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-0002501

Motivation Test test Test

Analyses of large physics data sets typically rely on extensive Monte Carlo (MC) simulations that aim to describe the underlying physical processes, backgrounds, and detector response. We propose to develop an entirely different approach, augmenting or replacing the traditional role of MC simulations with a Machine Learning (ML) based framework trained on data, therefore placing real data at the center of the data analysis effort. This framework, which employs a Deep Neural Network (DNN) and the recently developed Generative Adversarial Networks (GAN), will allow to generate MC samples that reproduce detector data entirely. The GAN and DNN based algorithms will be also used to perform transformation of the data point-cloud to a "truth level" point-cloud, with the ultimate goal of dissecting the data down to its basic information content in terms of known and novel physics particles and processes, opening an era of ML-based discovery physics.

Prototypical examples of the current use of MC simulations are found in high-energy and nuclear physics, for both elementary proton-proton (p+p) or electron-electron (e^++e^-) collisions systems, and for the complex collisions involving atomic nuclei (e+A, p+A, or A+A). Here, simulations are needed to model background contributions, their sources, and to determine the response matrix of detectors. Simulation are used also to separate signal and background particles, extracting parameters like efficiency, resolution, and purity. This information obtained from MC simulations enables the interpretation of measurements in real data. However, knowledge and models for the sum of all physics phenomena observed in data are typically incomplete, and no MC simulation can perfectly reproduce all physics processes. Simulations are continually updated through the release of new "tunes", as more precise understanding develops. The complexity of one A+A collisions (which can produce orders of magnitude more particles than one e^++e^-/p+p collision) also means that to-date, the richness of the underlying physics is much less well understood than for e^++e^-/p+p, making the MC tuning much less reliable. This is a major challenge for present and future experimental programs at the CERN Large Hadron Collider (LHC) and the RHIC and EIC facilities at Brookhaven National Laboratory.

Reliable MC modeling requires consideration of time-dependent variations in detector conditions and response, such as alignment between detectors, calibrations and detector aging. This often leads to the introduction of additional 'residual' corrections in the analysis process, accounting for aspects not understood in simulations. As a consequence the physics results carry additional uncertainties, diluting the power of interpretation and new physics knowledge they carry. High-accuracy/sensitivity and in general any discovery studies (*e.g.*, Higgs or neutrino measurements) are extremely sensitive to the reliability of the MC modeling.

Whether looking at signal or background in MC simulations, an additional factor is the statistical precision of the MC studies, which should be commensurate with that of the collected experimental data. Taking LHC as example, this leads to the additional challenge that improvements in accelerator performance and data collection require increasingly large data samples, far outpacing the increase in available computing resources (CPU and storage) for simulations. This challenge will become more pressing in the future "high luminosity" era of LHC operations, requiring resources for MC sample production that cannot be easily met (e.g., an estimated increase from 40 to 500PB of disk needs from 2020 to 2030). Availability of workforce and computing resources to address the simulation needs of many concurrent analyses, often involving multiple iterations, is rapidly becoming the major bottleneck in the overall data analysis enterprise. Current mitigation strategies to address high statistics MC sample needs include discarding part of the information content in the simulation process and using so-called fast simulations, in which the detector response is approximated through parametrizations. Such strategies inevitably lead to less overall accuracy.

Proposal Our plan has three main objectives as illustrated in the flow-chart in Fig. **??** (left):

Data to ML: To reproduce most accurately the collision environment (background, signal, and run condi-

tions), we plan to develop a GAN algorithm fed by real data and capable of separating/tagging background samples from signal samples in the data itself.

MC to ML: To shorten the simulation time, the GAN/DNN framework will replace the 'person made' and imperfect description of the detector response included in fast simulation through an automatic, data-driven generation of relevant parametrizations.

Data to MC: To uncover known and unknown physics processes in data, the GAN/DNN framework will be fed real events and tasked to disassemble them into basic sets of constituents, allowing a 'particle level' analysis close to the one possible at "truth" level in MC events. Importantly, this approach will make it possible to continue analysis of archived data from completed experiments, for which no new simulations are available.
Figure 1: Left: Flow-chart of proposed project. Right: Invariant mass distributions for D^0 meson in e^++e^- ALEPH data (blue), and from GAN (red).

