PROCEEDINGS OF SCIENCE

Generative models and seq2seq techniques for the flash-simulation of the LHCb experiment

Lucio Anderlini,[a] **Matteo Barbetti,**[b,*] **Simone Capelli,**[c,d] **Gloria Corti,**[e] **Adam Davis,**[f] **Denis Derkach**[g] **and Maurizio Martinelli**[c,d] **on behalf of the LHCb Simulation Project**

[a]*Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Firenze, Italy*
[b]*Istituto Nazionale di Fisica Nucleare (INFN), CNAF, Italy*
[c]*Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Milano-Bicocca, Italy*
[d]*Department of Physics, University of Milano-Bicocca, Italy*
[e]*European Organization for Nuclear Research (CERN), Switzerland*
[f]*Department of Physics and Astronomy, University of Manchester, United Kingdom*
[g]*Faculty of Computer Science, HSE University, Russia*

E-mail: Lucio.Anderlini@cern.ch, Matteo.Barbetti@cern.ch

Simulating detector and reconstruction effects on physics quantities is crucial for data analysis, but it is coming unsustainably costly for the upcoming HEP experiments. The most radical approach to speed-up detector simulation is Flash Simulation, as proposed by the LHCb collaboration in Lamarr, a software package implementing a novel simulation paradigm relying on Deep Generative Models and Seq2seq attention-driven techniques to deliver simulated samples. Thanks to its modular layout, Lamarr provides analysis-level quantities by applying a pipeline of machine learning-based modules that properly transforms the information resulting from physics generators. Good agreement is observed by comparing key reconstructed quantities obtained with Lamarr against those from the existing detailed Geant4-based simulation. Lamarr has been designated with dual capabilities: it can function as a stand-alone simulation framework, while also being seamlessly integrated into the LHCb simulation software.

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

[*] Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

https://pos.sissa.it/

PoS(ICHEP2024)1032

1. Introduction

The LHCb experiment [1] was designed to study heavy hadron decays, namely processes that involve particles containing b and c quarks, at the Large Hadron Collider (LHC). To this end, the detector is configured as a single-arm forward spectrometer covering the pseudorapidity range of $2 < \eta < 5$. It includes a Tracking system for momentum p and primary vertex (PV) precise measurements and a Particle Identification (PID) system. The PID system comprises RICH detectors for distinguishing charged hadrons, calorimeters for identifying photons, electrons and hadrons, and a dedicated system for muon identification named MUON.

The baseline for simulation at LHCb is the *Detailed Simulation* which relies on first-principle fine-tuned models for radiation-matter interactions occurring within the detector materials. The LHCb simulation software is built upon two main applications: Gauss and Boole. Gauss handles the *generation* and *simulation* phases, whose final goal is to compute the energy deposited by the long-lived particles capable of partially or fully traversing the detector active volumes. The energy deposits are then converted into raw data in a step called *digitization*, which is implemented in the Boole application.

The simulation of the radiation-matter interactions occurring within the detector, mainly handled by Geant4 [2], is the primary consumer of CPU resources at LHCb, accounting for more than 90% of the total computing budget during the LHC Run 2 (2015 – 2018). The new Run 3 LHCb detector [3] is collecting a larger amount of data with respect to the previous data-taking, thus putting severe pressure on the CPU resources to retain constant the ratio between statistical uncertainties on the collected and simulated data. Relying only on *Detailed Simulation* to meet the upcoming and future requests for simulated samples would far exceed the computing resources pledged to the experiment. Hence, a major transformation of the LHCb simulation software stack is needed, including not only code optimization and modernization but also the development of novel and faster simulation options.

2. Fast simulation VS. flash simulation

Simulating the multitude of physics processes happening within the active volume of a detector is extremely resource-expensive in terms of computing power. This challenge is shared across the High Energy Physics (HEP) community, which is collectively investing significant effort into studying alternative strategies to reduce the CPU cost of simulations. The LHCb Collaboration participates in this joint effort, developing solutions to optimize the simulation [4] and, where possible, reuse results from previous computations [5].

Replacing part of the Geant4 computations with results obtained through alternative methods (e.g., via parameterizations) can significantly reduce the simulation costs, particularly when these substitutions are made for the most resource-intensive computations. A common example is the simulation of the calorimeter response which, involving a cascade of secondary particles, dominates the CPU usage in HEP experiments. Whenever the computation of energy deposits relies on parameterizations [6, 7] instead of Geant4, literature refers to these techniques as *Fast Simulation*.

