Núcleo de
Computação
Científica

Applications of Machine Learning Techniques to the HL-LHC Experiments

Serrapilheira CALL 2/2018

High Energy Physics (HEP) explores the elementary particles, which are the fundamental constituents of matter. HEP experiments led to important discoveries, such as the detection of the Higgs boson at the CERN Large Hadron Collider - LHC. The High-Luminosity LHC (HL-LHC) is the next challenge in the HEP scenario, bringing the collider's instantaneous luminosity to 20 Hz/nb and increasing in 5 times the amount of additional pp interactions in the same or neighboring bunch crossings, referred to as pileup (PU). In order to deal with the increased amount of generated data and the complexity of the simulations, new techniques and frameworks have to be deployed and/or developed. In that scenario, the Deep Neural Networks (DNN) revolution can make a significant impact on HEP. These techniques are most promising when there are both a large amount of data and a high number of features. This project proposes an exploration of the usage of advanced machine learning techniques at the HL-LHC.

Coordinator: Thiago Rafael Fernandez Perez Tomei
Host Institutions: Center for Scientific Computing – NCC-Unesp

Contents

1 Introduction

High Energy Physics (HEP) explores the elementary particles, which are the fundamental constituents of matter, and their interactions. Elementary particles are the underlying structure at the inner kernel of matter and, at the same time, plays an essential role in the evolution of the Universe. The last century has shown that collider accelerators have been among the most powerful tools used to explore the deep structure of matter that enabled the development of a universal quantum field theory — the standard model. HEP experiments led to important discoveries that go from the identification of heavy quarks, passing by the discovery of the $W^{\pm}$ and Z^{0} bosons, up to the breakthrough represented by the recent discovery of the Higgs boson at CERN [1, 2], evidence for the Brout-Englert-Higgs mechanism through which fundamental particles acquire their mass.

The ***Large Hadron Collider*** (LHC) at CERN, the most advanced facility in operation, has opened new opportunities to explore the energy frontiers of physics, colliding protons at 13 TeV, which allows to explore distances of 10^{-18} meters. The collider has two large general-purpose detectors, ATLAS (A Toroidal LHC ApparatuS) [3] and CMS (Compact Muon Solenoid) [4], operating on the Swiss-French border near Geneva, Switzerland. Both collaborations are examples of truly international endeavors. For instance, the CMS collaboration comprises more than 210 institutions from 48 countries and has become a very successful example of large international cooperation. The impact, in terms of scientific results, can be attested by the number of papers that CMS has published: over 750 articles based on collider data since 2010, including the one that reported the discovery of the Higgs boson, which has almost 8,500 citations and led to the award of the 2013 Nobel Prize in Physics. As of this writing, the LHC has just finished its second operation run (Run 2) [5], having delivered more than 150 fb^{-1} of 13 TeV pp collision data to both ATLAS and CMS.

The LHC has, unsurprisingly, been put on the center stage of the European Strategy for Particle Physics [6]. It is also a highlight of the recent report from the Particle Physics Project Prioritization Panel (P5) from the U.S. Department of Energy [7]. The High-Luminosity LHC (HL-LHC) [8] is the next step of this strategy, bringing the instantaneous luminosity delivered to experiments to 20 Hz/nb. Over the proposed operation period of the HL-LHC, from 2026 to 2035, this will entail a twenty-fold increase of the produced data. All aspects of the experiments will have to undergo essential improvements, in what is known as the Phase-II upgrade, to cope with the increase of luminosity and to survive the radiation levels that the detector will experience during the HL-LHC period. The LHC schedule, including the HL-LHC era, is shown in Fig. 1. One particular aspect of the HL-LHC operation will be the much higher occurrence of additional pp interactions in the same or neighboring bunch crossings, referred to as pileup (PU). The average pileup on standard HL-LHC can reach up to 200, in comparison to PU$\sim$ 40 observed during the LHC Run 2. This leads to much more complex events that are difficult to treat at every level – simulation, data acquisition, reconstruction and analysis. A simulated HL-LHC event is shown in Fig. 2, extracted from Ref. [9].

