeV, and $20 < E_{\text{Radiator}} < 35$ GeV. There is a significant suppression of small-angle splittings in D^0–tagged jets suggesting the presence of dead-cone effect for charm quark emissions, which is most visible at lower radiator energy. This analysis was done by using iterative declustering techniques that allow to access the splittings at the smallest angles in deep levels of the clustering history.

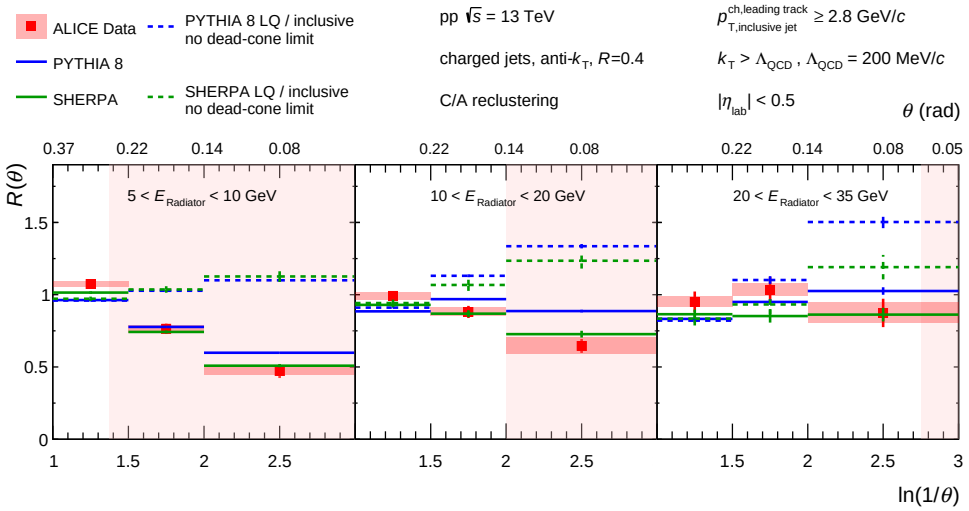

Figure 1.2: The measurement of the dead cone effect. Ratio of the angular distribution of splittings for D^0–tagged jets vs inclusive jets, $R(\theta)$, in $p + p$ collisions at $\sqrt{s} = 13$ TeV shown for three different radiator energy (E_{Radiator}) intervals. The data were compared with MC calculations from SHERPA and PYTHIA generators. Taken from [2].

1.2 Jets in heavy-ion collisions

Collisions of heavy ions at high energies are accompanied in the first moments by the creation of hot dense matter, which is called quark-gluon plasma (QGP). This dense strongly interacting medium of weakly bound partons is believed to have existed in extreme conditions at the beginning of the universe and may be accessed in the high energy colliders.

The quark-gluon plasma state exists after the heavy-ion collision for only a brief moment before the matter starts to hadronize. However, information about the properties of QGP can be obtained indirectly by using the so-called hard probes. The term hard probe usually refers to particles that are produced in processes with large momentum transfer on very short time scales ($\tau \sim 0.1$ fm/c), whose behavior can be described by the perturbative QCD theory. For example, jets and heavy quarks (c and b) can act as hard probes in the medium as they are produced in the initial stages of the collision and then pass through all stages of QGP evolution [20].

Figure 1.3 shows a simple illustration of the difference in production of jets depending on the presence or absence of the quark-gluon plasma in collisions. The left figure shows a proton-proton collision, where the QGP is not formed and the jets propagate freely in the vacuum. However, in the figure on the right, the jets from heavy-ion collision that undergo scattering in the medium are suppressed compared to those in vacuum.

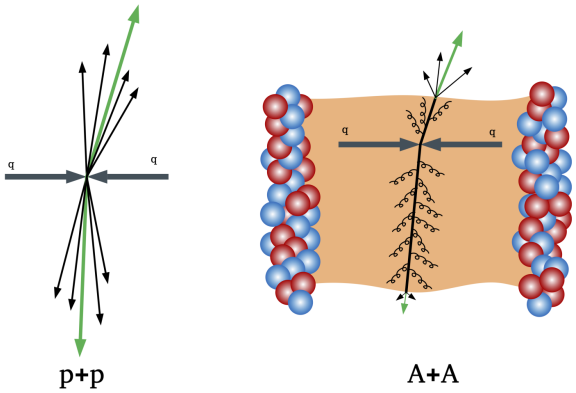

Figure 1.3: Illustration of the jet quenching effect. Jet production in proton-proton collision (left) and heavy-ion collision (right). The incoming quarks (q) are scattered off each other in the interaction and the outgoing jets of particles are shown here as arrows. In heavy-ion collision the jets travel through the QGP (orange region) and are suppressed as opposed to the jets in proton-proton collision. Taken from [3].

Overall, the interaction of hard probes with the medium and the transfer of energy during their passage through the medium can bring us valuable information about the QGP and the initial conditions during the collision. In jets, these interactions can result in experimentally observable effect commonly reffered as jet quenching, which can manifest itself as a suppression of particle production at large transverse momenta, modification of di-hadron correlations or modification of properties of reconstructed jets.

To quantify the effects of jet quenching on particle production in experimental analyses, following two observables are commonly used – nuclear modification factor R_{AA} ([4], [5]), and central-to-peripheral nuclear modification factor R_{CP} ([7], [21]).

The nuclear modification factor R_{AA} describes the suppression of particle production in nucleus-nucleus collisions with respect to the particle production in proton-proton collisions:

$$R_{AA} = \frac{1}{\langle N_{coll} \rangle} \frac{\mathrm{d}^2 N_{AA}/\mathrm{d}p_\mathrm{T} d\eta}{\mathrm{d}^2 N_{pp}/\mathrm{d}p_\mathrm{T} d\eta}, \tag{1.1}$$

where $\langle N_{coll} \rangle$ stands for a scaling factor representing an average number of binary nucleon-nucleon collisions based on the Glauber model predictions[1]. N_{AA} and N_{pp} are the numbers of particles produced in $A + A$ and $p + p$ collisions, respectively. They are both functions of transverse momentum p_T and pseudorapidity η.

In cases where no comparison with $p + p$ collisions is provided, the definition of central-to-peripheral nuclear modification factor R_{CP} can be useful. This method of observing suppression (or enhancement) of particle production is defined as:

$$R_\mathrm{CP} = \frac{\langle N^\mathrm{P}_\mathrm{coll} \rangle}{\langle N^\mathrm{C}_\mathrm{coll} \rangle} \frac{\mathrm{d}^2 N^\mathrm{C}_{AA}/\mathrm{d}p_\mathrm{T} d\eta}{\mathrm{d}^2 N^\mathrm{P}_{AA}/\mathrm{d}p_\mathrm{T} d\eta}. \tag{1.2}$$

In this equation, we compare particle spectra from central (C) and peripheral (P) nucleus-nucleus collisions scaled by $\langle N^\mathrm{C}_\mathrm{coll} \rangle$ and $\langle N^\mathrm{P}_\mathrm{coll} \rangle$. In central collisions, measured particles have small impact parameter and their mean pathlength through the medium is presumably longer. The particles from peripheral collisions are expected to have shorter in-medium pathlengths and therefore lose less energy.

The values of the nuclear modification factors R_AA and R_CP carry information about the enhancement or suppression of particle production in the experiment. $R_\mathrm{AA} > 1$ typically signifies enhanced particle production, whereas $R_\mathrm{AA} < 1$ means that particle production is suppressed (similarly for the R_CP factor.)

[1] See Appendix A or studies in [22], [15] for more information about Glauber model.

Let us now present several examples of the nuclear modification factor measurements using jets. Figure 1.4 shows a measurement of the nuclear modification factor R_{AA} from the ATLAS experiment at the LHC at center-of-mass energy of $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV per nucleon-nucleon pair [4] for resolution parameter R = 0.4. The results indicate a strong suppression of the jet production in Pb+Pb collisions, which is observed up to large jet transverse momentum (p_{T}). The most significant suppression of R_{AA} is present in the centrality bin $0-10\%$ (most central Pb+Pb collisions.)

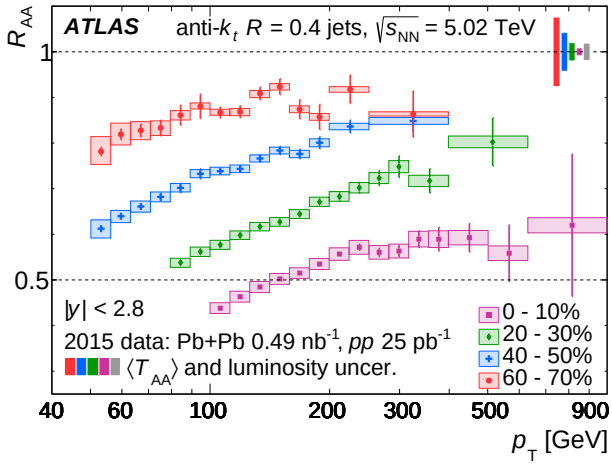

Figure 1.4: The R_{AA} values as a function of jet p_{T} for four centrality intervals measured in $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV Pb+Pb collisions by the ATLAS experiment. Taken from [4].

Another example of jet suppression measurement, this time from the CMS experiment at the LHC at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV [5], is depicted in Figure 1.5. Here the plots show the dependence of R_{AA} on the jet p_{T} for reconstructed jets with various jet radii R and for the first time up to a radius $R = 1.0$. The data manifest a strong suppression of high-p_{T} jets reconstructed with all distance parameters, implying that a significant amount of jet energy is scattered to large angles, even beyond $R = 1.0$.

Recent results reported by ALICE [6] (see figures 1.6 and 1.7) are using machine learning (ML) techniques to improve background subtraction for jets in heavy-ion collisions. The improvement in background subtraction of novel ML approach over the standard ALICE method is shown on the left plot in Figure 1.6. There is a reduction in the standard deviation of the δp_{T} distribution, which is most significant for central collisions and large R. Figure 1.7 (left) shows the results for R_{AA} of jets with resolution parameter $R = 0.6$ compared to different model predictions. This measurement successfully allowed to extend the reach of measurement in both p_{T} and R. The right plot in Figure 1.7 depicts a ratio of R_{AA} for jets with different resolution parameters $R = 0.6$ and $R = 0.2$. Here the larger R jets appear to be more suppressed, which suggests an R-dependence of jet quenching.

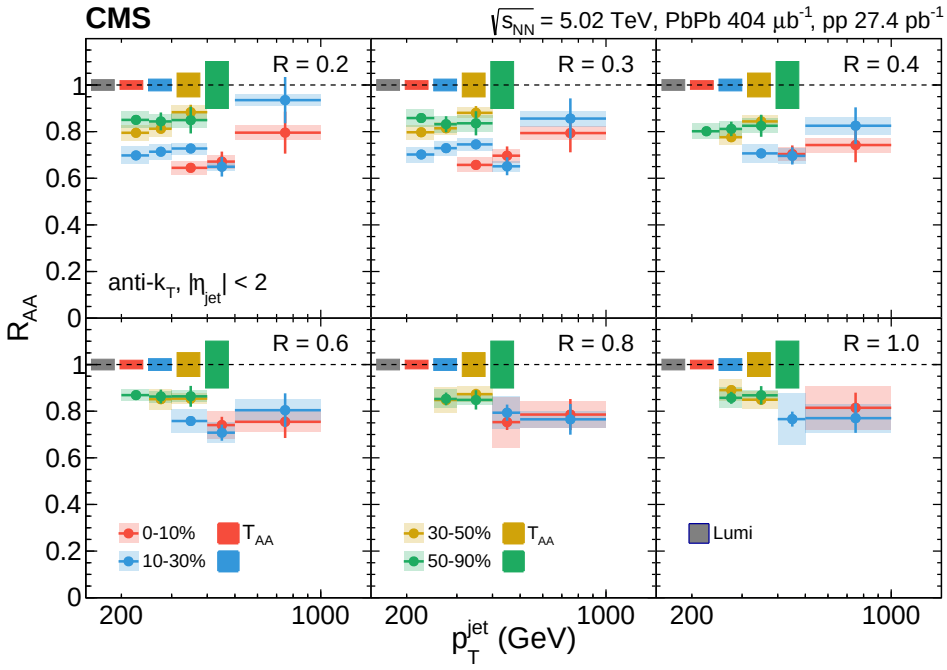

Figure 1.5: The R_{AA} for jets as functions of p_{T}^{jet} for various R and centrality intervals from the CMS experiment at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV. The statistical uncertainties are represented by vertical lines, the systematic uncertainties by shaded boxes and the global uncertainties as colored boxes on the dashed line. Taken from [5].

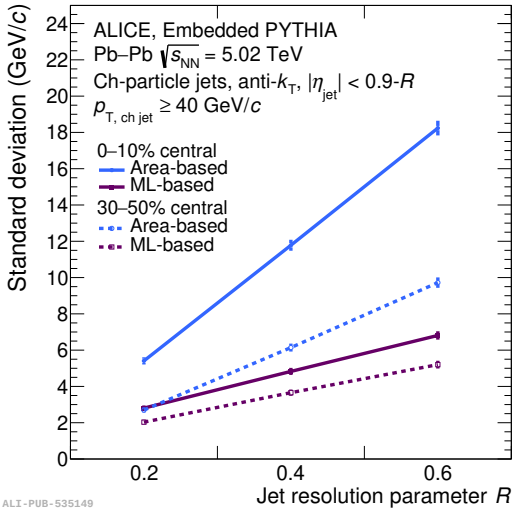

Figure 1.6: Improvement of background subtraction in R_{AA} measurement with ML based ALICE approach compared to traditional techniques. Standard deviation of the δp_{T} distribution (difference between the reconstructed and measured p_{T} of jet) as a function of jet resolution parameter R. Taken from [6].

