April 3, 2021

Redesigning the role and scope of Monte Carlo simulations for physics studies

Ivan Amos Cali[1], Yen-Jie Lee[1], Camelia Mironov[1],

[1]Massachusetts Institute of Technology, Laboratory for Nuclear Science, Cambridge, Massachusetts 02139

Particle and Nuclear Physics Research: Redesigning the role and scope of Monte Carlo simulations for physics studies

Principal Investigator:	Prof. Yen-Jie Lee
Institution:	Massachusetts Institute of Technology
Phone:	(617) 324-7418
E-mail:	yenjie@mit.edu

Funding Opportunity Announcement Number: DE-FOA-0002493

1 Motivation

The Monte Carlo (MC) simulations play several key roles when analyzing data. Whether one looks at 'simple' e^++e^-or p+p collision systems, or more complex ones in which ions are also involved (e+A, p+A, or A+A), MC samples, for a signal of interested, are needed to: i) model the background levels and its source (*i.e.*, from other physics processes, or from random combinations); ii) model the signal passing through the detector after it was produced in the collision (in order to assess how its efficiency and purity is affected by the reconstruction). With these two notions under control and understood from MC simulations, one then can be confident in the interpretation of the measurements in real data.

However, there is no perfect knowledge and explanations for all physics phenomena observed so far in data, and therefore, no MC event generator that can reproduce all Standard Model processes. Event generators are usually being updated and new tunes are released regularly, as newer and/or more precise measurements become available. This is a process that while easier (though not trivial) in e^++e^- and p+p collisions, it is mush more complicated for A+A collisions. The complexity of one A+A collisions (which can produce order of magnitudes more particles than an e^++e^-or a p+p collision at the same central of mass collision energy) means also, at present, that the richness of the underlying physics is much less known than that of a e^++e^-/p+p collision, so no tuning is fully reliable because the underlying source/cause of an observed effect is not known.

Moreover, in order for the detector response to be reliable, an underlying requirement of all MC studies is that one can model all the data-taking period conditions (*i.e.*, the 'run conditions'): alignments between detectors, calibrations, vertex distribution, etc. But, more often than not, data analyses have to introduce additional 'residual' corrections due to parts of data reconstruction or detector response not understood based on MC modeling. The end result is that the physics results carry additional 'uncertainties', diluting in this way the power of interpretation and the message these results carry.

Whether looking at signal or background in MC simulations, an additional factor that has to be factored in, is that the statistical precision of the MC studies has to keep up with that of data. But, taking LHC as example, this means in that in the span of 10-15 years, when the data size increased by factor of XX in p+p(2010 vs 2018) and YY in Pb+Pbcollisions (2010 vs 2018), the MC samples had also to be increased. And while it is fast to produce a p+pMC event (x sec/event to generate + fully reconstruct), it is not trivial to produce a Pb+Pbone (xh/event from generation to reconstruction). A tremendous amount of resources are requested, to address the needs of scientifically competing groups, prioritizing analyses, conferences, and publications depending on the availability of MC samples. Hundreds of PBs of MC sample have to be prioritized and generated each year, and one or two times per year also discarded, in order to produce new samples that reflect more studies, using data, that reproduce better the run conditions.

Mitigations to address the present situation (inaccuracies in describing data and incremental increased in sample sizes) include lowering the event content in the output file (to be able to store larger MC samples, but with less information and hence being more constrain in what can be done with it), using fast simulations in which the detector response is parametrized (making the simulation process faster but inescapable less accurate), and overall just buying more computing resources and storage.

This document proposes to place real events at the center of 'data analysis' instead of MC events. But also, for the purpose of 'unknown physics', to tune better the MC generators.

2 Proposal

Our plan is a 3-way high-way:

a) Data to ML. To reproduce most accurately the collision environment (and hence to remove the need for several MC campaigns and bypass the need for increasing resources), we plan to develop a ML algorithm that when is fed real data it separates/tag background samples from signal samples. This will produce the ultimate, most accurate background samples, with in situ knowledge of both physics and run conditions. b) MC to ML. To shorten the simulation time, the ML algorithm will help replace the 'person made' and imperfect description of the detector response included in fast simulation.

c) Data to MC. In order to understand the physics processes in data (old or novel), the ML algorithm will be fed real events and asked to dismantle them in each of its constituents. A 'particle level' analysis as the one possible in MC events.

3 Technical approach

We have already started looking at several ML algorythms. Preliminary results, using the open source data from ALEPH are shown in Fig. 1

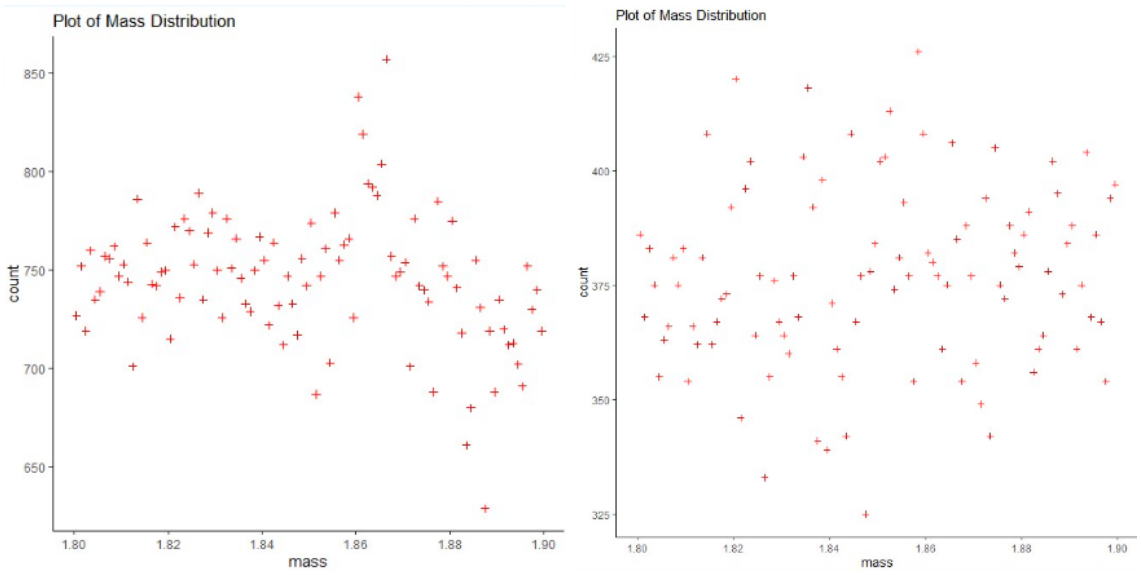

Figure 1: Left: Invariant mass distributions for D meson in e^++e^-ALEPH data. Right: Output from GAN, when trained with data from left.