In order to seize the opportunities and address the challenges posed by the HL-LHC, the HEP field is moving closer to the frontier of information technology and computer science [10]. The new techniques, tools and frameworks from the field of ***machine learning*** (ML) are ideally suited to the HL-LHC era [11]. The usage of ML techniques is not foreign to HEP; indeed, the ROOT data analysis

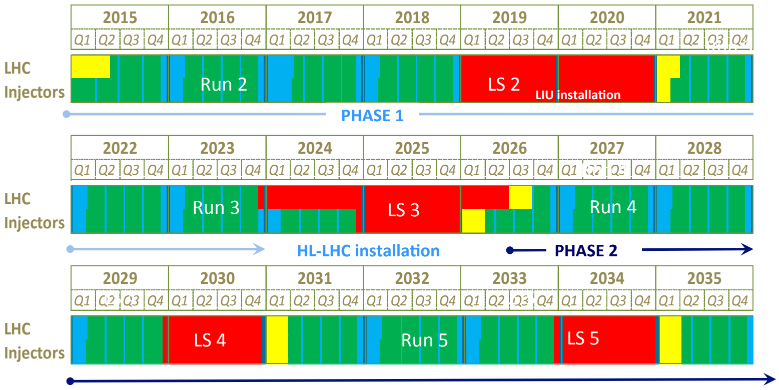


Figure 1: Long-term LHC Schedule including the Physics Programs (Run), Long Shutdown (LS), Beam Commissioning and Technical Stops.

framework has incorporated a Toolkit for Multivariate Analysis (TMVA) since 2013 [12]. Solutions based in machine learning have been extensively used in the LHC experiments, for identification of heavy-flavour jets [13], reconstruction and identification of τ leptons [14], energy reconstruction of electrons [15] and even high-level physics studies like searches for resonant $t\bar{t}$ [16] and supersymmetric partners of the top quark [17]. The Deep Neural Networks (DNN) revolution [18] has, however, made significant impact on HEP; it is particularly promising when there are both a large amount of data and a high number of features, as well as symmetries hidden in the data and complex nonlinear dependencies between inputs and outputs. All those are primary characteristics of HEP experiments' data and the HL-LHC will be no exception.

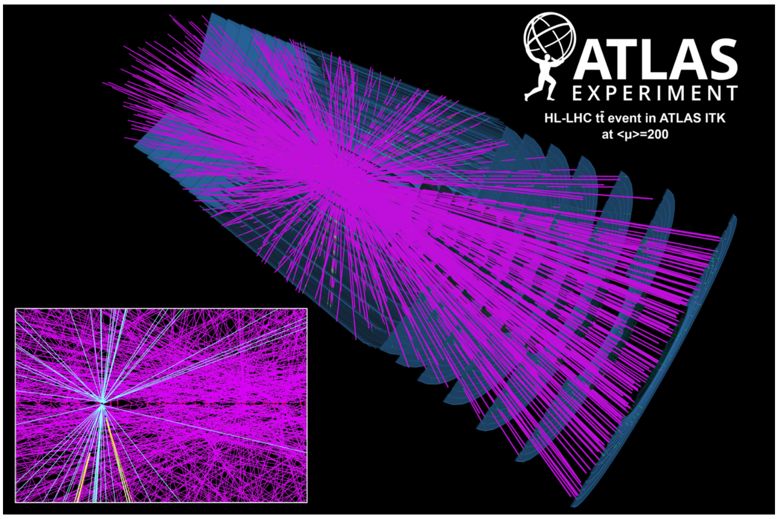


Figure 2: Simulation of a $t\bar{t}$ event, at average pileup of 200 collisions per bunch crossing. Image courtesy of the ATLAS Collaboration [9].

This project proposes an exploration of the usage of advanced machine learning techniques at the HL-LHC. Our primary goal will be the exploration, study, development and deployment of those techniques for simulating and analyzing HL-LHC pp collision events in the HL-LHC conditions. This work will be done in the context of the CMS collaboration; however, the results obtained during this project will be applicable to all sorts of HEP experiments, including future collider [19, 20] and non-collider experiments [21, 22].

