Wigner Research Centre for Physics
1121 Budapest, Konkoly-Thege Miklós út, 29-33.
1525 Budapest, Pf. 49
Tel/Fax: +36 1 3922222/2536
E-mail: vertesi.robert@wigner.hu

Review of Master Thesis

Identifying heavy-flavor jets using vectors of locally aggregated desctiptors

by *Georgij Ponimatkin, Bc.*

In his master thesis Georgij presents a method of selecting heavy-flavor jets in a realistic experimental environment using the novel JetVLAD machine learning technique that he developed with his co-authors. In recent years, research of flavor-dependent observables became one of the main focal topics in the high-energy community and, because of the relatively scarce production of heavy-flavor, its precise identification is a crucial question in most measurements. As a result of his work the author is able to identify heavy-flavor jets in PYTHIA proton-proton simulations based on reconstructed jet features, at $\sqrt{s}$=200 GeV with an 80% efficiency and an above 80% purity at the same time, which is a remarkable result. The originality and high quality of the current research is supported by a peer-reviewed paper (JINST 16 (2021) 03, P03017) to which the author was the main contributor, and two conference talks.

The thesis is written in English. Its use of language is generally professional, though there are minor mistakes (mostly missing or misplaced articles). Chapters 1, 2 and 3 provide introduction to high-energy heavy-flavor physics, as well as machine learning techniques in general and applied to this field of research. Chapter 4 details the author's own work: the JetVLAD model and the results obtained from it. Then Chapter 5 briefly summarizes the research and provides outlook toward application on data. The text is generally well structured and easy to follow. In the introductory part, however, I often found that the notations are inconsistent (eg. Δ in the text but θ in the Lund-plane figures, similarly the z and $z$bar variables) and some quantities or concepts are not defined (eg. ω in the case of the dead cone). It also makes understanding more difficult for the everyday physicist that many neural network concepts go undefined or unexplained (residual skip connections, DropOut method, cosine annealing, warm restart, convnet). Also, while the general quality of the thesis is high, there are a few inaccuracies and arguable statements especially in the introductory part. Some examples:

- In 1.2, the author interprets the requirement of process independence as the necessity of parton-level and hadron-level jets having the same properties. I think that may hold only in the case of a clean single-jet environment, otherwise jets can irrecoverably merge during hadronization.

- In Sec. 1.4, while the 1st direct dead cone measurement is indeed by ALICE, the depletion at small angles was indirectly measured in LEP data already, see DELPHI-2004-037 CONF 712.

- The two bottom panels in Fig. 4.1 are the same (ΔR). I believe the author's intention was to show the z-dependence similarly to the JINST paper.

Questions to the author:

1) Machine learning techniques have been used in heavy-flavor jet identification based on several jet descriptors, and they tend to yield stunning raw discrimination powers, eg. in arXiv:1909.01639. The usual caveat is that simulations are imperfect and therefore the performance of these methods under real circumstances is difficult to evaluate. In fact the crucial point in current measurements is not to obtain high purity and efficiency using conventional methods but rather to get the corresponding systematics under control. Discriminators based on the secondary vertex are well known to be robust, while tracking or jet structure features are less used because of model dependence. You are obviously aware of the problem as you mention that in real data analysis, systematics could be evaluated from different models. However, in heavy-ion collisions, jet structures may drastically change. Do you see any chance for the applicability of the technique in p+A or A+A collisions?

2) The high sensitivity to the thermal background means that the underlying event has to be known very well in a real measurement for the proper training. Fig. 4.8, however, is an extreme scenario. Do you have any more realistic way in mind to estimate the effect?

3) The "balanced" sample yielded a very high purity, and you consider it as an indicaton that higher purity can be achieved with a good trigger. Why do you believe that a two-step selection process can be more efficient than the ML trained in an unordered, unbiased way? (In case of the "balanced" sample the higher input purity of the same sort of jet samples result in a higher output purity. On the other hand any trigger will preselect jets based on some kinematic features and thus the samples will already be biased compared to unselected jets.)

4) Table 4.1 the $p_T$hat regions are relatively narrow around the p_T^{jet} ranges: a $p_T\text{hat}^{min}$=13 GeV/c corresponds to p_T^{jet}>15 GeV/c. However, the jet distributions are known to be affected up to $p_T^{jet}\sim 2\ p_T\text{hat}^{min}$ (or even higher in specific event activity classes). So, although this might not affect the internal jet structures too much, your choice does not look very conservative to me. Could you explain what motivated this decision?

In summary, Georgij's thesis deals with a timely topic and presents high-quality original research that is beyond expectations from a master student, despite the comments I worded above (none of which are major critics). Therefore, also assuming that the questions will be addressed during the defense, I propose a grade **"A" (excellent)**.

Budapest, 6th June 2021

Róbert Vértesi, Ph.D.
(reviewer)