A more radical approach is employed by *Flash Simulation* (also known as *Parametric* or *Ultra-Fast Simulation*) strategies, which replace the entire simulation step, and sometimes the

PoS(ICHEP2024)1032

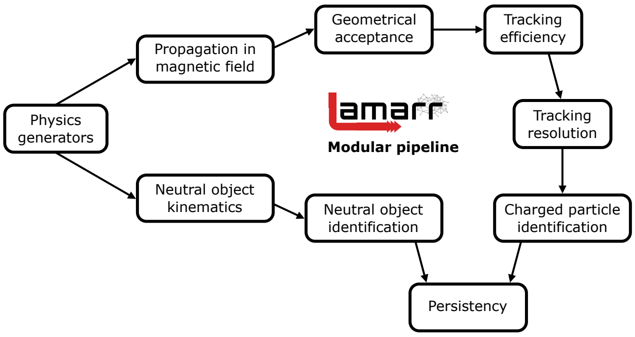
Physics generators
Propagation in magnetic field
Geometrical acceptance
Tracking efficiency
Tracking resolution
Charged particle identification
Neutral object kinematics
Neutral object identification
Persistency
lamarr
Modular pipeline

Figure 1: Scheme of the Lamarr modular pipeline. According to the charge of the particles provided by the physics generator, two sets of parameterizations are defined: the charged particles are passed through the Tracking and PID models, while the neutral ones follow a different path where the calorimeter modeling plays a key role.

PoS(ICHEP2024)1032

subsequent digitization and reconstruction steps, with parameterizations, typically implemented using Deep Generative Models. Following pioneering works that demonstrated the validity and reliability of using *Generative Adversarial Networks* (GAN) to reproduce the high-level response of RICH detectors [8], LHCb has begun developing a novel simulation framework based on the flash-simulation paradigm, called Lamarr [9] and discussed in the next Section.

3. Lamarr: the flash-simulation option for the LHCb experiment

Lamarr [9] is the official flash-simulation framework for LHCb, designed to offer the fastest option for simulation. The Lamarr framework consists of multiple pipelines of parameterizations that follow one another to reproduce the high-level response of the LHCb sub-detectors when traversed by quasi-stable particles as provided by the MC physics generators. Most of these parameterizations rely on Deep Generative Models, and in particular on GANs. The Lamarr pipeline can be logically split into two separated chains, based on the charge of the generated particles. A schematic representation of the Lamarr modular pipeline is shown in Figure 1.

3.1 The charged particles pipeline

When detected and properly reconstructed, a charged particle traversing the LHCb spectrometer gives origin to a track. Charged particles with successfully reconstructed tracks are assigned to a PID interpretation based on the response of the RICH detectors, the calorimeters, and the MUON system. Under the assumption that each quasi-stable generated particle is associated with at most one reconstructed candidate, the response of any detectors can be described in terms of two types of parameterizations: *efficiency*, which models when to exclude particles due to failed reconstruction, and *resolution*, which describes, in general, how to use generator-level information to reproduce reconstructed quantities. Assuming that k reconstructed objects originated from k generated particles (k-to-k relation) is sufficient to accurately parameterize Tracking and PID response.

Given a set of quasi-stable particles (e.g., muons, pions, kaons, protons) provided by the MC physics generators, the first step is to determine which of these particles fall within the fiducial volume of the LHCb spectrometer (*geometrical acceptance*). Lamarr relies on a neural network (NN) trained to solve this binary classification problem, allowing to select the subset of particles that succeed in traversing the detector. A second NN model is then employed to identify which of the particles in acceptance are associated with a reconstructed track, parameterizing the *tracking efficiency*. At this stage, the set of reconstructed tracks still rely on the generator-level