2 Methodology

The methodology to develop this research will be focused on applied Artificial Intelligence; the scientific focus of the project is still High Energy Physics. In this sense, we plan to develop new machine learning models using the most powerful and popular frameworks for machine learning, such as Keras [23] and Tensorflow [24]. For the hardware resources needed for that endeavour, we will leverage the computing resources at the Center for Scientific Computing (NCC-Unesp) [25].

2.1 Generative Models for Detector Simulation

We are particularly interested in investigating ways to speed-up the simulation methods for the benefit of the HEP community. Currently, almost all event simulations are done with the Geant4 framework [26], which can take minutes for complex events like those expected at the HL-LHC. Previous work in that area has resulted in approaches like the ATLAS FastCaloSim [27] and the CMS FastSim [28], in which the authors report large speed-ups on the calorimeter simulation. Both approaches promote the speed-up by optimizing algorithms for simulating the various components of the calorimeter, but at the expense of sacrificing some of the accuracy of the simulation. NCC has also participated in the Geant V project, where improvement on the overall level of parallelization of the simulation code [29] comes from the use of modern hardware, such as accelerators and GPUs.

An alternative approach is to employ machine learning techniques to improve the performance of simulations. We will investigate the usage of generative models, such as (deep) generative adversarial networks (GAN) [30], for simulation of physics events. These models work by simultaneously training two neural networks, with one network being responsible for building synthetic images and the other being responsible for classifying these images as fake or real. This approach has already shown some results [31], where the authors use deep generative networks for simulating the ATLAS liquid argon calorimeter; some results are shown in Fig. 3. We intend to investigate the adherence of these models to the calorimeter simulation including the state of art GANs such as the Wasserstein GANs (WGAN) [32] or the Deep Convolutional GANs (DCGAN) [33].

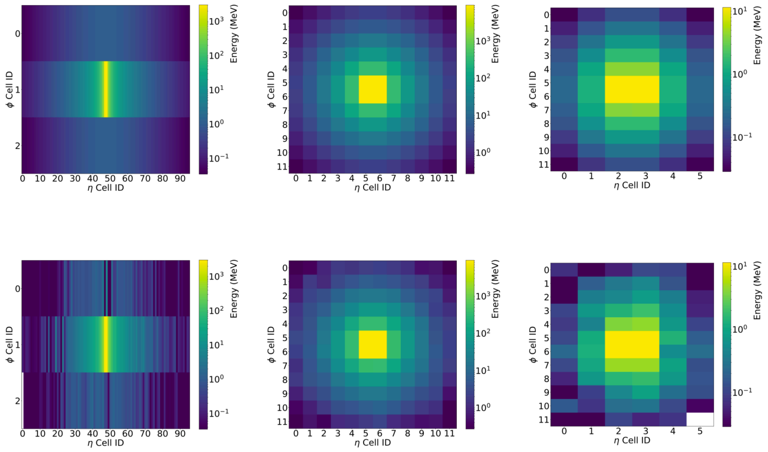


Figure 3: An application of ML techniques to a HEP problem: simulating an e^+ incident in a three-layer liquid argon calorimeter with the Geant4 framework (top) and with a special purpose GAN (bottom). Figure extracted from Ref. [31].

2.2 Classification and Regression Models for Physics Analysis

The task of reconstructing and classifying the outgoing particles after the collision is challenging. This is due both to the high energy and to the high particle multiplicity from the pileup. HEP experiments typically deploy selection algorithms that consist in a sequence of parameterized binary hypothesis tests. This approach is easily generalized to BDTs, the most common form of multivariate analysis techniques used in high energy physics such as b-tagging [13] and electron energy reconstruction [15].

One area that has synergy with previous works done at NCC is the identification of high-energy particle jets coming from the decay of heavy objects (W, Z, H bosons) into quarks. At high energies, those decays lead to collimated jets that appear overlaid in detector, making it hard to distinguish from single jets originated from pure quantum chromodynamics processes. Recently, a new approach [34] was proposed to address this problem that transforms the data retrieved from the detector into images, making it easier to deploy computer vision and machine learning techniques – an example is shown in Fig. 4. We intend to investigate the usage of modern machine learning techniques for performing the jet classification task. We will evaluate techniques that go beyond image classification, such as the Extreme Gradient Boost (XGBoost).