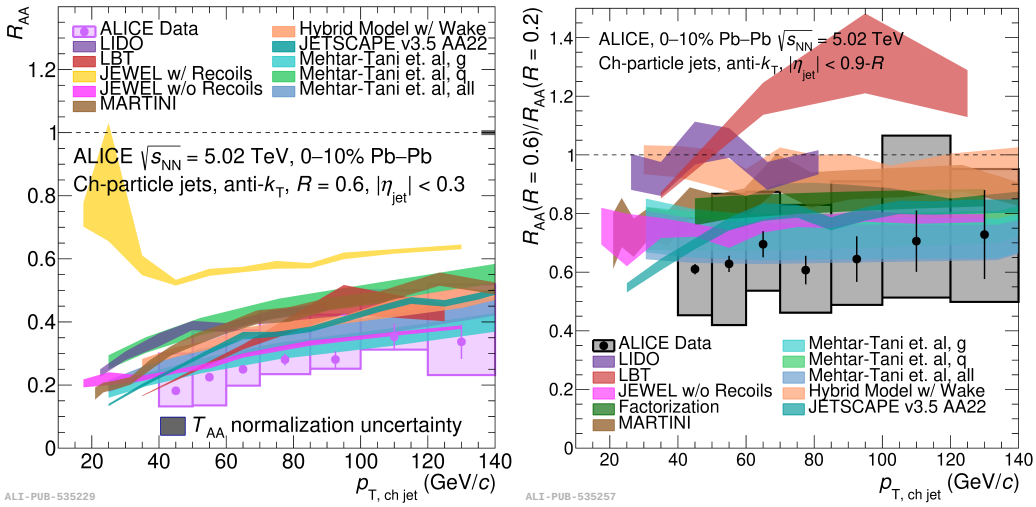

Figure 1.7: (Left:) R_{AA} for jets with resolution parameter R=0.6 in central collisions. (Right:) Ratio of jet R_{AA} using $R = 0.6$ as the numerator and $R = 0.2$ as the denominator. The measurements are compared to different model predictions. Taken from [6].

As the last example of jet suppression measurements in this section, it is convenient to compare nuclear modification factor measurements from the LHC and RHIC. Figure 1.8 shows a comparison of the R_{CP} factor measurements for charged-particle jets and hadrons in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV (RHIC) and Pb+Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV (LHC). We can see a visible suppression of particle production in the medium and the same effect is present in the production of reconstructed jets. Although the conditions in these two colliders are different as the temperature and density of the created QGP are more extreme at the LHC, the results showed up to be quantitatively comparable for both RHIC and LHC energies.

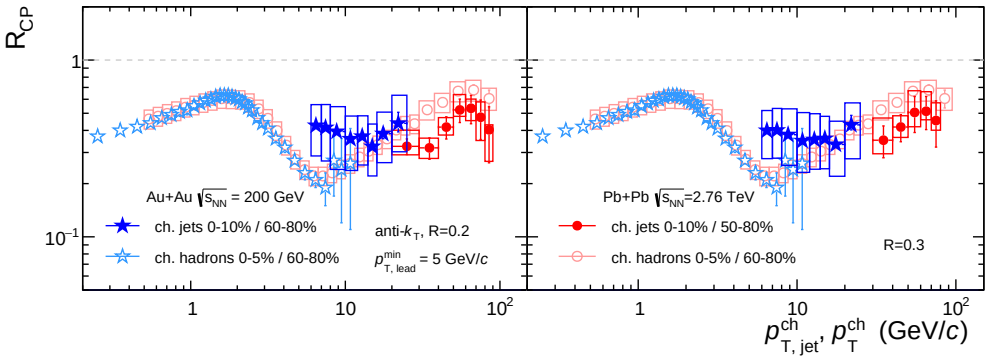

Figure 1.8: Comparison of the charged hadron and charged-particle jet R_{CP} factors in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV (RHIC) and Pb+Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV (LHC). The results are shown for two different jet resolution parameters $R = 0.2$ and $R = 0.3$. From [7].

One of the effects of jet quenching in the medium is the modification of properties of the reconstructed jets. Changes in the jet substructure show up, for example, when measuring the radial distribution of heavy flavor particle production in jets. The first study dedicated to the charm quark diffusion with respect to the jet axis in heavy-ion collisions was conducted by the CMS collaboration (LHC) [8]. The results in Figure 1.9 show radial distribution of D^0 mesons in jets as a function of the distance from the jet axis. When comparing Pb+Pb collisions to the $p+p$ results, the D^0 meson distribution for D^0 mesons with transverse momentum in the range of $4 < p_T^D < 20$ GeV/c indicates on average a larger distance with respect to the jet axis. At higher p_T^D, the Pb+Pb and $p+p$ radial distributions appear to be similar. This could be attributed to the effects of medium quenching of charm quarks at lower p_T.

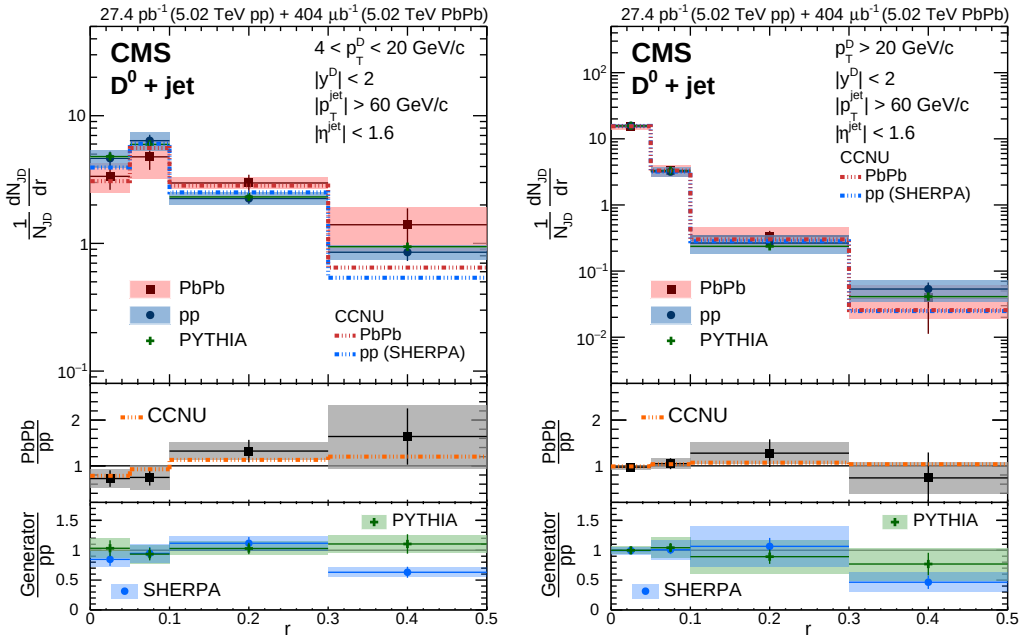

Figure 1.9: Distributions of D^0 mesons in jets as a function of the distance from the jet axis (r) measured in $p+p$ and Pb+Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV by the CMS experiment. The measurement is performed in the D^0 meson transverse momentum range $4 < p_T^D < 20$ GeV/c (left) and $p_T^D > 20$ GeV/c (right). The ratios of the D^0 meson radial distributions in Pb+Pb and $p+p$ collisions are shown in the middle panels and the bottom panels show the ratios of the D^0 meson radial distributions of $p+p$ over MC event generators. Taken from [8].

Another measurement that investigates jet shape modification, but at lower collision energy, was performed by the STAR experiment at RHIC in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV [9]. Here, the measurement is focused on low p_{T} D^0–tagged jets. The results in Figure 1.10 (left) show that the radial profile of D^0 in jets is consistent for different centralities. The right plot in Figure 1.10 indicates that the p_{T} spectra of D^0–tagged jets are suppressed for central and mid-central collisions at low p_{T} and the nuclear modification factor R_{CP} was found to increase with the jet p_{T}.

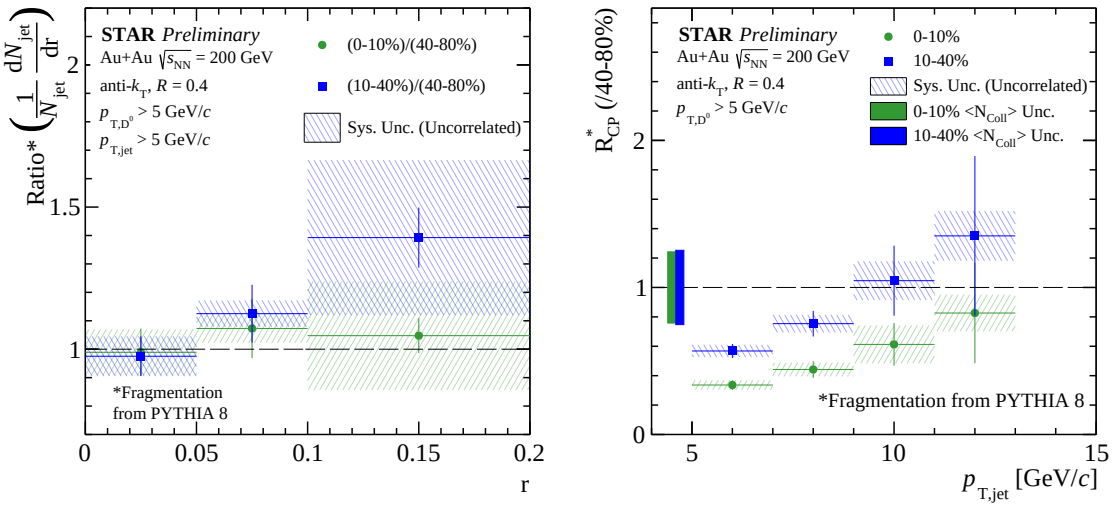

Figure 1.10: Ratio of D^0 radial profiles in central and mid-central Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV with respect to D^0 radial profile in peripheral events as a function of the distance from the jet axis (r) in different centrality bins (left). Nuclear modification factor R_{CP} for D^0 jets as a function of jet p_{T} (right). Taken from [9].

Chapter 2

Jet algorithms

As mentioned earlier, jets can be thought of as collimated showers of high energy particles originating in parton scattering. We can distinguish three different levels of a jet: parton-level jet (described by perturbative QCD in theoretical physics), particle jet (a collection of hadrons) and calorimeter jet, which is registered by the detector and represented by the measured kinematics of the outgoing particles in the experiment.

The goal of experimental physicists is to find a link between those types of jets and re-establish the original correlations between scattered particles. In order to reconstruct jets, we introduce jet algorithms.

2.1 General characteristics of jet algorithms

A jet algorithm represents a set of rules for grouping particles into jets. The ideal jet algorithm should be strictly defined and it should satisfy certain properties based on both experimental and theoretical requirements for jet reconstruction.

Some of the desired conditions that an ideal jet algorithm should meet are for example:

- maximal reconstruction efficiency (identification of physically interesting jets)
- computational efficiency (effective use of computational resources)
- independence of the properties of detector (detector segmentation, resolution)
- order independence (the same jets are found at parton, particle, and detector level)
- theoretically correct behavior (the algorithm should be infrared and collinear safe)

Infrared and collinear safety

Let us briefly illustrate the meaning of the two above-mentioned theoretical attributes of the ideal jet algorithm – infrared and collinear safety.

Infrared safety in jet algorithms suggests that the jet finding should not be affected by adding soft emissions to the event. An example of the violation of infrared safety is depicted in Fig. 2.1, where the jets are represented by cones with arrows that are proportional to their energy and direction. We can see that the number and shape of reconstructed jets have changed only due to the presence of soft gluon radiation.

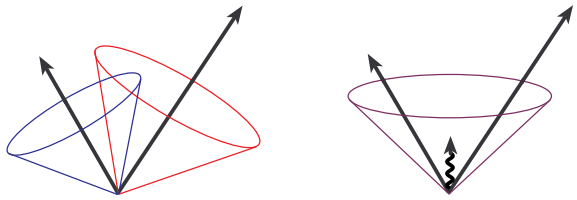

Figure 2.1: An illustration of infrared sensitivity in jet algorithm. Here the presence of soft gluon radiation (shown in figure on the right) results in merging of the two original jets (left figure). Taken from [10].

Collinear safety means that the jet finding is insensitive to collinear radiation. One possible collinear problem can occur in a case when the energy of the measured particle is distributed among several detector towers. The particle is then interpreted by the algorithm as two collinear particles with smaller energies and the jet may not be succesfully reconstructed as demonstrated in Fig. 2.2.

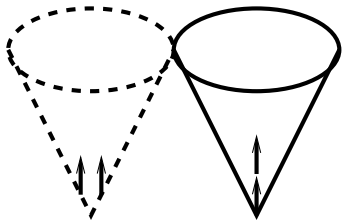

Figure 2.2: Collinear sensitivity in jet reconstruction. The configuration on the left fails to produce a jet because the energy of its seed particle (around which the jet is reconstructed) is distributed among several detector towers. However, the configuration on the right produces a jet due to the narrower distribution of energy in detector. Taken from [10].

Another example of a collinear problem is depicted in Fig. 2.3. This problem is caused by the sensitivity of the algorithm to the particle energy ordering. We can see that different jets are reconstructed depending on the presence or absence of collinear splitting of the most energetic particle in the jet.

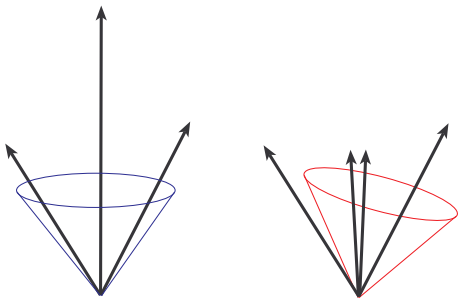

Figure 2.3: Another example of collinear sensitivity in jet reconstruction. A different jet is reconstructed due to the presence of collinear splitting. From [10].

Both infrared and collinear sensitivity are unwanted features in jet algorithms as they can lead to unpredictable behavior – soft emissions as well as collinear splittings can occur randomly and their properties are difficult to predict. This can lead to discrepancies between the interpretations of partonic and observed jets.

2.2 Classes of jet reconstruction algorithms

There are two main categories of jet algorithms – cone and sequential recombination algorithms. In the following text, we are going to introduce their basic properties and provide some examples of the most popular jet reconstruction algorithms.

2.2.1 Cone algorithms

The class of cone algorithms is based on interpretation of a jet as a cone with a radius R defined as:

$$\Delta R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2} < R, \tag{2.1}$$

where i is a so-called seed particle and j depicts a particle that lies within a circle of radius R, η_i is pseudorapidity and ϕ_i azimuthal angle.

