UNIVERSITÀ
DEGLI STUDI
FIRENZE

Scuola di
Scienze Matematiche
Fisiche e Naturali
Corso di Laurea Magistrale
in Scienze Fisiche e Astrofisiche

Relatore:
Dott. Lucio Anderlini
`anderlinil@fi.infn.it`

Correlatore:
Prof. Piergiulio Lenzi
`piergiulio.lenzi@unifi.it`

Candidato:
Matteo Barbetti
`matteo.barbetti@stud.unifi.it`

Techniques for parametric simulation with deep neural networks and implementation for the LHCb experiment at CERN and its future upgrades

The LHCb experiment is one of the four detector along the accelerator ring of the Large Hadron Collider (LHC) at CERN, and is dedicated to the study of heavy flavour physics in *pp* collisions. Its primary goal is to look for indirect evidence of phenomena beyond the Standard Model in CP-violation and in rare decays of b- and c-hadrons. In order to improve its statistical power, starting from the Run 3 of LHC, LHCb will operate with a fully software trigger system, which will provide datasets at least one order of magnitude larger. This will allow to reach unprecedented accuracy as long as the Collaboration will be able provide simulated samples as large. As a direct consequence, the production of such simulated samples will dominate the computing effort of the experiment. Reproducing accurately all the physics processes from the *pp* collisions to the radiation-matter interactions within the detectors (the *full* simulation approach) is already now incapable to sustain the analysis demands of the various physics groups, and it is therefore necessary to adopt faster solutions to take full advantage of the upgraded detector. *Ultra-fast* simulation needs lower computing resources, renouncing to reproduce radiation-matter interactions and parameterizing directly the high-level response of the detector. The LHCb subsystems are based on various physics processes, also very different, that make building high-level parameterizations non-trivial: for example, the *Particle Identification* (PID) system combines information from RICH, calorimeter and muon detectors. This task can be carried out effectively by *Generative Adversarial Networks* (GAN), a powerful class of deep learning algorithms able to reproduce highly faithful and diverse probability distributions thanks to a *generative model* learned directly from data.

A large part of this thesis has concerned the development and implementation of *state-of-the-art* GAN algorithms to provide the high-level response of the PID subsystems of LHCb. These neural networks were trained over the calibration samples collected in 2016, in order to provide datasets composed by an unbiased selection of long-lived particles. I have modified the learning procedure to subtract statistically the residual background within the training data, and I have developed an *independent* algorithm capable to measure the quality of the generated samples. This strategy has allowed to build models capable not only to parameterize the high-level response of specific detectors (such as RICH detectors and muon system) to different particles traversing them, but also to reproduce the distributions of variables resulting from the combination of various detector responses. Therefore, given a few basic information such as the particle type, its kinematics and the total number of tracks within the detector, the obtained models are able to synthesize accurately a wide range of probability distributions representing the response obtained from a single detector or from their combination.

The second important personal contribution is related to the design and development of the `mambah` framework, a Python package aimed to provide and manage *user friendly* data structures for High Energy Physics applications. All `mambah` objects were designed to take full advantage of a *batch-grained* framework, using the most modern softwares for parallel computing and exploiting efficiently hardware accelerators, such as GPUs or FPGA. Within `mambah` project, I have dealt with the implementation of database management functions and to the design of the simulation framework based on `mambah`, named `mambah.sim`. The `mambah.sim` module allows to selectively generate particles with the kinematics set by the *pp* collision and to propagate them within the detector defining custom parameterization functions for efficiencies and resolutions: I have developed the parameterization for the PID system of LHCb. I have proved the correctness of the implemented models, showing the generalization capabilities of GANs in describing decay channels different from the one of training. Lastly, I have shown that the samples produced by `mambah.sim` are competitive with the *full* simulated ones, while ensuring a significant reduction of the computing cost.