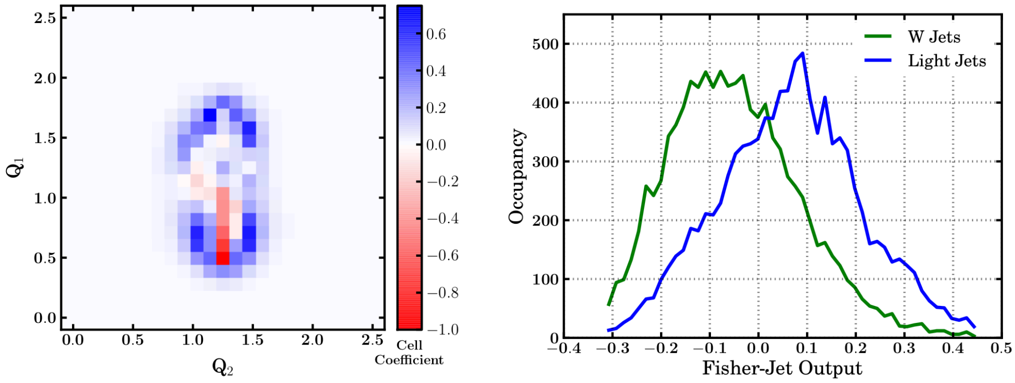
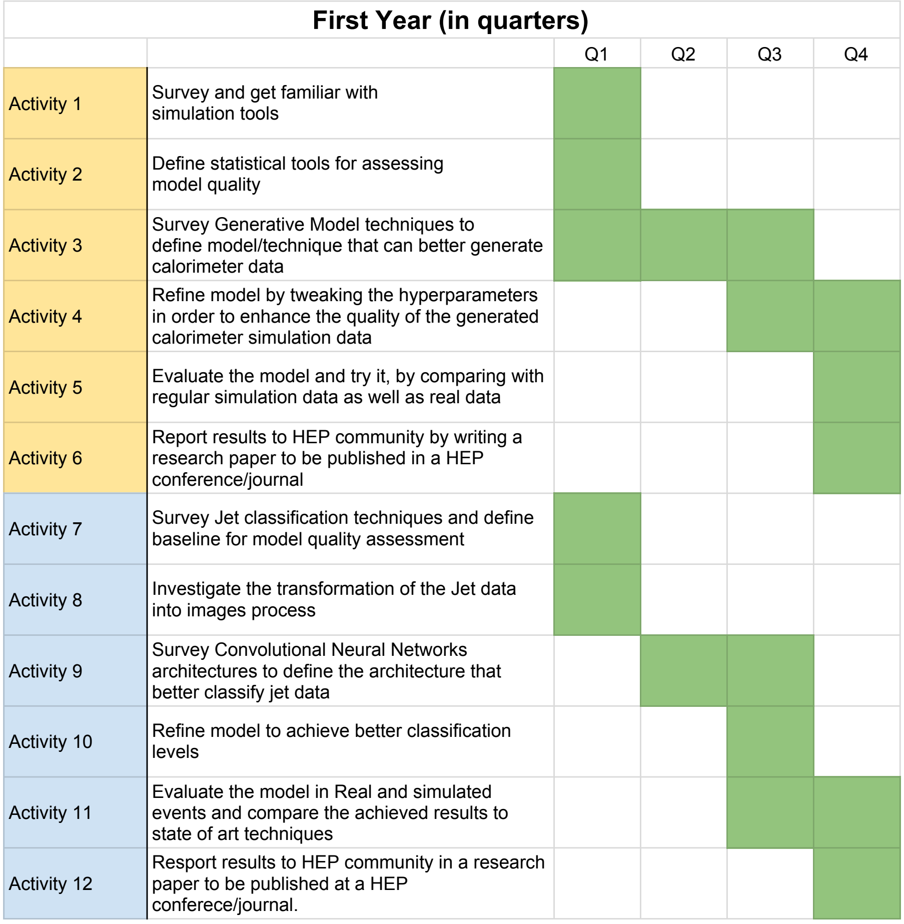


Figure 4: An application of ML techniques (computer vision) to a HEP problem: the usage of Fisher's linear discriminant (FLD)to discriminate between regular hadronic jets (light-jets) from a pp collision to those initiated by the hadronic decay of a highly boosted W boson (W-jets). Left: FLD presented as an image. Right: distributions of the discriminant output when applied to W-jets and Light-jets. Figure extracted from Ref. [34].