Jet cone algorithms usually start by finding seed particles with p_{T} larger than some predefined fixed value. The momentum of the seed particle i sets the initial direction of the jet cone. Then by summing the momenta of all the particles j in the circle with radius R (according to Eq. 2.1), one gets a new direction of jet axis. By repeating this process, a stable cone is found as a representation of the jet.

Although the cone algorithms have a relatively straightforward implementation and an intuitive geometrical meaning (a cone with radius R in the $\eta - \phi$ space of the detector), they are generally not favored in jet reconstruction due to their infrared and collinear unsafety. There have been various attempts to solve this issue in the past. One notable solution is the SISCone seedless algorithm [23], which is an example of a cone algorithm that does not use the concept of seed particle and is infrared and collinear safe.

2.2.2 Sequential recombination algorithms

Nowadays, the most popular approach to jet reconstruction is to use the family of the generalised-k$_t$ algorithms [24], a part of the class of sequential recombination algorithms. They are widely used thanks to their infrared and collinear safety and computing speed. Sequential recombination algorithms are based on the measurement of distances between particles. The basic schematic process of the algorithms can be described as follows:

1. Distances d_{ij} between particles i, j and a distance d_{iB} between entity and beam are calculated as follows:

$$d_{ij} = \min(p_{Ti}^{2k}, p_{Tj}^{2k}) \frac{\Delta_{ij}^2}{R^2} \,, \tag{2.2a}$$

$$d_{iB} = p_{Ti}^{2k} \,, \tag{2.2b}$$

   where $\Delta_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$. Here p_{Ti}, η_i and ϕ_i represent transverse momentum, pseudorapidity and azimuthal angle, respectively. R and k are parameters.

2. Minimum distance is found as $d_{min} = \min \{d_{ij}, d_{iB}\}$.

3. If $d_{min} = d_{ij}$, the particles are recombined into a new object and the steps are repeated from step 1. If $d_{min} = d_{iB}$, object i is called a jet.

The parameter R in equation (2.2a) is called jet resolution parameter and represents the "radius" of a jet. The purpose of parameter k is to include the information about particle energy in distance measurement. The algorithm with $k = 1$ is called k_t algorithm. Value $k = 0$ represents the Cambridge/Aachen algorithm and for $k = -1$, we obtain the anti-k_t algorithm [11].

Cambridge/Aachen algorithm

Due to the value of parameter $k = 0$ in Eq. (2.2), Cambridge/Aachen algorithm is energy-independent in clustering and only the configuration of particles in space is considered. Cambridge/Aachen algorithm can be useful for example to study the jet substructure as it keeps track of hierarchy in angles.

k_t algorithm

The recombination process of k_t algorithm starts by clustering of soft particles (particles with low p_{T}). This can lead to a formation of larger irregular shapes because such a recombination of soft particles at the beginning tends to significantly change the jet direction. Thanks to the sensitivity to soft particles, k_t algorithm may be used in heavy-ion physics for the estimation of background or in the studies of jet substructure.

Anti-k_t algorithm

This algorithm initiates recombination with hard particles (particles with large p_{T}), so the jet direction remains relatively stable and the anti-k_t algorithm leads to creation of jets with circular shapes. This approach is less likely to illustrate the realistic process of hadronization, but it is resilient with respect to soft radiation.

Figure 2.4 shows the behavior of four different jet reconstruction algorithms in a simulated parton-level event from Herwig generator [25] that contains a large number of random soft particles ($\sim 10^4$). As we already mentioned, k_t and Cambridge/Aachen algorithms tend to produce irregular shapes whereas anti-k_t algorithm clusters around the most energetic particles and creates more regular circles.

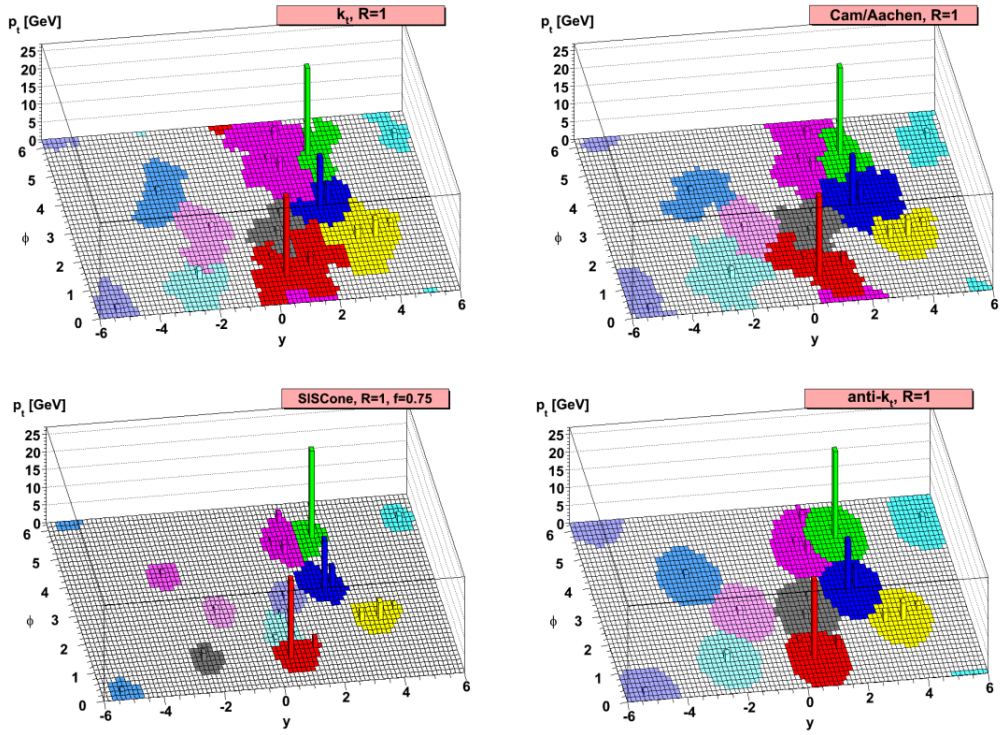

Figure 2.4: A comparison of jet areas using four different reconstruction algorithms (k_t, Cambridge/Aachen, SISCone and anti-k_t algorithm). Taken from [11].

Chapter 3

Machine learning

Machine learning (ML) is a branch of computer science that deals with algorithms, which use large datasets to learn statistical relationships within the data and use those relationships to generate predictions for new data.

Since machine learning has recently undergone a massive development, its use has spread to many fields such as engineering, computer vision and science. Particle physics is no exception, as physicists usually have to deal with large amounts of data and computationally demanding procedures in their work. Applying ML techniques to some tasks can lead to increased performance compared to the traditional algorithms as well as increased computational efficiency.

The purpose of this chapter is to introduce the basic concepts of machine learning techniques that are relevant to the analysis part of this thesis. Therefore, we would like to discuss mainly supervised learning with an emphasis on the jet classification problem and the JetVLAD model architecture.

3.1 Supervised machine learning

Supervised learning approach is designed to learn to predict targets (labels) given inputs. In other words, it is trained to learn functional mapping from input space X to labeled space Y:

$$f\colon X \to Y. \tag{3.1}$$

Here, in our case, X stands for a set of jets with chosen mathematical representation and Y is a set of corresponding jet flavors. Particularly, the dataset used for training can be described as a collection of N pairs $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is a jet represented in a matrix form and $y_i \in \{0, 1\}$ is a binary label, where 0 is a background jet and 1 corresponds to a jet of interest. The exact meaning of label values is dependent on a given task.

The mapping f can be described by a parametric function $f \equiv f(\boldsymbol{x}, \boldsymbol{\theta})$. The parameters $\boldsymbol{\theta}$ can be estimated from the training data in a process called model training. The model is taught to predict a jet flavor $\hat{y} \equiv f(\boldsymbol{x}, \boldsymbol{\theta})$ given an input jet $\boldsymbol{x}$, which is then compared to the ground truth jet flavor y via the loss function. The concept of loss function is used to find the best parametrization of the model [26].

The loss function L measures the model error between the predicted value $\hat{y}$ and ground truth value y during the training phase and is used to guide the model training towards the optimal model parametrization. More formally, we can write the loss function as:

$$L \equiv L(f(\boldsymbol{x}, \boldsymbol{\theta}), y). \tag{3.2}$$

During the training, we randomly sample N data points from the training dataset and calculate the average loss function as:

$$\hat{L} = \frac{1}{N} \sum_{i=1}^{N} L(f(\boldsymbol{x_i}, \boldsymbol{\theta}), y_i). \tag{3.3}$$

The $\hat{L}$ is used as a proxy for full loss function calculated over the entire training dataset, which is computationally infeasible. The optimal parameters $\boldsymbol{\theta}$ are then found by repeatedly computing and minimizing $\hat{L}$ using the tools of mathematical optimization.

One of the basic optimization algorithms is called gradient descent. The algorithm iteratively estimates optimal model parameters by computing the update in the direction of negative gradient of loss function with respect to the parameter $\boldsymbol{\theta}$. The basic version of gradient descent uses a derivative of loss function averaged over the entire dataset. A commonly used modification of this method is stochastic gradient descent (SGD) [27], which reduces computational cost at each iteration by using average loss function $\hat{L}$ defined in Eq. (3.3).

For the purpose of model training, the data can be divided into three groups: training, testing and validation. The model parameters are optimized on the training dataset. Validation dataset is used in order to tune model architecture and guide model training. The final model performance is provided by the testing dataset, when the model is confronted with previously unseen data to prevent overfitting of the model.

3.2 JetVLAD model architecture

The machine learning model used in this work, the JetVLAD model, is derived from the NetVLAD architecture [12], which uses the concept of a neural network. Therefore, let us first introduce the basic properties of neural networks and the NetVLAD architecture.

A neural network can basically be described as a combination of N layers, where each layer takes input of the previous one and applies an affine transformation to it. These layers in fact represent algebraic operations that are followed by an activation function to add non-linearity. This allows to express non-linear tendencies in the data. One of the most used activation functions is the rectified linear unit (ReLU) function [28]. Given an element x, it can be represented as:

$$\mathrm{ReLU}(x) = \max(x, 0), \tag{3.4}$$

which means that the ReLU function keeps only positive elements and discards all negative elements.

An example of the application of neural networks is their use in the field of computer vision. A special case of the neural network layer is the NetVLAD [12] adaptive layer. NetVLAD is inspired by the **V**ector of **L**ocally **A**ggregated **D**escriptors (VLAD) representation, which was originally designed for visual recognition problems.

NetVLAD allows to aggregate a set of descriptors (vectors) and returns a fixed-length feature vector that characterizes it. We assume a set of n descriptors, where each one is represented by a d-dimensional vector $\{\boldsymbol{x_i}\}$. If there are k clusters in the input space with trainable parameters $\{\boldsymbol{c_k}\}$, $\{\boldsymbol{w_k}\}$ and $\{b_k\}$, then we get a $d \times k$-dimensional matrix $\boldsymbol{V}$ of the NetVLAD output:

$$\mathbf{V}_{j,k} = \sum_{i=1}^{n} \frac{e^{\mathbf{w}_k^T \mathbf{x}_i + b_k}}{\sum_{k'} e^{\mathbf{w}_{k'}^T \mathbf{x}_i + b_{k'}}} (\mathbf{x}_{i,j} - \mathbf{c}_{k,j}), \tag{3.5}$$

where $\mathbf{x}_{i,j}$ is a j-th element of the i-th descriptor and $\mathbf{c}_{k,j}$ is the j-th element of the k-th cluster center vector. This matrix is then L^2 normalized in columns, reshaped into a vector and again L^2 normalized in order to get the final feature vector.

A useful feature of NetVLAD is that it can work with unordered set of inputs. It is also resistant to noise, as it was designed to recognize landmarks with a variable number of background objects in place localization. Thanks to these properties, the NetVLAD layer became a suitable foundation for our jet classification task, where a variable number of background and signal objects are present within a jet.

In the analysis part of this work, we use the JetVLAD model [13], whose architecture is based on the previously described NetVLAD layer. This allows to use directly measured variables to produce the so-called particle descriptors. The jet is then described as a set of particles:

$$\mathcal{J} = \{(p_{T,i}, \eta_i, \phi_i, \dots)\}_{i=1}^{n}, \tag{3.6}$$

where n corresponds to the total number of jet constituents, where $p_{T,i}$ is the track transverse momentum, η_i is the track pseudorapidity and ϕ_i corresponds to the track azimuthal angle. The jet defined in such a way is then processed by the JetVLAD model which predicts a most probable jet flavor for the given jet (light or heavy-flavor jet in our case). The schematic representation of the JetVLAD model is shown in Figure 3.1.

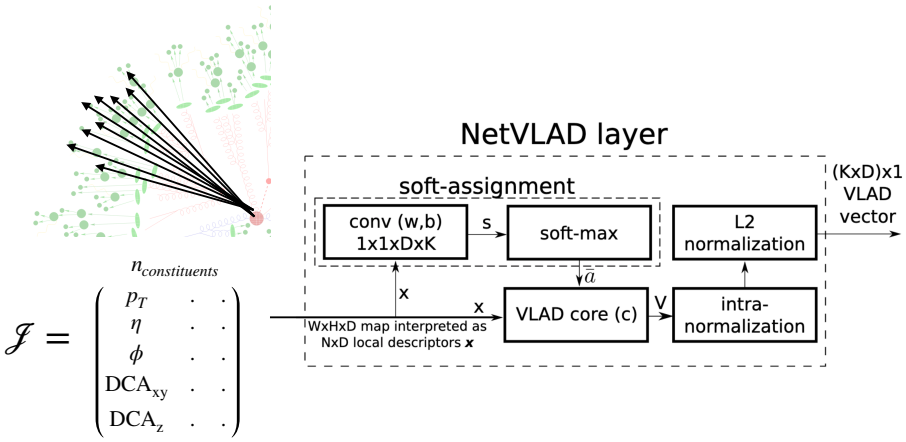

Figure 3.1: JetVLAD model architecture, based on [12].

In previous work, the JetVLAD model was applied to $p + p$ collisions at RHIC energies [13]. However, the applicability of the JetVLAD model to higher collision energies, collision systems, as well as different tagging schemes has not been explored. Hence, it may be an interesting topic to investigate the performance of the model in these regimes.