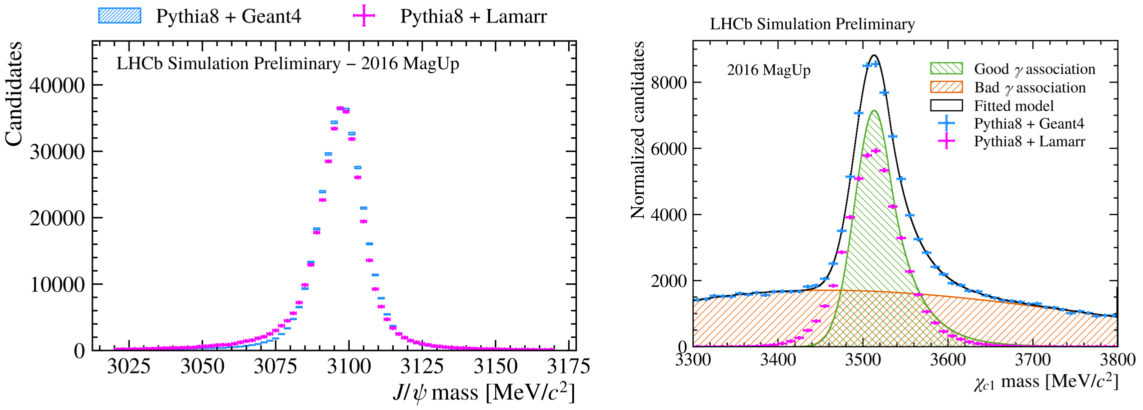

Figure 2: Validation plots for the invariant mass of J/ψ (left) and χ_{c1} (right) coming from $B^+ \to \chi_{c1}K^+$ decays with $\chi_{c1} \to J/\psi\gamma$ and $J/\psi \to p\bar{p}$. The *detailed simulated* samples obtained with Geant4 are compared with Lamarr *flash simulated* samples.

information and does not account for the *resolution effects* and the *reconstruction errors*. In Lamarr these contributions are parameterized using GANs trained with the generator-level information as input conditions. Then, the PID interpretation is added to each of the reconstructed tracks. The high-level response of the Tracking system, as derived from the previous steps of the Lamarr pipeline, is combined with a shallow model for the detector occupancy to parameterize the PID system. Two specialized classes of GAN models are employed to reproduce the high-level response of the RICH detectors and the MUON system. Finally, the outputs of the RICH, MUON, and Tracking models are stacked together with an efficiency model for muon candidates to parameterize the *global response* of the LHCb PID system. Further details on model design, training strategies, and performance can be found in Ref. [10].

Among the various validation campaigns performed during the development of Lamarr, we briefly discuss here the performance on simulated $B^+ \to \chi_{c1}K^+$ decays with $\chi_{c1} \to J/\psi\gamma$ and $J/\psi \to p\bar{p}$. Sample of B-mesons and their decay products are produced with Pythia8 [11] and EvtGen [12] respectively. In the detailed simulation, the quasi-stable particles so produced (i.e., kaons, photons, and protons) are propagated through the spectrometer with Geant4 which allows to compute the energy deposits, which are then combined into a set of reconstructed quantities by the LHCb data processing software. The same set of generated particles are also processed with Lamarr to produce flash simulated samples of the LHCb detector. The pipeline described above to parameterize the Tracking and PID systems succeeds in describing what results from the Detailed Simulation even if the models were trained using decay channels different from the ones used for validation. As a demonstration, in Figure 2 (left), we report the accurate agreement between the invariant mass of J/ψ resulting from standard simulations (in cyan) and the ones produced via Lamarr (in magenta).

3.2 The neutral particles pipeline

As part of the PID system, the primary role of the LHCb calorimeters is to enable the separation of photons from π^0 candidates and to contribute to the identification of electrons. The reconstruction algorithms allow to distinguish neutral from charged particles by studying the absence (or

PoS(ICHEP2024)1032

presence) of tracks in front of the energy deposits collected within the calorimeters, which are grouped into clusters. The electromagnetic calorimeter (ECAL) response is typically concerned by bremsstrahlung radiation, converted photons, or merged π^0, which may lead to having n generated particles responsible for m reconstructed candidates (in general with $n \neq m$). This leads to ambiguous relations between the generated particles and the associated reconstructed candidates, making it challenging to define a flash-simulation pipeline for the ECAL response.