3 Schedule

The schedule for this project is divided into the first-year schedule and the one for the extension.

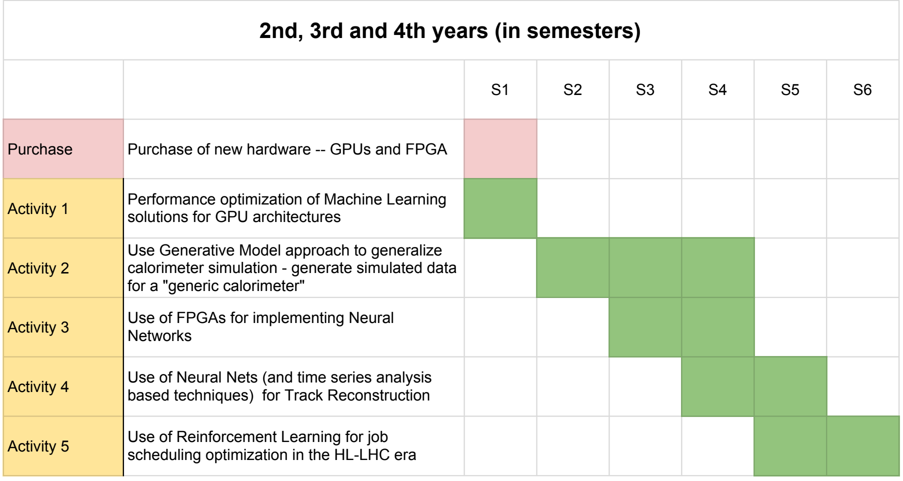
3.1 First-year schedule



First Year (in quarters)

		Q1	Q2	Q3	Q4
Activity 1	Survey and get familiar with simulation tools	🟩			
Activity 2	Define statistical tools for assessing model quality	🟩			
Activity 3	Survey Generative Model techniques to define model/technique that can better generate calorimeter data	🟩	🟩	🟩	
Activity 4	Refine model by tweaking the hyperparameters in order to enhance the quality of the generated calorimeter simulation data			🟩	🟩
Activity 5	Evaluate the model and try it, by comparing with regular simulation data as well as real data				🟩
Activity 6	Report results to HEP community by writing a research paper to be published in a HEP conference/journal				🟩
Activity 7	Survey Jet classification techniques and define baseline for model quality assessment	🟩			
Activity 8	Investigate the transformation of the Jet data into images process	🟩			
Activity 9	Survey Convolutional Neural Networks architectures to define the architecture that better classify jet data		🟩	🟩	
Activity 10	Refine model to achieve better classification levels			🟩	
Activity 11	Evaluate the model in Real and simulated events and compare the achieved results to state of art techniques			🟩	🟩
Activity 12	Resport results to HEP community in a research paper to be published at a HEP conferece/journal.				🟩

3.2 Extended schedule



2nd, 3rd and 4th years (in semesters)

		S1	S2	S3	S4	S5	S6
Purchase	Purchase of new hardware -- GPUs and FPGA	🟥					
Activity 1	Performance optimization of Machine Learning solutions for GPU architectures	🟩					
Activity 2	Use Generative Model approach to generalize calorimeter simulation - generate simulated data for a "generic calorimeter"		🟩	🟩	🟩		
Activity 3	Use of FPGAs for implementing Neural Networks			🟩	🟩		
Activity 4	Use of Neural Nets (and time series analysis based techniques) for Track Reconstruction				🟩	🟩	
Activity 5	Use of Reinforcement Learning for job scheduling optimization in the HL-LHC era					🟩	🟩

4 Budget

The budget for this project will be entirely spent in computing hardware and research personpower. We will hire at two levels: Early Stage Researchers (ESR) are in the first four years of their research and have not been awarded a Ph.D, while Experienced Researchers (ER) are in possession of a doctoral degree and have at least four years of full-time equivalent research experience. For the budgeting of computing hardware, we have used retail prices available as of this writing, considering the exchange rate of 1 USD = 4 BRL.