Chapter 4

Application of JetVLAD model to $p + p$ collisions

In this chapter, we are going to explore the performance of the JetVLAD model as a tool for tagging heavy-flavor jets in proton-proton collisions. The performance of the JetVLAD model has been recently studied on simulated data in $p + p$ collisions at the centre-of-mass energy of $\sqrt{s} = 200$ GeV and reported in [13], which is the starting point for our further studies.

Our goal is to investigate the performance of the JetVLAD model at higher collision energies ($\sqrt{s} = 510$ GeV and $\sqrt{s} = 7$ TeV) and with different tagging approaches (parton tagging and D^0 tagging).

4.1 Datasets and inputs

In our analysis, we used simulated data generated by the PYTHIA8 event generator [29]. Datasets of $p + p$ collisions were generated at the centre-of-mass energy $\sqrt{s} = 200$ GeV and $\sqrt{s} = 510$ GeV, which correspond to the typical RHIC energies. Additionally, datasets for $\sqrt{s} = 7$ TeV were generated to show preliminary results for collision energies achievable at the LHC. In order to well represent the realistic jet flavor ratio of heavy-flavor and light flavor jets, cross-section weighted samples were used. The simulated events were smeared by the fast simulation approach of STAR detector as in [13]. This allows us to include the effects of finite detector resolution without requiring a computationally expensive full detector simulation. The jets were reconstructed from charged particles with $p_{\mathrm{T}} > 0.2$ GeV using the anti-k_{T} clustering algorithm with resolution parameter $R = 0.4$.

The generated datasets were split into three parts for training, testing and validation. As a type of input for training, we used the "Tracking + Vertexing" input from [13] with corresponding input variables ($p_{\mathrm{T}}, \eta, \phi$, DCA_{xy}, DCA_z), since it turned out to be the optimal combination of input features. Here p_{T} is the track transverse momentum, η is the track pseudorapidity, ϕ corresponds to the track azimuthal angle and DCA_{xy}, DCA_{xy} are the distances of the closest approach from the track to the primary vertex in $x - y$ and z planes.

For each jet p_{T} bin, the $\hat{p_{\mathrm{T}}}$ ranges of the Pythia HardQCD processes were set as $\hat{p_{\mathrm{T},min}} = p_{\mathrm{T},min}^{jet} - 2$ and $\hat{p_{\mathrm{T},max}} = p_{\mathrm{T},max}^{jet} + 2$, where $p_{\mathrm{T},min}^{jet}$ is the minimum jet p_{T} in the bin and $p_{\mathrm{T},max}^{jet}$ is the maximum jet p_{T} in the bin. Each jet p_{T} bin dataset for testing and validation contains 5×10^5 generated events with jets, the corresponding training datasets contain 5×10^6 events.

In our analysis, we used two different jet labeling approaches – parton tagging and D^0 tagging. The first approach is conceptually simple, where the jet flavor is identified by the heaviest parton that lies withing the reconstructed jet cone. Here the jets originating from c and b quarks are identified as heavy-flavor jets. However, this method does not reflect real-world experimental setup very well, since partons are experimentally unavailable. The second mentioned approach, D^0 tagging, is based on a method called ghost association. This approach allows one to generate datasets that are closer to the experimentally observed data, since it is using information about D^0 meson being present within the jet to establish the jet flavor. Usually, this is done by manual reconstruction of a D^0 meson via its decay into daughter particles, kaons and pions, which is computationally challenging. We simplify this process by requiring a massless D^0 to be present in the jet cone, which alleviates the need to reconstruct kaons and pions.

4.2 Classification metrics

In order to evaluate the performance of the JetVLAD model, we need to introduce a set of metrics to quantify it.

The first metric that we use is called *Efficiency* or true positive rate (TPR). It is defined as a ratio of positively identified heavy-flavor jets (TP) to the total number of heavy-flavor jets in the testing sample (P). This metric tells us the percentage of signal jets that the model extracts from the sample and it can be expressed as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}}. \tag{4.1}$$

Another relevant metric is mis-identification probability or false positive rate (FPR), which is given by the ratio of false-positive samples (FP) identified in the testing sample to the total number of background objects (N) in the testing sample:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}}. \tag{4.2}$$

In jet physics, however, we prefer to use another related metric – the *Background Rejection* (REJ). It measures how much of the true background will be rejected per one false-positive detection as:

$$\mathrm{REJ} = \frac{1}{\mathrm{FPR}}. \tag{4.3}$$

This metric is particularly useful for heavy-flavor jet classification, where the signal is much smaller than the background due to the difference in production cross-section.

The last metric relevant to us is called *Purity*. This metric tells us the extent of contamination of the signal by false-positive objects and is given by the following equation:

$$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{4.4}$$

where TP is a number of true positive objects and FP is a number of false positive objects found in the testing sample.

4.3 JetVLAD performance and analysis results

In the first part of our analysis, we aim to reproduce the approach used in [13] to provide a starting point for comparison with other datasets and methods. Therefore, we selected datasets of generated $p + p$ collisions at $\sqrt{s} = 200$ GeV and used the parton tagging approach for tagging heavy-flavor jets. The results are demonstrated in Figure 4.1. In the left plot, we can see the purity vs efficiency curves and the right plot shows the curves for background rejection. Various jet p_{T} selections $[5-10], [10-15], [15-20], [20-25]$ and $[25-40]$ GeV/c are demonstrated by different colors. The results show us a very good tagging performance, as at 80% efficiency it the model achieves almost 80% purity and large background rejection factor. The effect of varying jet p_T mostly manifests itself only in the background rejection, where the effect behaves as expected – the higher the jet p_T, the greater background rejection.

To quantify and compare the results within different datasets, we choose two working points based on efficiencies of 80% and 50%. As shown in Tab. 4.1, the signal purity is remaining relatively consistent at the given efficiency, while we find a consistent trend of increasing background rejection with increasing jet p_{T}.

Further on, we introduce a modification to the jet tagging approach in the datasets to explore the applicability of the JetVLAD model for tagging D^0 jets. The plots in Figure 4.2 demonstrate the resulting model performance at the same energy as before ($\sqrt{s} = 200$ GeV) with datasets based on D^0 tagging approach instead. Here the curves for the jet p_T range of $20 - 25$ GeV/c from parton tagging approach (dashed blue curve in both plots) are included to demonstrate the difference in the model performance. There is an apparent decrease in the model performance in the case of D^0 tagging approach, as also quantified in Tab. 4.2. However, it is expected that this tagging approach may better reflect the experimental method and it helps to test the robustness and applicability of the JetVLAD model in this direction.

In order to investigate the performance of the JetVLAD model at different collision energies, datasets with the center-of-mass energy $\sqrt{s} = 510$ GeV were generated and the model was trained on these samples. Again, the datasets were produced for both parton and D^0 tagging approaches. The results in Figure 4.3 (parton tagging) and Figure 4.4 (D^0 tagging) show us a similar performance in purity and rejection as in the case of results for $\sqrt{s} = 200$ GeV. The performance of the model is lower but still consistent with our expectation because of the different production cross-section in the samples.

It may be also interesting to test the model performance for much higher energies achievable at the LHC. Figure 4.5 therefore shows preliminary results for $p + p$ collisions at $\sqrt{s} = 7$ TeV using parton tagging approach. Here, we used different jet p_T ranges due to the higher occurence of background objects in low-p_T jets at the LHC collision energies. The results show a good model performance as it achieved $\sim 70\%$ purity at the efficiency of 80%. However, it should also be noted that we used the same configuration as in the case of RHIC energies (STAR detector parameters) and hence the dataset does not fully reflect the LHC environment.

Our study in $p + p$ collisions across available energies from RHIC to LHC demonstrates that JetVLAD model is a powerful tool for identifying heavy-flavor jets that works well at different collision energies as well as heavy-flavor jet tagging approaches. This makes JetVLAD model a promising candidate for heavy-flavor ML based tagging measurements in real experimental environments.

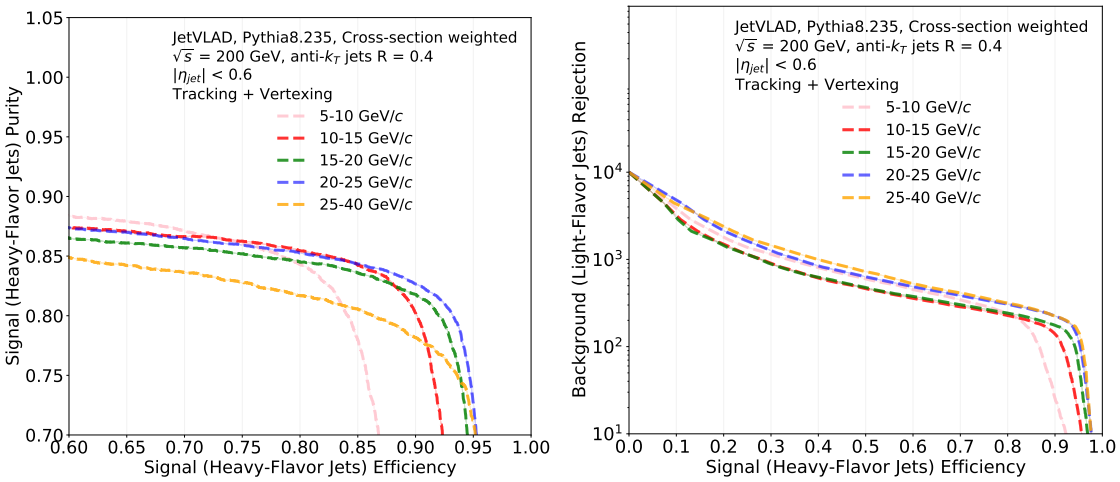

Figure 4.1: Signal purity (left) and background rejection (right) vs efficiency for parton tagging approach at the c.m.s. energy of $\sqrt{s}$ = 200 GeV. Different jet p_T selections are shown separately by the colored curves.

$\sqrt{s}$ [GeV]	Range in jet p_{T} [GeV/c]	Tagging efficiency	Signal Purity	Background Rejection
200 (parton tagging)	[5-10]	80 %	83 %	223
		50 %	88 %	540
	[10-15]	80 %	85 %	223
		50 %	88 %	476
	[15-20]	80 %	85 %	259
		50 %	88 %	506
	[20-25]	80 %	85 %	310
		50 %	88 %	624
	[25-40]	80 %	81 %	322
		50 %	85 %	677

Table 4.1: JetVLAD classification performance in purity and rejection for different jet p_{T} ranges with two working points based on efficiencies of 80% and 50%. Results are shown for datasets generated at $\sqrt{s}$ = 200 GeV using the parton tagging approach.

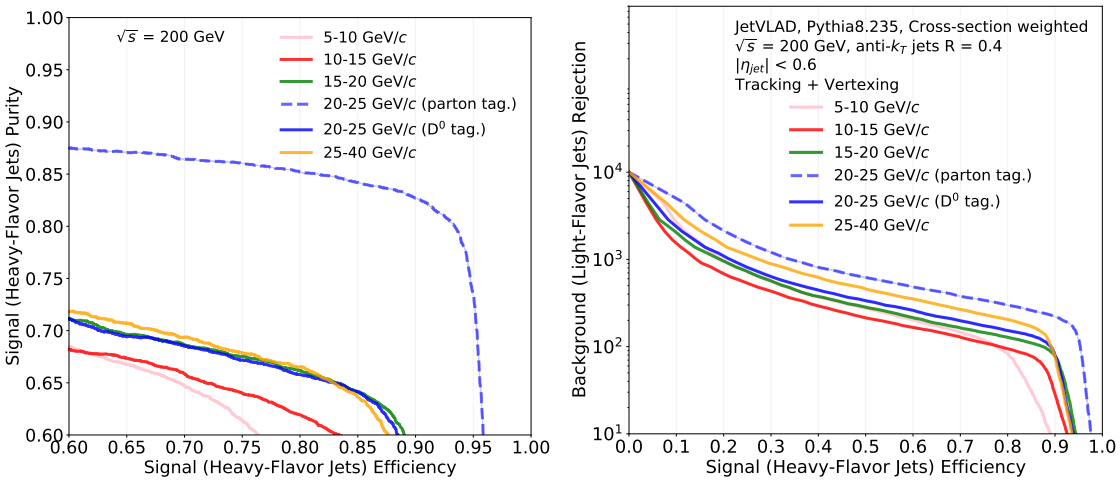

Figure 4.2: Signal purity (left) and background rejection (right) vs efficiency for D^0 tagging approach at $\sqrt{s} = 200$ GeV. The blue dashed curves from parton tagging (jet p_T bin $20 - 25$ GeV) were included for comparison.

$\sqrt{s}$ [GeV]	Range in jet p_T [GeV/c]	Tagging efficiency	Signal Purity	Background Rejection
200 (D^0 tagging)	[5-10]	80 %	54 %	84
		50 %	71 %	278
	[10-15]	80 %	62 %	94
		50 %	70 %	211
	[15-20]	80 %	66 %	127
		50 %	73 %	288
	[20-25]	80 %	66 %	152
		50 %	73 %	336
	[25-40]	80 %	67 %	206
		50 %	74 %	449

Table 4.2: JetVLAD classification performance in purity and rejection for different jet p_T ranges with two working points based on efficiencies of 80% and 50%. Results are shown for datasets generated at $\sqrt{s} = 200$ GeV using the D^0 tagging approach.