However, the pipeline described above for charged particles provides a clear and effective model for the detectors high-level response, and it is therefore difficult to abandon this approach. This is especially true for signal particles, where reliable parameterizations are crucial. In Lamarr, signal photons are parameterized following the k-to-k assumption valid after having enforced unambiguous relations under some geometrical constraints. This allows to describe the high-level response of the ECAL detector in terms of acceptance and efficiency, relying on NN models, and resolution, based on the output of specialized GANs. While these parameterizations accurately reproduces what expected from detailed simulated samples once imposed the same geometrical constraints (see Ref. [10]), having k-to-k relations proves to be an oversimplified assumption that leads to several limitations in the description of the ECAL response. The problem becomes evident by comparing detailed and flash simulated $B^+ \to \chi_{c1} K^+$ decays with $\chi_{c1} \to J/\psi \gamma$ and $J/\psi \to p\bar{p}$. In particular, looking at the χ_{c1} invariant mass in Figure 2 (right), we notice an inconsistency between what results from standard simulations (in cyan) and what obtained with Lamarr (in magenta). This mismodeling stems from an *ephemeral* relation between neutral particles and reconstructed candidates, which leads to a fraction of "wrong" clusters being associated with the signal photons by construction. This explains the plateau reported in orange in the χ_{c1} mass plot, which Lamarr is unable to reproduce due to the geometrical constraints underlying the k-to-k assumption. To address these limitations, it is necessary to face the n-to-m problem by directly parameterizing the particle-to-particle correlations. A preliminary attempt has been made for background photons using *Seq2seq* and *Graph2graph* models, as described in Refs. [9, 10]. We are currently working on an extension of those models, aimed to precisely describe the response of the ECAL detector once traversed by signal photons, and powered by a Decoder-only Transformer architecture [13].

4. Conclusion

Evolving simulation techniques and their integration in the LHCb software stack are essential to meet the growing demand for simulated samples expected during Run 3 and beyond. The flash-simulation paradigm offers a valuable solution for reducing the pressure on the pledged CPU resources, without significantly compromising the description of the uncertainties introduced in the detection and reconstruction phases. These techniques, powered by Deep Generative Models, are made available to LHCb via the novel Lamarr framework. The parameterizations for the Tracking and the charged PID systems are robust and have been repeatedly validated on decay channel samples that represent only a small fraction of the datasets used for training. However, for modeling the ECAL response, there is no alternative than directly facing the particle-to-particle correlation problem. Ongoing investigations are focused on parameterizing the n-to-m relations using Seq2seq and Graph2graph models.

PoS(ICHEP2024)1032

Acknowledgements

The work presented in this paper is performed in the framework of Spoke 0 and Spoke 2 of the ICSC project – *Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing*, funded by the NextGenerationEU European initiative through the Italian Ministry of University and Research, PNRR Mission 4, Component 2: Investment 1.4, Project code CN00000013 - CUP I53C21000340006.

References

[1] LHCb collaboration, *LHCb Detector Performance*, *Int. J. Mod. Phys. A* **30** (2015) 1530022 [1412.6352].

[2] Geant4 collaboration, *Geant4 developments and applications*, *IEEE Trans. Nucl. Sci.* **53** (2006) 270.

[3] LHCb collaboration, *The LHCb Upgrade I*, *JINST* **19** (2024) P05065 [2305.10515].

[4] M. Mazurek, G. Corti and M. Kmie, *Performance of the Gaussino CaloChallenge compatible infrastructure for ML-based fast simulation in the LHCb Experiment*, in *22nd International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT)*, 2024.

[5] D. Müller et al., *ReDecay: A novel approach to speed up the simulation at LHCb*, *Eur. Phys. J. C* **78** (2018) 1009 [1810.10362].

[6] LHCb collaboration, *Calorimeter fast simulation based on hit libraries LHCb Gauss framework*, *EPJ Web Conf.* **214** (2019) 02040.

[7] V. Chekalina et al., *Generative Models for Fast Calorimeter Simulation: the LHCb case*, *EPJ Web Conf.* **214** (2019) 02034 [1812.01319].

[8] LHCb collaboration, *Towards Reliable Neural Generative Modeling of Detectors*, *J. Phys. Conf. Ser.* **2438** (2023) 012130 [2204.09947].

[9] L. Anderlini et al., *The LHCb ultra-fast simulation option, Lamarr design and validation*, *EPJ Web Conf.* **295** (2024) 03040 [2309.13213].

[10] M. Barbetti, "The flash-simulation paradigm and its implementation based on Deep Generative Models for the LHCb experiment at CERN." CERN-THESIS-2024-108, 2024.

[11] T. Sjostrand, S. Mrenna and P.Z. Skands, *A Brief Introduction to PYTHIA 8.1*, *Comput. Phys. Commun.* **178** (2008) 852 [0710.3820].

[12] D.J. Lange, *The EvtGen particle decay simulation package*, *Nucl. Instrum. Meth. A* **462** (2001) 152.

[13] A. Vaswani et al., *Attention Is All You Need*, in *31st International Conference on Neural Information Processing Systems (NeurIPS)*, 6, 2017 [1706.03762].

PoS(ICHEP2024)1032