4.1 First-year budget

For the first year, as mentioned above we plan on using the computing resources already present at NCC-Unesp for our preliminary studies. Those include:

- the GridUnesp computing cluster, with 288 TB of storage and 3104 computing cores that are able to reach 77 TFlops;
- specialty hardware, like Intel® Xeon Phi™ manycore processors and NVIDIA GeFORCE RTX and TITAN V GPUs.

Therefore, the entirety of the first-year budget will be allocated in personpower. We will hire two team members, at the ESR level, for a period of twelve months. Table 1 summarizes the first-year budget.

Table 1: First-year budget.

Item	Monthly Cost (BRL)	Annual Cost (BRL)
ESR 1	4000	48000
ESR 2	4000	48000
Total		**96000**

4.2 Extended budget

For the extended period (second, third and fourth years), the two aforementioned researchers will be retained. However, at that point in the project lifetime there should be significantly better ML hardware available in the market[1]. Although there are turnkey GPU solutions in the market, they tend to be outside the budget possibilities of this project. We propose instead to buy discrete GPUs and install them in a regular rackmount server; we also propose buying FPGA cards for studying high-performance inference with neural networks. For the third and fourth year, having gone through the hardware upgrade, we go back to an "entirely personpower" budget allocation; three additional researchers will be hired, two at the ESR and one at the ER level. Additionally, 3.3% of the budget is allocated to academic overhead that may be used to cover conference expenses, hardware supplies, among others. Table 2 summarizes the extended budget.

[1] For reference, in 2018 NVIDIA launched their new Tesla V100 GPUs; when compared to their previous offering, the Tesla P100, the new hardware was 4 times more efficient for deep learning workflows.

Table 2: Extended budget for years 2–4.

Item	Monthly Cost (BRL)	Annual Cost (BRL)
4U server	—	40000
2020 NVIDIA GPU (4)	—	4×30000
FPGA (devel. kit)	—	10000
FPGA (deploy. version)	—	30000
ESR 1	4000	48000
ESR 2	4000	48000
Year 2 total		**296000**
ESR 1	4000	48000
ESR 2	4000	48000
ESR 3	4000	48000
ESR 4	4000	48000
ER	8000	96000
Overhead	—	10000
Year 3 total		**298000**
ESR 1	4000	48000
ESR 2	4000	48000
ESR 3	4000	48000
ESR 4	4000	48000
ER	8000	96000
Overhead	—	10000
Year 4 total		**298000**
Extended budget total		**892000**

5 References

[1] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", ***Phys.Lett.*** **B716** (2012) 1–29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[2] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", ***Phys. Lett.*** **B716** (2012) 30–61, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

[3] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", ***JINST*** **3** (2008) S08003, doi:10.1088/1748-0221/3/08/S08003.

[4] CMS Collaboration, "The CMS experiment at the CERN LHC", ***JINST*** **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[5] "LHC Report: Another run is over and LS2 has just begun…". https://home.cern/news/news/accelerators/lhc-report-another-run-over-and-ls2-has-just-begun.

[6] T. Radford, J. Gillies, C. Jakobsson et al., "Accelerating science and innovation: societal benefits of European research in Particle Physics", May, 2013. https://cds.cern.ch/record/1551933.

[7] Particle Physics Project Prioritization Panel (P5), "Building for Discovery: Strategic Plan for U.S. Particle Physics in the Global Context", May, 2014. https://science.energy.gov/~/media/hep/hepap/pdf/May-2014/FINAL_P5_Report_053014.pdf.

[8] G. Apollinari, I. Béjar Alonso, O. Brüning et al., "High-Luminosity Large Hadron Collider (HL-LHC) : Preliminary Design Report", 2015. doi:10.5170/CERN-2015-005.

[9] "ATLAS Experiment – Public Results: Event Displays from Upgrade Physics Simulated Data". https://twiki.cern.ch/twiki/bin/view/AtlasPublic/UpgradeEventDisplays.

[10] A. A. Alves et al., "A Roadmap for HEP Software and Computing R&D for the 2020s.", (Dec, 2017). HSF-CWP-2017-001.