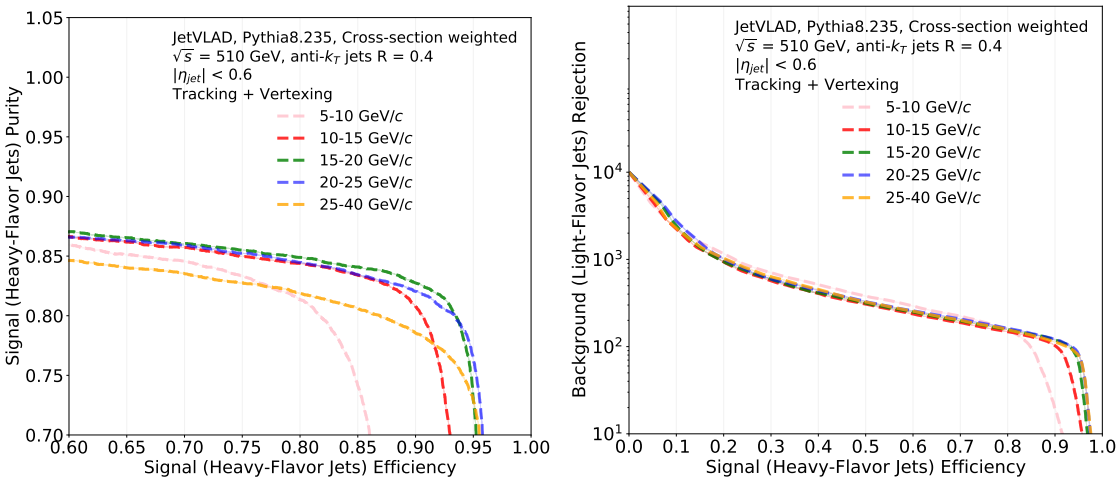

Figure 4.3: Signal purity (left) and background rejection (right) vs efficiency for parton tagging approach at the c.m.s. energy of $\sqrt{s} = 510$ GeV. Different jet p_T selections are shown separately by the colored curves.

$\sqrt{s}$ [GeV]	Range in jet p_{T} [GeV/c]	Tagging efficiency	Signal Purity	Background Rejection
510 (parton tagging)	[5-10]	80 %	81 %	157
		50 %	87 %	385
	[10-15]	80 %	84 %	149
		50 %	87 %	310
	[15-20]	80 %	85 %	161
		50 %	88 %	323
	[20-25]	80 %	84 %	161
		50 %	87 %	323
	[25-40]	80 %	82 %	157
		50 %	85 %	323

Table 4.3: JetVLAD classification performance in purity and rejection for different jet p_{T} ranges with two working points based on efficiencies of 80% and 50%. Results are shown for datasets generated at $\sqrt{s} = 510$ GeV using the parton tagging approach.

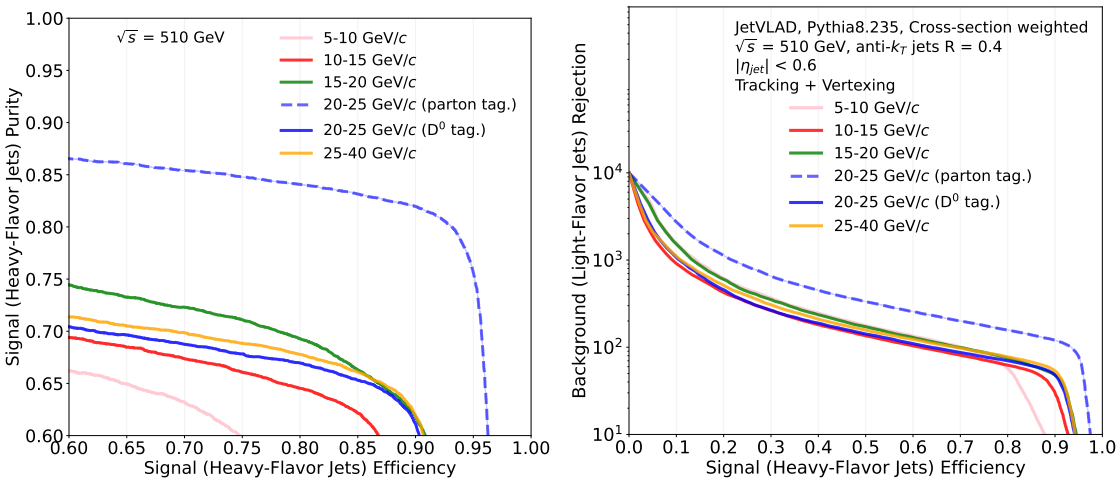

Figure 4.4: Signal purity (left) and background rejection (right) vs efficiency for D^0 tagging approach at $\sqrt{s} = 510$ GeV. The blue dashed curve from parton tagging (jet p_T bin $20 - 25$ GeV) was included for comparison.

$\sqrt{s}$ [GeV]	Range in jet p_T [GeV/c]	Tagging efficiency	Signal Purity	Background Rejection
510 (D^0 tagging)	[5-10]	80 %	53 %	56
		50 %	68 %	179
	[10-15]	80 %	65 %	63
		50 %	71 %	136
	[15-20]	80 %	69 %	75
		50 %	76 %	171
	[20-25]	80 %	67 %	70
		50 %	72 %	141
	[25-40]	80 %	68 %	78
		50 %	73 %	157

Table 4.4: Signal purity (left) and background rejection (right) vs efficiency for D^0 tagging approach at the c.m.s. energy of $\sqrt{s} = 510$ GeV. The blue dashed curves from parton tagging (jet p_T bin $20 - 25$ GeV) were included for comparison.

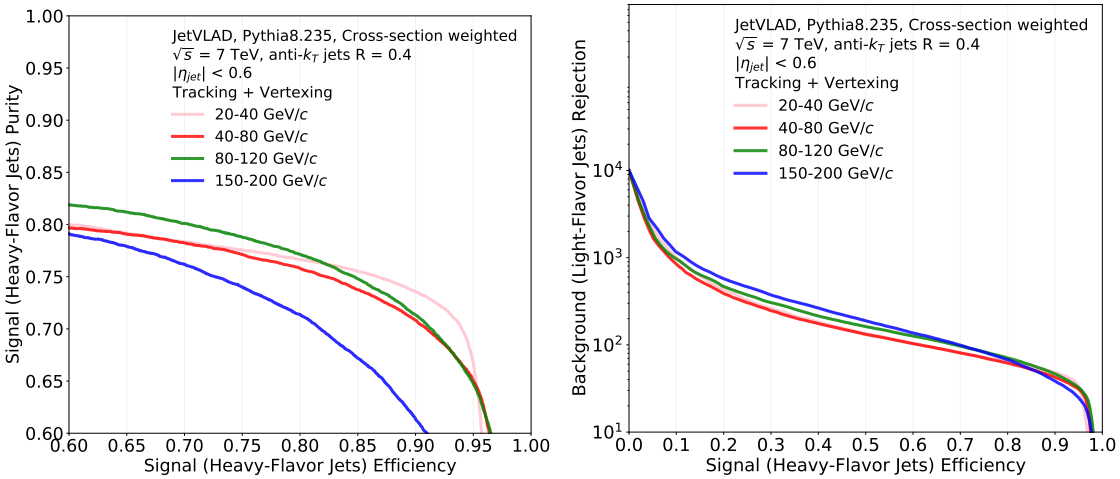

Figure 4.5: Signal purity (left) and background rejection (right) vs efficiency for parton tagging approach at $\sqrt{s}$ = 7 TeV. Different jet p_T selections are shown separately by the colored curves.

$\sqrt{s}$ [TeV]	Range in jet p_{T} [GeV/c]	Tagging efficiency	Signal Purity	Background Rejection
7 (parton tagging)	[20-40]	80 % 50 %	72 % 84 %	90 278
	[40-80]	80 % 50 %	74 % 84 %	94 278
	[80-120]	80 % 50 %	74 % 84 %	95 288
	[150-200]	80 % 50 %	73 % 83 %	100 288

Table 4.5: JetVLAD classification performance in purity and rejection for different jet p_{T} ranges with two working points based on efficiencies of 80% and 50%. Results are shown for datasets generated at $\sqrt{s}$ = 7 TeV using the parton tagging approach.

Chapter 5

Application of JetVLAD model to heavy-ion collisions

This chapter begins with an introduction of the event generator called JETSCAPE [30]. The applicability of the JetVLAD model to the datasets of simulated proton-proton and heavy-ion collisions in JETSCAPE generator is then discussed in section 5.2 of this chapter.

5.1 Introduction to JETSCAPE framework

The **J**et **E**nergy-loss **T**omography with a **S**tatistically and **C**omputationally **A**dvanced **P**rogram **E**nvelope (JETSCAPE) [30] is a comprehensive framework dedicated to the development of Monte Carlo (MC) event generators with an emphasis on the physics of heavy-ion collisions. JETSCAPE also includes powerful statistical tools that allow to conduct Monte Carlo studies of heavy-ion collisions.

JETSCAPE serves as a framework for general-purpose Monte Carlo simulations in heavy-ion collisions. The versatility of this framework allows to simulate the whole evolution of heavy-ion event, which is convenient for studying various aspects of heavy-ion collisions. The JETSCAPE framework consists of a complex system of interacting generators (physics modules) that are directed by the core framework. Thanks to its structure, the JETSCAPE framework allows to produce events and simultaneously analyze and check the observables against experimental predictions. This is also a useful feature for heavy-ion collisions as there are usually multiple stages of collisions with different physics evolving side by side.

As mentioned, JETSCAPE incorporates numerous physics modules (generators) to cover different stages of heavy-ion collisions. Let us mention a few examples of important JETSCAPE modules. The initial state module of JETSCAPE called TRENTO [31] is responsible for the determination of the initial state geometry. The hydrodynamical evolution of the medium in JETSCAPE is governed by hydrodynamic modules such as CLVisc [32] and MUSIC [33]. The main event generator used for hard scattering is PYTHIA8 [29]. As energy loss modules in JESTCAPE, which are also responsible for the simulation of jet quenching effects, we can name for example MATTER [34], MARTINI [35] and LBT [36] etc. For more details on different JETSCAPE modules, we refer the reader to the JETSCAPE manual [30].

JETSCAPE as an event generator allows us to study, for example, the fluid dynamical evolution of the quark-gluon plasma, the transport and medium induced modifications of jets, and other aspects of heavy-ion collisions. Additionally, JETSCAPE framework also contains a powerful statistical toolkit, which includes advanced statistical analysis tools based on Bayesian techniques for calibration and comparison of simulated data with experimental data.

5.2 JETSCAPE analysis

In this section, we would like to test the applicability of the JetVLAD model to the datasets of jets generated by using the JETSCAPE event generator for two configurations – jets in the vacuum ($p + p$ collisions) and jets in the presence of QGP medium (to simulate heavy-ion collisions) at RHIC collision energies.

It has been shown that without secondary vertex information (DCA_{xy} and DCA_{z}), the performance of the JetVLAD model drops down significantly. This was demonstrated in the study of the influence of different input features on the JetVLAD model performance in [13] and it is depicted in the Figure 5.1. In addition, it is expected that the performance of the model will decrease in heavy-ion collisions due to the presence of the QGP medium. The experimentally more realistic D^0 tagging approach also results in the decrease of the model performance (as observed in the analysis described in the previous chapter, section 4.3).

With all that being said, it is therefore very important to provide the highest quality data possible for the model training in order to achieve the best performance of the model. However, it turned out that the current version of the JETSCAPE event generator does not contain the secondary vertex information. It was thus necessary to come up with a temporary solution to our task.

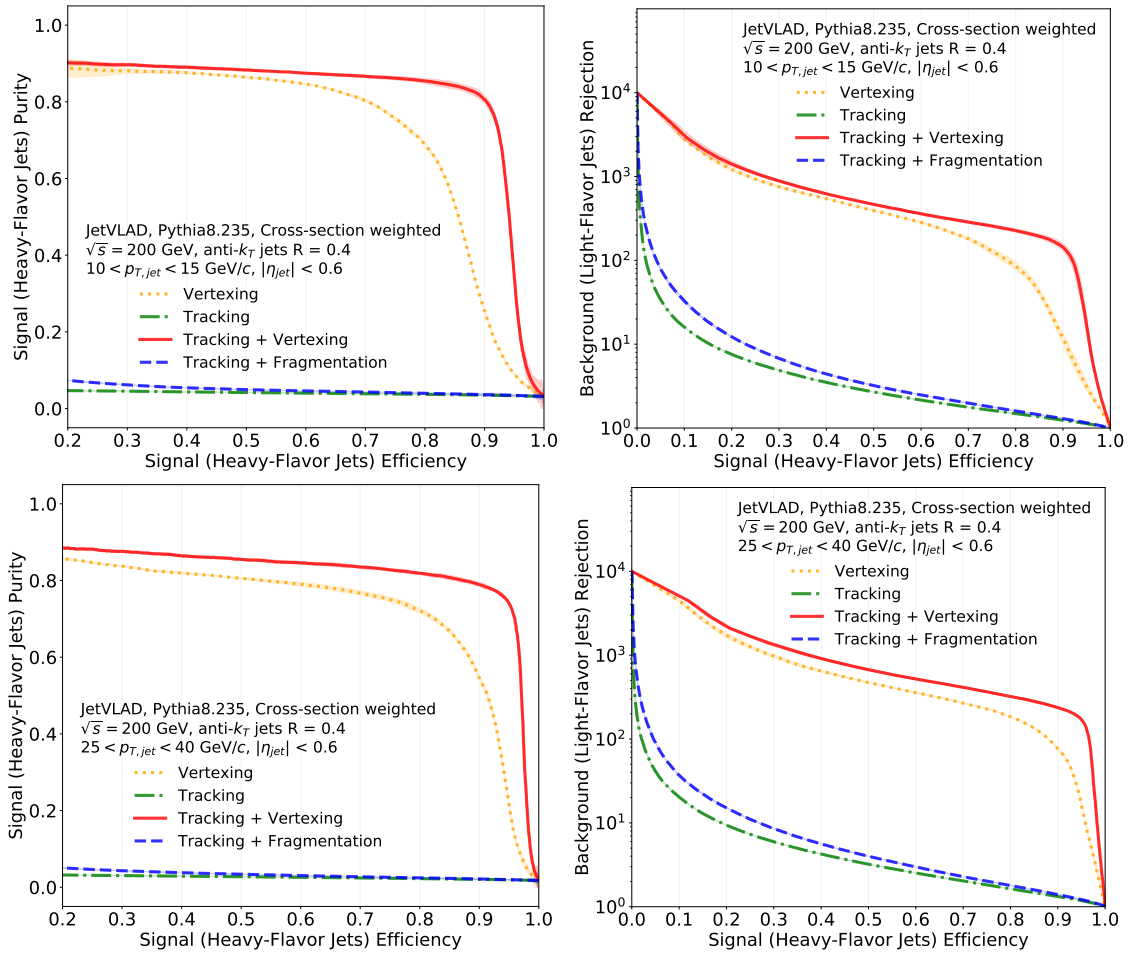

Figure 5.1: Purity and background rejection vs efficiency curves shown for different types of inputs. The top and bottom panels show jets with $10 < p_T < 15$ and $25 < p_T < 40$ GeV/c. The "Tracking+Vertexing" input with the corresponding variables ($p_{\mathrm{T}}, \eta, \phi$, DCA_{xy}, DCA_z) shows up to be the best input option for model training as it reaches the highest performance of all. Taken from [13].