We have performed feasibility studies for this new approach, evaluating so far one GAN algorithm, tableGan (a convolutional neural network known for its performance and flexibility in representing data), with which we have created first demonstrators using cloud computing resources. These evaluations were performed using e^++e^- archived data from the ALEPH experiment (which ended data taking in 2000), for which the background levels are low, and the theoretical descriptions are more precise. The input and output point-clouds are based on particle momentum vector, type and charge, as well as event multiplicity. Preliminary results are shown in Fig. **??** (right). Importantly this demonstrates that while the input for the tableGAN training was based on single particle observables, the GAN based algorithm could also describe correlated two-particle observables, such as the invariant mass distribution shown. This includes both the bulk (combinatorial) background shape and fine details in the simulation such as the D^0 meson resonance peak. In our approach, one could then use the output particles to perform other physics analysis (*e.g.*, reconstruct other mesons decays). While still at an early stage, the ML implementation today is already demonstrating that it could generate e^++e^- events that are similar to real data. Before deciding on the final approach, other algorithms will be evaluated. These will include ctGAN, which was demonstrated to generate synthetic tabular data with high fidelity; α-GAN, a hybrid approach combining a likelihood based model (variational auto-encoders, VAEs), and an implicit generative model (GANs) to combine an adversarial loss with a data reconstruction loss; VQ-VAE-2, a vector quantized VAE model, which uses hierarchical multi-scale latent maps for achieving an increased resolution; InterFaceGAN, which interprets the latent semantics learned by GANs for semantic face editing.

The proposal aims to conduct foundational research to develop reliable and efficient ML tools and approaches to exploit new computational technologies in order to find solutions for one of the most challenging problems in nuclear, particle and neutrino physics, and many areas of high-precision research in general: the size of the MC samples and the accuracy of MC modelling. The proposed research also aims to profoundly change data analysis, employing graph and network algorithms for discovery: training the machine to uncover yet-unknown physics.

Table 1: PI and Senior/Key Personnel and their Institutional Affiliations

Last Name	**First Name**	**Title**	**Institution**
Cali	Ivan Amos	Research Scientist	Massachusetts Institute of Technology
Chen	Yi	Senior Postdoc	Massachusetts Institute of Technology
Lee	Yen-Jie	PI	Massachusetts Institute of Technology
Mironov	Camelia	Research Scientist	Massachusetts Institute of Technology
Roland	Gunther	Professor	Massachusetts Institute of Technology

Table 2: Collaborators, Co-editors, and Graduate and Postdoctoral Advisors and Advisees of the PI and Senior/Key Personnel

Last Name	**First Name**	**Title**	**Institution**
Bi	Ran	Postdoc	CU Boulder
Busza	Wit	Professor, Emeritus	Massachusetts Institute of Technology
Chang	Paoti	Professor	National Taiwan University
Chien	Yang-Ting	Senior Researcher	Stony Brook University
Citron	Zvi	Senior Lecturer	Ben-Gurion University of the Negev
Dainese	Andrea	Physicist	INFN - Sezione di Padova
Dellacasa	Giuseppe	Professor	Università del Piemonte Orientale
Dong	Xin	Staff Scientist	Lawrence Berkeley National Laboratory
Granier de Cassagnac	Raphael	Scientist	CNRS
Grosse-Oetringhaus	Jan Fiete	Staff	CERN
Hofman	David	Professor	UIC
Huang	Jin	Physicist	Brookhaven National Laboratory
Innocenti	Gian Michele	Staff	CERN
Jowett	John	Accelerator Physicist	CERN
Kunde	Gerd	Scientist	LANL
Liu	Ming	Scientist	Los Alamos National Laboratory
Manzari	Vito	Scientist	INFN
Margetis	Spyridon	Professor	Kent State University
McGinn	Christopher	Postdoc	CU Boulder
Morrison	Dave	Physicist	Brookhaven National Laboratory
Nguyen	Matthew	Scientist	CNRS
Rapp	Ralf	Professor	Texas A&M University
Tatar	Kaya	Fellow	CERN
Thaler	Jesse	Associate Professor	Massachusetts Institute of Technology
Veres	Gabor	Professor	ELTE Institute of Physics
Wang	Jing	Postdoc	Massachusetts Institute of Technology
Wang	Minzu	Professor	National Taiwan University
Weidermann	Urs	Staff	CERN
Winn	Michael	Staff Scientist	CEA/IRFU, DPhN
Wyslouch	Boleslaw	Professor	Massachusetts Institute of Technology