Ph.D. Annual Report

Ph.D. Student: Martino Tanasini
Tutors: Andrea Coccaro, Carlo Schiavi

13 settembre 2021

1 Research Activity

My research activity can be divided in two main arguments:

Qualification Task: In order to be listed among the authors of all papers published by the ATLAS Collaboration, every scientist in ATLAS must have completed a Qualification Task (QT). QTs are service tasks useful to the group by which they are assigned, but they are also meant to help the qualifying student familiarize with the instruments pivotal to its future job in the group. My QT takes place in the Flavour Tagging Group in the context of a new release of Athena, the offline reconstruction software developed by the ATLAS collaboration.

Since Athena is a very complex software, in which different modules take on different parts of the reconstruction and analysis of registered events, a full validation of each new change is needed, in order to see how these changes propagate from functionality to functionality.

Flavour Tagging is the process of identifying the flavour of a quark from which an hadronic jet originated. Flavour Tagging in ATLAS is done by means of an algorithm called DL1r, a Deep Neural Network (DNN) that combines the outputs of some so-called "low level" algorithms, that are based on the track and secondary vertex information in the jet, and produce three scalar outputs. These outputs represent the probability of the jet to be a b-jet (a jet containing a b-hadron), a c-jet (a jet containing a c-hadron) or a light-flavour jet (a jet that doesn't contain b-hadrons, c-hadrons or tau leptons). For DL1r to unleash its full discriminating power, each low level algorithm has to be maintained, developed, tuned and validated each time a new release is being implemented. My QT consists in preparing the tuning, for the new release, of the low-level algorithms devoted to the reconstruction of secondary and tertiary vertices, namely SingleSecondaryVertexFinder (SSVF) and JetFitter. In order to do so, I have to:

- Contribute to the release 22 retraining of the TrackClassificationTool (TCT), a Bosted Decision Tree (BDT) that operates on charged-particle tracks contained in the jet under study and assigns to each track a value related to its probability of being produced by the decay products of a b or c-hadron
- Tune the cut applied to the TCT's output, which is used to filter the tracks used by SSVF and JetFitter, and provide different working points of these algorithms based on different values of the cut
- Identify a sensible set of Figures of Merit (FOMs) which may help validating the effects of the changes in SSVF and JetFitter
- Develop two NNs, based respectively on the outputs of SV1 (the b-tagging algorithm based on the secondary vertex reconstructed by SSVF) and JetFitter, in order to evaluate their performance in the new release in concert with the aforementioned FOMs. The FOMs must allow human observers to monitor the physics of the low level algorithms while the NNs take care of extracting their full discrimination power by automating their tuning.

While my QT is still ongoing, I worked on most of these objectives and reported on my work various times to the software and algorithms subgroups of the FTAG group. I learned how to modify the cut on the TCT output and defined various FOMs to test the effects of this change. I developed the NNs based on SSVF and JetFitter output and tested it in the new release. I did a lot of studies and discussions on the best way of using a software developed by the FTAG group for optimisation tasks (the FlavourTaggingPerformanceFramework). As a byproduct, I managed to insert a shortcut in the framework's workflow that allows to add to its output a lot of validation plots that otherwise would have had to be manually reproduced by users.

Graph Neural Networks for TruthTagging The study of the decay channel of the Higgs boson into a $b\bar{b}$ pair when produced in association with a W and Z boson, first observed in 2018, requires to analyze events in which two jets are tagged as b-jets. As for any discriminant separating signal from background, the cuts adopted on DL1r output are such that part of the b-jets are not b-tagged, while part of the c and light-flavour jets are mistagged as b-jets. When an analysis requires events to contain more than one b-tagged jet, inefficiencies multiply. Thus, if the efficiency of b-tagging a b-jet is 70%, the efficiency of tagging events with two b-jets is roughly around 49%. If the probability of mistagging a light-flavour jet as a b-jet is around 1%, the probability of jointly mistagging two in one event is around 0.01%. This very large background event rejection factor has anyway the side effect that huge amount of simulated data is needed in order to model the rate of expected background events in restricted regions of the phase space, if one wants to apply cuts on the b-tagging discriminant variables.

A different approach, named Truth Tagging, consists in weighting the events with their probability of being selected. In order to do so, a good parametrization of the dependency of the efficiency function of the tagger on all the relevant variables is needed. Unfortunately, such dependency is highly non-trivial: among the relevant variables surely figure the kinematic of the jet, but there are others and there is no way to know all of them a-priori, nor to guess the functional form of the efficiency function.

In the past $VHbb$ analyses, parametrization was achieved via efficiency maps binning the dependency of the efficiency on the impulse and direction of each jet. Nevertheless, maps are unable of capturing higher dimensional dependencies due to the curse of dimensionality, i.e. the fact that pantagruelic samples are needed to populate maps when dimensions increase.

Graph Neural Networks (GNNs) provide an alternative solution to the problem. GNNs are Machine Learning algorithms that operate on graphs, which are defined as sets of vertexes and nodes each coming with its own vector representation. They are capable of representing the data and predict some of their proprieties by capturing correlations embedded in the graph structure.

GNNs can be applied to the problem of truth tagging by representing each event as a graph, in which vertexes are the jets and edges are the angles between their directions. The vector representations they build are then fed to a NN which outputs the probability of each jet of being b-tagged. The GNN+NN system is jointly updated backpropagating on an ad-hoc cross entropy loss function.

I contributed to the training of the GNNs for the latest $VHb\bar{b}$ analysis. I presented my results to the group working on the next $VHb\bar{b}$ paper, and my models will be used for the analysis. Furthermore, I joined the effort for developing GNNs to model the efficiency of charm taggers. Such networks are designed with the aforementioned philosophy, but with a different definition of tag. Actually, by changing the way in which the three DL1r scalar outputs are combined, one can use them to tag c-jets instead of b-jets. The latest analysis will take advantage of this for jointly studying the $VHb\bar{b}$ and $VHc\bar{c}$ Higgs boson's decay channels. Thus, the new GNNs will be used for the joint $VHb\bar{b}$-$VHc\bar{c}$ strategy adopted for the upcoming paper.

2 Attended Courses and Exams Given:

Experimental Particle Physics: In order to pass the exam, I presented a lesson about "ATLAS ITk for HighLumi"

Theoretical Physics: Exam not given yet

Advanced Computational Physics: Exam not given yet

3 Conference Presentations

Congresso Nazionale della Società Italiana di Fisica 2021: *"Nuove tecniche di btagging ad alto impulso senza l'utilizzo di tracce"*. 13/9/2021 - 18/9/2021 (online conference).

ATLAS Young Italia 2021: *"Flavour Tagging improvements with the use of ML algorithms"*. 27/9/2021-29/9/2021 (online conference).