The task was solved in a multi-step approach. A simplified diagram of the data generation process is depicted in Figure 5.2. The first step was to generate the events by JETSCAPE generator with non-decaying D^0 meson setup. In the second step, the D^0 meson decay was simulated using EvtGen [37] generator and each D^0 meson was decayed and replaced in the original dataset by its decay daughters (kaons and pions). The last step was the reconstruction of jets based on the presence of decay daughters in the jet cone. The final data were used to train and test the JetVLAD model performance on the datasets. Let us now break this pipeline down into individual steps.

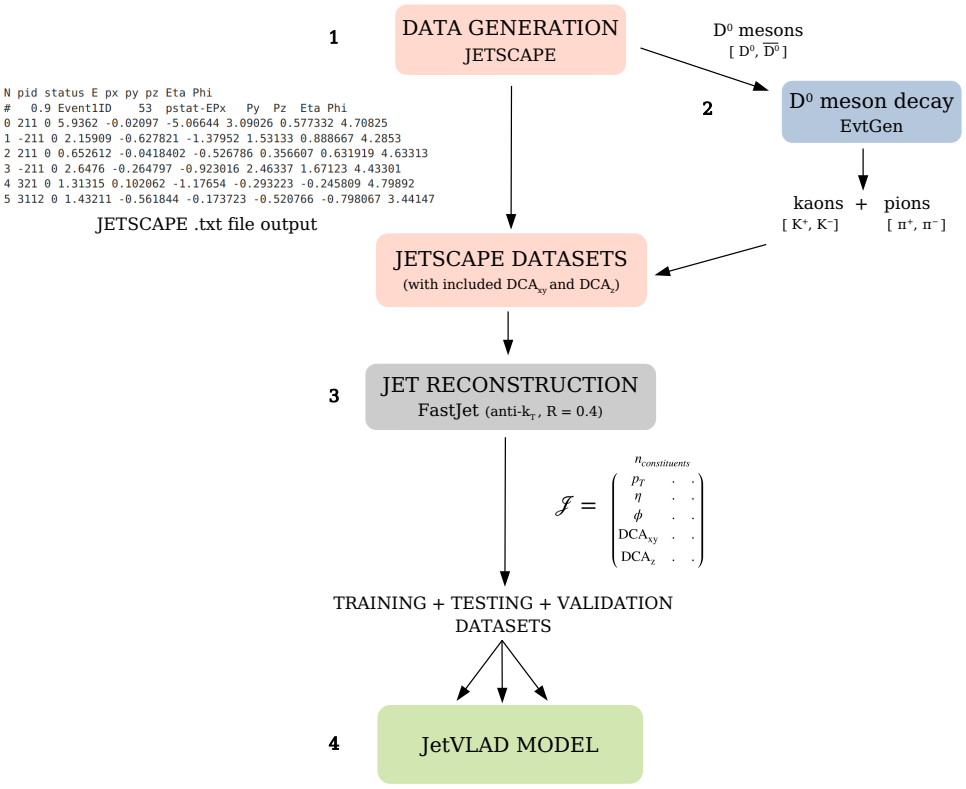

Figure 5.2: Simplified diagram of the analysis process with individual steps from JETSCAPE to the JetVLAD model.

JETSCAPE data generation. For the purpose of this task, the two different datasets were chosen – one setup with vacuum jets and the other with jets in the medium. While generating the events, the D^0 mesons were set to be stable, non-decaying particles. For both datasets a total of 2×10^6 events were generated at the c.m.s. energy of $\sqrt{s} = 200$ GeV. The number of generated events was chosen with respect to the higher computational cost of running JETSCAPE simulations, especially in the case of jets in the medium. For more detailed information about JETSCAPE setup, see the attached configuration files in Appendix B.

Evtgen – D^0 decay. EvtGen [37] is an event generator, which is used for simulations of heavy-flavor particle decays such as decays of B and D mesons. In this analysis, the D^0 mesons are decayed via the hadronic channel $D^0 \to K^- + \pi^+$, which has the branching ratio $BR = (3.89 \pm 0.04)\%$ [38]. The complementary decay $\overline{D^0} \to K^+ + \pi^-$ was also included. The DCA_{xy} and DCA_z for D^0 meson daughter particles were obtained from the EvtGen generator. For the other particles, we used artificial values from a Gaussian distribution based on the DCA spectra for light particles generated from PYTHIA events. The p_T probability distribution of D^0 mesons in p+p and heavy-ion collisions can be seen in Fig. 5.3. The p_T distribution of D^0 mesons tends to be more shifted towards low transverse momentum particles in the presence of the medium.

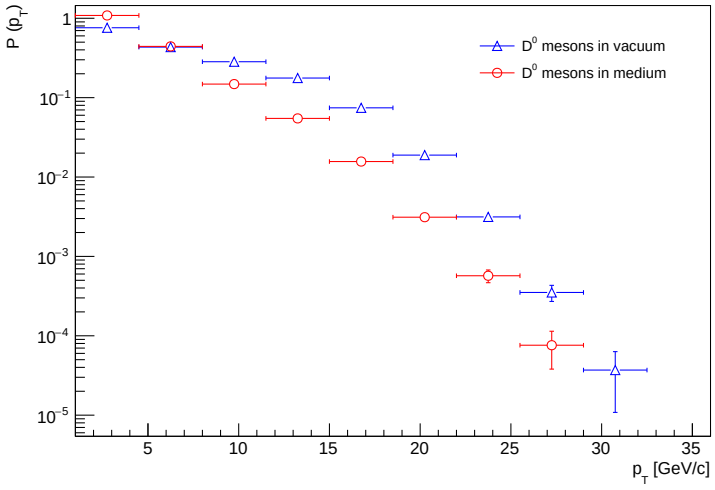

Figure 5.3: D^0 meson p_T probability distribution $P(p_T)$ in vacuum and medium JETSCAPE events. In total, about 10^5 of D^0 and $\overline{D^0}$ mesons were obtained for each of the JETSCAPE dataset.

Jet reconstruction. The jets were reconstructed from charged particles in the events using anti-k_T clustering algorithm with resolution parameter $R = 0.4$, and with the p_T cut on jet constituent particles $p_T > 0.2$ GeV. In order to increase the number of jets retrieved from JETSCAPE datasets, the jets were reconstructed in the jet p_T range $[5 - 100]$ GeV/c without further jet p_T selections. For each dataset, about 10^6 jets and 2×10^4 D^0–tagged jets were obtained. Figure 5.4 shows a comparison of D^0–tagged jet p_T probability distribution in medium and in vacuum, respectively. Again, there is a noticeable increase in the number of low transverse momentum D^0–tagged jet in the presence of the medium.

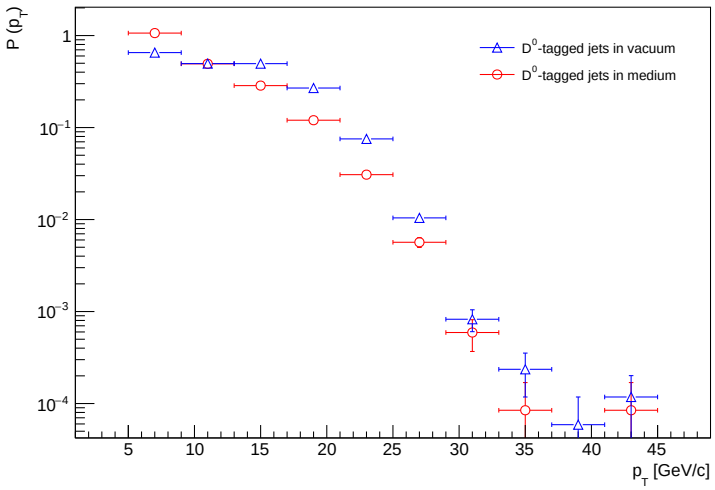

Figure 5.4: Comparison of D^0–tagged jet p_T probability distribution $P(p_T)$ in medium and in vacuum events, reconstructed from the JETSCAPE datasets.

JetVLAD model performance and discussion of results Both datasets were split as 80:10:10 for training, testing and validation. Then the JetVLAD model was trained on these samples using the "Tracking+Vertexing" input with variables ($p_{\mathrm{T}}, \eta, \phi$, DCA_{xy}, DCA_{z}).

Figure 5.5 and 5.6 show the resulting performance of the JetVLAD model in tagging D^0 jets in vacuum and medium JETSCAPE events, respectively. To compare the results within the two datasets, the working points based on efficiencies of 80% and 50% are shown in Tab. 5.1 and Tab. 5.2. The results for vacuum jets and medium jets show up to be comparable with slightly better performance in the case of vacuum events. There is, however, a significant decrease in the model performance in both datasets compared to the previous results for JetVLAD with pure PYTHIA $p+p$ events.

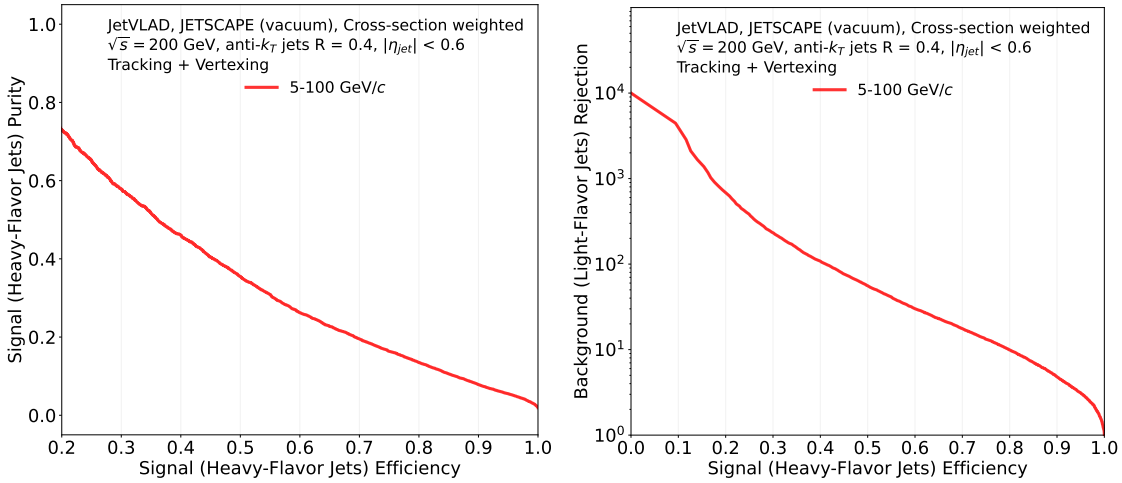

Figure 5.5: Signal purity (left) and background rejection (right) vs efficiency for D^0–tagged jets in the vacuum in JETSCAPE generated events at $\sqrt{s} = 200$ GeV.

$\sqrt{s}$ [GeV]	Range in jet p_{T} [GeV/c]	Tagging efficiency	Signal Purity	Background Rejection
200	[5-100]	80 % 50 %	16 % 46 %	12 86

Table 5.1: JetVLAD classification performance in purity and rejection for working points at efficiencies of 80% and 50%. Jets in vacuum JETSCAPE events.

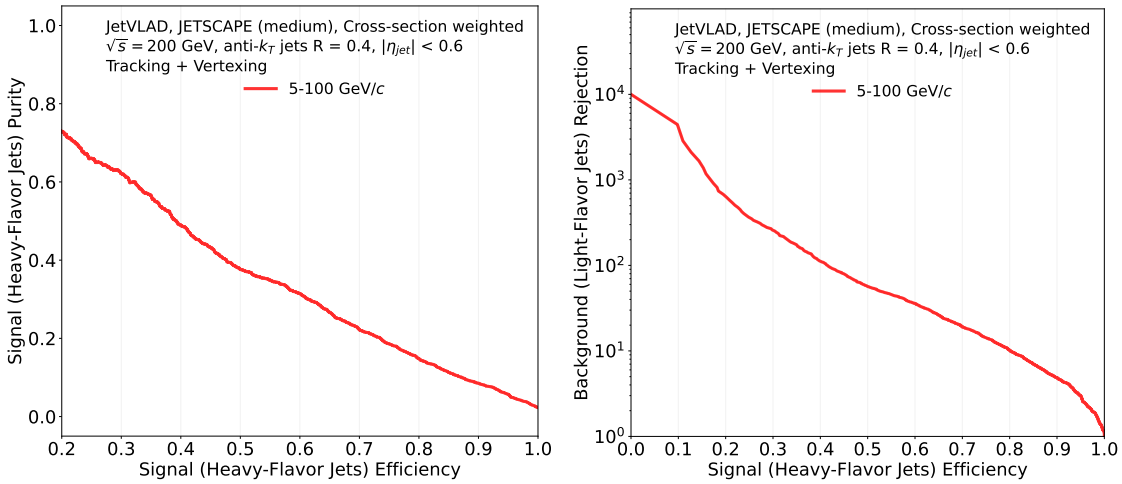

Figure 5.6: Signal purity (left) and background rejection (right) vs efficiency for D^0–tagged jets in the medium in JETSCAPE generated events at $\sqrt{s} = 200$ GeV.

$\sqrt{s}$ [GeV]	Range in jet p_T [GeV/c]	Tagging efficiency	Signal Purity	Background Rejection
200	[5-100]	80 % 50 %	15 % 38 %	10 57

Table 5.2: JetVLAD classification performance in purity and rejection for working points at efficiencies of 80% and 50%. Jets in medium JETSCAPE events.