[11] A. Radovic, M. Williams, D. Rousseau et al., "Machine learning at the energy and intensity frontiers of particle physics", ***Nature*** **560** (2018), no. 7716, 41–48, doi:10.1038/s41586-018-0361-2.

[12] "TMVA: The Toolkit for Multivariate Data Analysis". https://root.cern.ch/tmva.

[13] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", ***JINST*** **13** (2018), no. 05, P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[14] CMS Collaboration, "Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV", ***JINST*** **13** (2018), no. 10, P10005, doi:10.1088/1748-0221/13/10/P10005, arXiv:1809.02816.

[15] CMS Collaboration, "Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at √s = 8 TeV", ***JINST*** **10** (2015), no. 06, P06005, doi:10.1088/1748-0221/10/06/P06005, arXiv:1502.02701.

[16] CMS Collaboration, "Search for resonant $t\bar{t}$ production in proton-proton collisions at $\sqrt{s} = 13$ TeV", ***Submitted to: JHEP*** (2018) arXiv:1810.05905.

[17] CMS Collaboration, "Search for top squarks decaying via four-body or chargino-mediated modes in single-lepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV", ***JHEP*** **09** (2018) 065, doi:10.1007/JHEP09(2018)065, arXiv:1805.05784.

[18] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning". MIT Press, 2016. http://www.deeplearningbook.org.

[19] TLEP Design Study Working Group Collaboration, "First Look at the Physics Case of TLEP", ***JHEP*** **01** (2014) 164, doi:10.1007/JHEP01(2014)164, arXiv:1308.6176.

[20] M. Mangano, "Physics at the FCC-hh, a 100 TeV pp collider", doi:10.23731/CYRM-2017-003, arXiv:1710.06353.

[21] DUNE Collaboration, "Long-Baseline Neutrino Facility (LBNF) and Deep Underground Neutrino Experiment (DUNE)", arXiv:1512.06148.

[22] Hyper-Kamiokande Collaboration, "Hyper-Kamiokande Design Report", arXiv:1805.04163.

[23] F. Chollet et al., "Keras". https://keras.io, 2015.

[24] M. Abadi, A. Agarwal, P. Barham et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems". https://www.tensorflow.org, 2015.

[25] Center for Scientific Computing at UNESP, "Center of Excellence in Machine Learning". https://coe-ml.ncc.unesp.br/.

[26] GEANT4 Collaboration, "GEANT4: A Simulation toolkit", ***Nucl. Instrum. Meth.*** **A506** (2003) 250–303, doi:10.1016/S0168-9002(03)01368-8.

[27] ATLAS Collaboration, "The simulation principle and performance of the ATLAS fast calorimeter simulation FastCaloSim". https://cds.cern.ch/record/1300517, 2010.

[28] CMS Collaboration, "The fast simulation of the CMS detector at LHC", ***J. Phys. Conf. Ser.*** **331** (2011) 032049, doi:10.1088/1742-6596/331/3/032049.

[29] G. Amadio et al., "The GeantV project: preparing the future of simulation", ***J. Phys. Conf. Ser.*** **664** (2015), no. 7, 072006, doi:10.1088/1742-6596/664/7/072006.

[30] I. Goodfellow et al., "Generative Adversarial Nets", in ***Advances in Neural Information Processing Systems 27***, Z. Ghahramani, M. Welling, C. Cortes et al., eds., pp. 2672–2680. Curran Associates, Inc., 2014.

[31] M. Paganini, L. de Oliveira, and B. Nachman, "CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks", ***Phys. Rev.*** **D97** (2018), no. 1, 014021, doi:10.1103/PhysRevD.97.014021, arXiv:1712.10321.

[32] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks", in ***Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017***, pp. 214–223. 2017.

[33] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", ***CoRR*** **abs/1511.06434** (2015) arXiv:1511.06434.

[34] J. Cogan, M. Kagan, E. Strauss et al., "Jet-Images: Computer Vision Inspired Techniques for Jet Tagging", ***JHEP*** **02** (2015) 118, doi:10.1007/JHEP02(2015)118, arXiv:1407.5675.