Since it is known that the DCA input has a large impact on the model performance, we will further investigate, how our choice of secondary vertex sampling affects the final performance of the JetVLAD model. Figure 5.7 shows a comparison of different assumptions about secondary vertex of light particles in JETSCAPE vacuum events. As a baseline, we set PYTHIA8 based sampling from Gaussian distributions with $(\mu_z = 0.07, \sigma_z = 0.6)$ mm and $(\mu_{xy} = 0.14, \sigma_{xy} = 1.2)$ mm that were also used in our analysis. Then three other cases are explored – first two, where we decrease values of σ parameters by factor of 5 and 10 and a third edge case, where we assume that all light particles, which are not coming from the D^0 decay, have secondary vertex location at origin. We observe, that the edge case gets almost 100% efficiency for 100% purity, which may be caused by the fact that JetVLAD model simply count the number of non-zero vertices to obtain the jet tag. The cases of $\sigma/5$ and $\sigma/10$ represent an intermediate performance as compared to the previous one, which is caused by the fact that light-flavor particles are still relatively easy to distinguish. The drop in the performance in our final result is then very likely caused by high level of similarity between vertices of light-flavor and heavy-flavor originating particles as generated by EvtGen.

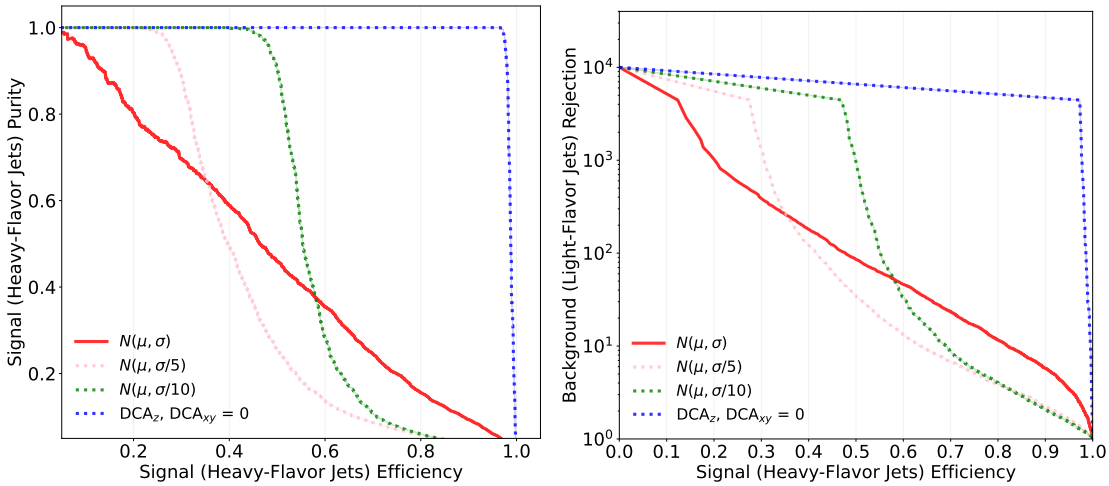

Figure 5.7: Signal purity (left) and background rejection (right) vs efficiency for D^0–tagged jets in the vacuum in JETSCAPE generated events at $\sqrt{s} = 200$ GeV. Each curve represents a different assumption about secondary vertex of light particles that are not coming from the decay of D^0 meson. $N(\mu, \sigma)$ represent dataset trained on secondary vertex distribution fitted to the PYTHIA8 data. $N(\mu, \sigma/5)$ and $N(\mu, \sigma/10)$ represent datasets with narrower spread of the vertex. Lastly, DCA_z, $DCA_{xy} = 0$, represent the edge case, where secondary vertices of light particles are assumed to be zero.

In addition, it is expected that the artificial vertexing is inducing an unrealistic bias into the dataset. Due to the used Gaussian distribution of DCA for light particles, the data is missing the original correlations between the generated particles in the events. As a consequence, each jet is represented as an ensemble of particles with independently sampled Gaussian DCA values with few outliers coming from D^0 decays. Hence, it is difficult for the model to classify such jets due to the low amount of useful signal being present. This is in contrast to result in section 4.3 of the previous chapter, where the JetVLAD model achieves good performance, which shows that unrealistic vertexing may be the main cause of the decrease in performance.

Another way to improve the performance of machine learning algorithms is to increase the amount of data used for training. However, data generation in JETSCAPE is computationally very demanding, especially for the datasets of jets in the medium. Moreover, the whole analysis procedure involved several complex steps due to the necessity to artificially add vertex information to the original JETSCAPE datasets. The low number of jets available for model training may therefore also play a role in the final model performance and it would be beneficial to used larger datasets to explore these effects. At this point, however, the absence of original vertexing information is still expected to have a significant impact on the results, which will remain even for larger amount of data. This is assumed because similar results were obtained for jets in medium and in vacuum, although there are differences in the number of particles in the events, p_{T} distributions of particles etc.

In order to draw a firm conclusion, it would be necessary to apply the JetVLAD model on JETSCAPE datasets with true secondary vertex information, which is not possible at the moment as the JETSCAPE framework does not yet provide it.

Conclusion

In this thesis, we explored how the recently introduced JetVLAD model can be applied towards tagging heavy-flavor jets in $p + p$ and heavy-ion collisions using different definitions of jet tagging approaches.

In the first chapter, the relevant experimental observables and measurements in jet physics were introduced. The second chapter served as an introduction to the topic of jet reconstruction algorithms. The third chapter was dedicated to the basic concepts of machine learning theory with an emphasis on the jet classification problem and the formulation of the JetVLAD model architecture.

In the fourth chapter, the application of the JetVLAD model to PYTHIA generated $p+p$ events for center-of-mass energies $\sqrt{s} = 200$ GeV and $\sqrt{s} = 510$ GeV with different jet tagging approaches (parton and D^0 tagging) was explored. In both cases, the model was found to achieve high classification performance and the experimentally more realistic D^0 tagging approach showed worse performance compared to the parton tagging approach. As last, we performed preliminary studies of the JetVLAD performance in $p+p$ collisions at $\sqrt{s} = 7$ TeV. The preliminary results show promising performance, however, more rigorous simulations of $p + p$ collisions and the detector induced smearing are needed.

In the last chapter, the possibility of applying the JetVLAD model to data from the JETSCAPE event generator was discussed. We explored the importance of the secondary vertex information in the data and presented a temporary solution to the estimation of secondary vertex, which is currently missing in the JETSCAPE framework. Then, the performance of the JetVLAD model was tested for two different datasets – one setup with vacuum jets ($p+p$ collisions) and the other with jets in the medium (to simulate heavy-ion collisions). It has been shown that at this point it is difficult for the JetVLAD model to learn how to distinguish light-flavor jets from the D^0 jets, which may be caused by the simplified secondary vertex simulation and perhaps also due to the lack of available statistics in the data.

The next steps in this research cannot be properly done until the correct secondary vertexing is implemented in the JETSCAPE framework. Once available, the JetVLAD studies from chapter 5 will need to be repeated with the new data and the JetVLAD model may need to be tuned in order to achieve the best possible performance on these novel data.

Appendix A

Glauber model and collision centrality

When colliding heavy nuclei, the resulting particle production depends on how the nuclei encounter each other. In order to describe the relative position of the nuclei, we introduce a quantity called the collision parameter b, which can be defined as the distance between the geometric centre of the two colliding nuclei in the transverse plane. Figure A.1 shows the state before and after the collision of two heavy nuclei. It can be seen that not all nucleons in the nuclei are involved in the collision – one can distinguish the so-called participants and spectators of the collision, whose ratio depends on the collision parameter b.

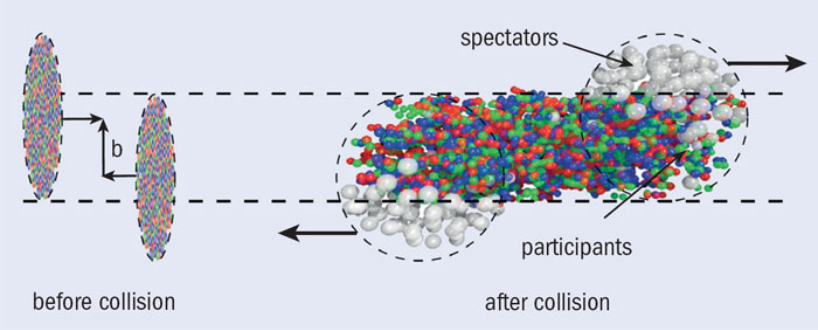

Figure A.1: Collision of two heavy ions with collision parameter b. Taken from [14].

However, neither the collision parameter b nor the number of participants in a collision N_{part} can be directly measured in experiments.

For this reason, we introduce the so-called Glauber model [15], [22], which allows to estimate these geometric quantities by theoretical calculations. There are two main approaches to Glauber model calculations – the optical model and the Monte Carlo Glauber model. The optical Glauber model is derived from the integration of Wood-Saxon distributions. The Monte Carlo Glauber model, unlike the previous one, assigns nucleons to specific positions in a coordinate system and uses Monte-Carlo simulations for the evaluation of the model.

The more theoretical and didactic, optical Glauber model, will be briefly introduced in the following section.

The optical Glauber model describes a nucleus-nucleus collision as a superposition of many independent nucleon collisions (N-N collisions). The assumptions of the model are: sufficiently high nucleon energy (so that the particles to move along direct trajectories); independent motion of the nucleons in the nucleus; and the assumption that the forces between the nucleons are negligible compared to the size of the nucleus.

In the Glauber model, a parameterization of the nucleus density is given by the the Woods-Saxon distribution as follows:

$$\rho(r) = \frac{\rho_0}{1 + \exp\left(\frac{r-R}{c}\right)}, \tag{A.1}$$

where r is the distance from the nucleon center, ρ_0 corresponds to the density at the center of the nucleus, and R is the mean radius of the nucleus. The parameters ρ_0 a c can be determined from scattering experiments by [20].

Let us now consider the collision of nuclei A (target) and B (projectile) with collision parameter $\mathbf{b}$ in Fig. A.2, where A and B are the numbers of nucleons.

We introduce the function $\hat{T}_A(s) = \int \hat{\rho}_A(\mathbf{s}, z_A) dz_A$, where $\mathbf{s}$ is the distance from the center of the projectile A, z is the axis in the beam direction and $\hat{\rho}_A$ is the volume probability of the occurence of a nucleon at the point $(\mathbf{s}, z_A)$. An analogous relation holds for projectile B. The total number of nucleon-nucleon collisions N_{coll} is then defined in Glauber's optical model as:

$$N_{\text{coll}}(\mathbf{b}) = AB \int \hat{T}_A(\mathbf{s})\, \hat{T}_B(\mathbf{s}-\mathbf{b})\ \sigma_{\text{inel}}^{\text{NN}}\, d^2s, \tag{A.2}$$

where z is the axis in the beam direction, $\mathbf{b}$ is the collision parameter and $\sigma_{\text{inel}}^{\text{NN}}$ denotes the inelastic effective cross-section for nucleon-nucleon collisions.

The number of participants in the collision N_{part} is given by:

$$\begin{aligned} N_{\text{part}}(\mathbf{b}) &= A \int \hat{T}_A(\mathbf{s}) \left\{1 - \left[1 - \hat{T}_B(\mathbf{s}-\mathbf{b})\, \sigma_{\text{inel}}^{\text{NN}}\right]^B\right\} d^2s + \\ &\quad B \int \hat{T}_B(\mathbf{s}-\mathbf{b}) \left\{1 - \left[1 - \hat{T}_A(\mathbf{s})\, \sigma_{\text{inel}}^{\text{NN}}\right]^A\right\} d^2s. \end{aligned} \tag{A.3}$$

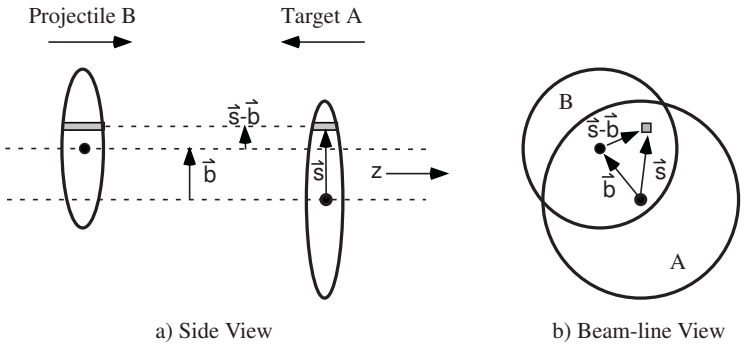

Figure A.2: Collision diagram in Glauber's optical model. Taken from [15].

The collision centrality can be determined from experimentally measurable quantities – for example, the collision multiplicity (number of produced particles). Figure A.3 shows the dependence of the differential cross-section on the number of produced charged particles (N_{ch}). The distribution is split into intervals of centralities such that the number of charged particles per bin corresponds to a certain percentage of geometric overlap in the collision.

The most central collisions correspond to values of centralities $0 - 5\%$ and small values of the parameter b. On the other hand, collisions with the highest values of centralities and large values of the parameter b are called peripheral collisions.

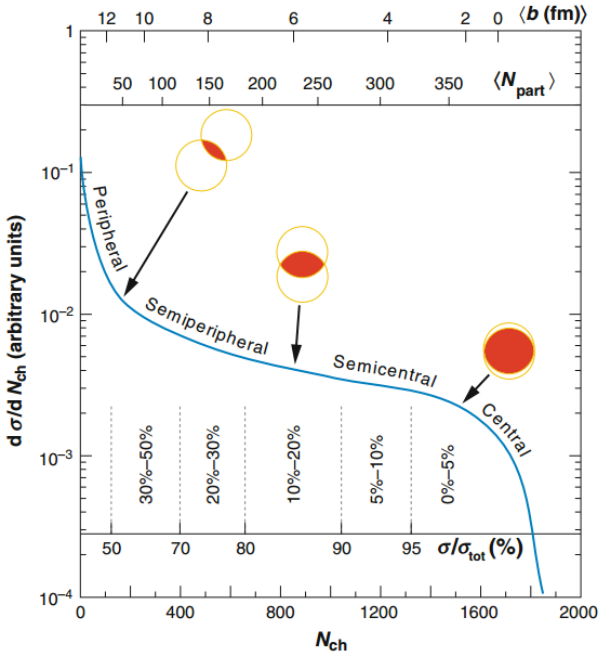

Figure A.3: Differential cross-section as a function of the number of produced charged particles N_{ch} and determination of centrality using calculations from the Glauber model. From [16].

Appendix B

JETSCAPE – configuration files

```
<?xml version="1.0"?>

<jetscape>

    <nEvents> 100000 </nEvents>

    <outputFilename>vacuum_run_1</outputFilename>
    <JetScapeWriterAscii> on </JetScapeWriterAscii>

    <Random>
        <seed>0</seed>
    </Random>


    <!-- Hard Process -->
    <Hard>
        <PythiaGun>
        <pTHatMin>18</pTHatMin>
        <pTHatMax>27</pTHatMax>
        <eCM>200</eCM>
            <LinesToRead>
            HardQCD:gg2gg = off
            HardQCD:gg2qqbar = off
            HardQCD:qg2qg = off
            HardQCD:qq2qq = off
            HardQCD:qqbar2gg = off
            HardQCD:qqbar2qqbarNew = off
            Charmonium:all = on
            HardQCD:gg2ccbar = on
            HardQCD:qqbar2ccbar = on
            Bottomonium:all = off
            HardQCD:gg2bbbar = off
            HardQCD:qqbar2bbbar = off
        </LinesToRead>
        </PythiaGun>
    </Hard>

    <!--Eloss Modules -->
    <Eloss>
        <Matter>
        <Q0> 1.0 </Q0>
        <in_vac> 1 </in_vac>
        <vir_factor> 0.25 </vir_factor>
        <recoil_on> 0 </recoil_on>
        <broadening_on> 0 </broadening_on>
        <brick_med> 0 </brick_med>
        </Matter>
    </Eloss>

    <!-- Jet Hadronization Module -->
    <JetHadronization>
        <name>colorless</name>
    </JetHadronization>

</jetscape>
```

Figure B.1: An example of JETSCAPE configuration file for vacuum events.

```
    <?xml version="1.0"?>

    <jetscape>

        <nEvents> 100000 </nEvents>
        <setReuseHydro>true</setReuseHydro>
        <nReuseHydro> 100000 </nReuseHydro>
        <outputFilename>medium_run_1</outputFilename>
        <JetScapeWriterAscii> on </JetScapeWriterAscii>
        <nEvents_printout>1</nEvents_printout>
        <Random>
            <seed>0</seed>
        </Random>
        <!--  Inital State Module   -->
        <IS>
                <initial_profile_path>../../SummerSchool2021/Jul23_Jets/test_hydro_profile</-
initial_profile_path>
        </IS>
        <!-- Hard Process -->
        <Hard>
            <PythiaGun>
            <pTHatMin>18</pTHatMin>
            <pTHatMax>27</pTHatMax>
            <eCM>200</eCM>
                <LinesToRead>
                HardQCD:gg2gg = off
                HardQCD:gg2qqbar = off
                HardQCD:qg2qg = off
                HardQCD:qq2qq = off
                HardQCD:qqbar2gg = off
                HardQCD:qqbar2qqbarNew = off
                Charmonium:all = on
                HardQCD:gg2ccbar = on
                HardQCD:qqbar2ccbar = on
                Bottomonium:all = off
                HardQCD:gg2bbbar = off
                HardQCD:qqbar2bbbar = off
            </LinesToRead>
            </PythiaGun>
        </Hard>
        <!-- Preequilibrium Dynamics Module  -->
        <Preequilibrium>
                <NullPreDynamics/>
        </Preequilibrium>
        <!--  Hydro  Module   -->
        <Hydro>
                <hydro_from_file>
                        <boost_invariant_>1</boost_invariant_>
                        <read_in_multiple_hydro>1</read_in_multiple_hydro>
                        <hydro_files_folder>../../SummerSchool2021/Jul23_Jets/-
test_hydro_profile</hydro_files_folder>
                </hydro_from_file>
        </Hydro>
        <!-- Eloss Modules  -->
        <Eloss>
                <deltaT>0.1</deltaT>
                <formTime>-0.1</formTime>
                <maxT>250</maxT>
                <tStart>0.6</tStart>
        <!--  Start time of jet quenching, proper time, fm/c    -->
        <mutex>ON</mutex>
```

Figure B.2: An example of JETSCAPE configuration file (part 1/2) for medium events. Based on the configuration files from the official JETSCAPE github [17].

```
    <Matter>
            <name>Matter</name>
            <useHybridHad>0</useHybridHad>
            <matter_on>1</matter_on>
            <Q0>2.0</Q0>
            <vir_factor>0.25</vir_factor>
            <in_vac>0</in_vac>
            <recoil_on>1</recoil_on>
            <broadening_on>1</broadening_on>
            <brick_med>0</brick_med>
            <!--  Set brick_med to 1 while using Brick Hydro module -->
            <T0>0.16</T0>
            <hydro_Tc>0.16</hydro_Tc>
            <qhat0>-1.0</qhat0>
            <!--  If Type=0, 1, 5,6,7 set qhat0 as negative since alphas will be used -->
            <alphas>0.25</alphas>
    </Matter>
    <Lbt>
            <name>Lbt</name>
            <Q0>2.0</Q0>
            <in_vac>0</in_vac>
            <only_leading>0</only_leading>
            <hydro_Tc>0.16</hydro_Tc>
            <alphas>0.25</alphas>
    </Lbt>
    </Eloss>
    <!--  Jet Hadronization Module  -->
            <JetHadronization>
            <name>colorless</name>
    <!--
<take_recoil>1</take_recoil>
<eCMforHadronization>2510</eCMforHadronization>
 -->
    </JetHadronization>
    </jetscape>
```

Figure B.3: An example of JETSCAPE configuration file (part 2/2) for medium events. Based on the configuration files from the official JETSCAPE github [17].

Bibliography

[1] M. Aaboud *et al.*, "Measurement of the inclusive jet cross-sections in proton-proton collisions at $\sqrt{s}$ = 8 TeV with the ATLAS detector," *JHEP*, vol. 09, p. 020, 2017.

[2] S. Acharya *et al.*, "Direct observation of the dead-cone effect in quantum chromodynamics," *Nature*, vol. 605, no. 7910, pp. 440–446, 2022. [Erratum: Nature 607, E22 (2022)].

[3] ATLAS Experiment, "Looking inside trillion degree matter with ATLAS at the LHC, web." `https://atlas.cern/Updates/Feature/Heavy-Ion-Physics` [Online; 1.12.2022].

[4] M. Aaboud *et al.*, "Measurement of the nuclear modification factor for inclusive jets in Pb+Pb collisions at $\sqrt{s_{\mathrm{NN}}}$ = 5.02 TeV with the ATLAS detector," *Phys. Lett. B*, vol. 790, pp. 108–128, 2019.

[5] A. M. Sirunyan *et al.*, "First measurement of large area jet transverse momentum spectra in heavy-ion collisions," *JHEP*, vol. 05, p. 284, 2021.

[6] "Measurement of the radius dependence of charged-particle jet suppression in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}}$ = 5.02 TeV," *arXiv:2303.00592*, 3 2023.

[7] J. Adam *et al.*, "Measurement of inclusive charged-particle jet production in Au + Au collisions at $\sqrt{s_{NN}}$ =200 GeV," *Phys. Rev. C*, vol. 102, no. 5, p. 054913, 2020.

[8] A. M. Sirunyan *et al.*, "Studies of charm quark diffusion inside jets using PbPb and pp collisions at $\sqrt{s_{\mathrm{NN}}}$ = 5.02 TeV," *Phys. Rev. Lett.*, vol. 125, no. 10, p. 102001, 2020.

[9] D. Roy, "An Investigation of Charm Quark Jet Spectrum and Shape Modifications in Au+Au Collisions at $\sqrt{s_{NN}}$ = 200 GeV," *Acta Phys. Polon. Supp.*, vol. 16, no. 1, p. 111, 2023.

[10] G. C. Blazey *et al.*, "Run II jet physics," in *Physics at Run II: QCD and Weak Boson Physics Workshop: Final General Meeting*, pp. 47–77, 5 2000.

[11] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm," *JHEP*, vol. 04, p. 063, 2008.

[12] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[13] J. Bielčíková, R. Kunnawalkam Elayavalli, G. Ponimatkin, J. H. Putschke, and J. Sivic, "Identifying Heavy-Flavor Jets Using Vectors of Locally Aggregated Descriptors," *JINST*, vol. 16, no. 03, p. P03017, 2021.

[14] Toia, A., University of Padua/INFN, "Participants and spectators at the heavy-ion fireball, web." `https://cerncourier.com/a/participants-and-spectators-at-the-heavy-ion-fireball` [Online; 1.4.2023].

[15] M. L. Miller, K. Reygers, S. J. Sanders, and P. Steinberg, "Glauber modeling in high energy nuclear collisions," *Ann. Rev. Nucl. Part. Sci.*, vol. 57, pp. 205–243, 2007.

[16] N. Science and Technology, "STAR Experiment, Event Triggering, web." `https://nsww.org/projects/bnl/star/trigger-system-e.php` [Online; 1.4.2023].

[17] Yasuki Tachibana, "JETSCAPE github, Jet Session from JETSCAPE Summer School 2021." `https://github.com/JETSCAPE/SummerSchool2021/tree/master/Jul23_Jets/config` [Online; 1.5.2023].

[18] Y. L. Dokshitzer and D. E. Kharzeev, "Heavy quark colorimetry of QCD matter," *Phys. Lett. B*, vol. 519, pp. 199–206, 2001.

[19] L. Cunqueiro and M. Płoskoń, "Searching for the dead cone effects with iterative declustering of heavy-flavor jets," *Phys. Rev. D*, vol. 99, no. 7, p. 074027, 2019.

[20] S. Sarkar, H. Satz, and B. Sinha, eds., *The Physics of the Quark-Gluon Plasma*, vol. 785. 2010.

[21] L. Adamczyk *et al.*, "Beam Energy Dependence of Jet-Quenching Effects in Au+Au Collisions at $\sqrt{s_{\rm NN}} = 7.7$, 11.5, 14.5, 19.6, 27, 39, and 62.4 GeV," *Phys. Rev. Lett.*, vol. 121, no. 3, p. 032301, 2018.

[22] P. Shukla, "Glauber model for heavy ion collisions from low-energies to high-energies," *arXiv:nucl-th/0112039*, 12 2001.

[23] G. P. Salam and G. Soyez, "A Practical Seedless Infrared-Safe Cone jet algorithm," *JHEP*, vol. 05, p. 086, 2007.

[24] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual," *Eur. Phys. J. C*, vol. 72, p. 1896, 2012.

[25] G. Corcella, I. Knowles, G. Marchesini, S. Moretti, K. Odagiri, P. Richardson, M. Seymour, and B. Webber, "Herwig 6.5 release note.," 02 2010.

[26] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," *arXiv preprint arXiv:2106.11342*, 2021.

[27] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization Methods for Large-Scale Machine Learning," *arXiv e-prints*, p. arXiv:1606.04838, June 2016.

[28] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," *arXiv e-prints*, p. arXiv:1803.08375, Mar. 2018.

[29] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, "An introduction to PYTHIA 8.2," *Comput. Phys. Commun.*, vol. 191, pp. 159–177, 2015.

[30] J. H. Putschke *et al.*, "The JETSCAPE framework," *arXiv:1903.07706*, 3 2019.

[31] J. S. Moreland, J. E. Bernhard, and S. A. Bass, "Alternative ansatz to wounded nucleon and binary collision scaling in high-energy nuclear collisions," *Phys. Rev. C*, vol. 92, no. 1, p. 011901, 2015.

[32] L.-G. Pang, H. Petersen, and X.-N. Wang, "Pseudorapidity distribution and decorrelation of anisotropic flow within the open-computing-language implementation CLVisc hydrodynamics," *Phys. Rev. C*, vol. 97, no. 6, p. 064918, 2018.

[33] B. Schenke, S. Jeon, and C. Gale, "(3+1)d hydrodynamic simulation of relativistic heavy-ion collisions," *Phys. Rev. C*, vol. 82, p. 014903, Jul 2010.

[34] A. Majumder, "Incorporating Space-Time Within Medium-Modified Jet Event Generators," *Phys. Rev. C*, vol. 88, p. 014909, 2013.

[35] B. Schenke, C. Gale, and S. Jeon, "MARTINI: An Event generator for relativistic heavy-ion collisions," *Phys. Rev. C*, vol. 80, p. 054913, 2009.

[36] S. Cao, T. Luo, G.-Y. Qin, and X.-N. Wang, "Heavy and light flavor jet quenching at RHIC and LHC energies," *Phys. Lett. B*, vol. 777, pp. 255–259, 2018.

[37] A. Ryd, D. Lange, N. Kuznetsova, S. Versille, M. Rotondo, D. P. Kirkby, F. K. Wuerthwein, and A. Ishikawa, "EvtGen: A Monte Carlo Generator for B-Physics," 5 2005.

[38] PDG, "The Review of Particle Physics (2020)." `https://pdg.lbl.gov/2016/tables/rpp2016-tab-mesons-charm.pdf` [Online; 1.12.2022].

[39] S. Acharya *et al.*, "Measurement of the production of charm jets tagged with D^0 mesons in pp collisions at $\sqrt{s} = 7$ TeV," *JHEP*, vol. 08, p. 133, 2019.

[40] "Search for medium effects using jets from bottom quarks in PbPb collisions at $\sqrt{s_{NN}} = 5.02$ TeV," *arXiv:2210.08547*, 10 2022.

[41] G. P. Salam, "Towards Jetography," *Eur. Phys. J.*, vol. C67, pp. 637–686, 2010.

[42] ML4Jets, "ML4Jets, Identifying Heavy-Flavor Jets Using Vectors of Locally Aggregated Descriptors, presentation." `https://indico.cern.ch/event/980214/contributions/4413492/` [Online; 1.5.2023].

[43] Ponimatkin G., "JetVLAD tagging model, software." https://github.com/ponimatkin/NetVLAD-tagger-pytorch [Online; 1.5.2023].

[44] The JETSCAPE Collaboration, "JETSCAPE github, software." https://github.com/JETSCAPE [Online; 1.3.2023].