EPJ Web of Conferences **364**, 01024 (2026) https://doi.org/10.1051/epjconf/202636401024
Quark Matter 2025

Machine learning as a technique for physics analysis

Hannah Bossi[1,*]

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

Abstract. High-energy experimental facilities such as the Relativistic Heavy Ion Collider (RHIC) and the Large Hadron Collider (LHC) are collecting more data and making more complex measurements than ever before. Machine learning has proven to be a valuable tool for these efforts that can be used throughout the pipeline from data collection to analysis. Such techniques will become necessary at future facilities such as the Electron Ion Collider (EIC) and the High Luminosity LHC (HL-LHC). These proceedings summarize a selection of recent developments on the use of machine learning as a technique for physics analysis and provide an outlook for future use.

1 What is Machine Learning? Why is it useful for physics?

The first electronic computers were developed in the 1940s with the purpose of performing tasks that were difficult and time consuming for humans to perform accurately. However, soon the need arose for computers to perform the opposite function; that is, for computers to do computations that humans are naturally good at, but on larger datasets. A majority of such computations rely on pattern recognition, a task the human brain is particularly skilled at. To fulfill this need, algorithms were created that mimic tasks that humans are skilled at by algorithmically replicating how humans learn. Such algorithms became the basis from which modern machine learning (ML), where algorithms improve over time, was born.

To mimic human learning, the fundamental building blocks of many modern day ML algorithms directly borrow the structure of neural networks inside the human brain. The networks inside the brain are composed of neurons connected by synapses. As one learns, the connection between some neurons become stronger and others become weaker. In ML, each piece has an analog, where neurons are replaced by nodes connected by weighted couplings that become either stronger or weaker as the network is trained, directly analogous to learning. The breakthrough that lead to the rise of modern artificial neural networks (ANNs) is the idea that collections of these node-like systems can give rise to computational collective behavior.

John Hopefield was the first to do this [1] with the so-called “Hopefield network", which is a fully-connected network where the inputs and outputs of the network come from all nodes. An improvement upon this idea was the Boltzmann machine [2] that further subdivided nodes into so-called visible nodes that are either input or output nodes (or both) and hidden nodes that are neither an input nor an output, but participate in the intermediate components of the network. These two foundational developments lead to the development of modern ANNs and was awarded the Nobel Prize in physics in 2024 where John Hopfield was awarded the

*e-mail: hannah.bossi@cern.ch

prize for the Hopfield network and Geoffrey Hinton for the Boltzmann machine [1]. This award serves as an acknowledgment of the vast impact that ML has had in academic spheres and in society over the last decades.

The primary goal of experimental measurements is to extract physics information from the available data. The traditional approach to do this is to make a selection using a series of boolean decisions motivated by physics or experimental constraints and then perform a statistical analysis on the selected data. However, as the datasets and observables of interest get more complex, the optimal decision becomes difficult to derive from expert knowledge alone. ML algorithms provide a natural solution by considering multiple variables simultaneously. In recent decades there has been a huge rise in the number of ML papers [2]. Many factors account for this rise, including an increase in computing power, collections of large datasets, as well as development of modern ML algorithms.

2 How is ML used as an analysis technique?

These proceedings are focused on how ML is used as an analysis technique, with special attention to its uses in heavy ion collisions. In a heavy-ion collision, the higher particle multiplicity leads to a higher complexity that adds difficulty for ML applications. In addition, there is a large variation amongst simulations, leading to a strong dependence on the simulation used in training. This can be be problematic for ML approaches applied to data, where this dependence must be accounted for with a systematic uncertainty. Despite these difficulties, ML has proven to be an incredibly useful tool for providing new insights into the physics of the quark gluon plasma (QGP). For a more complete review of QGP probes, see Ref. [3].

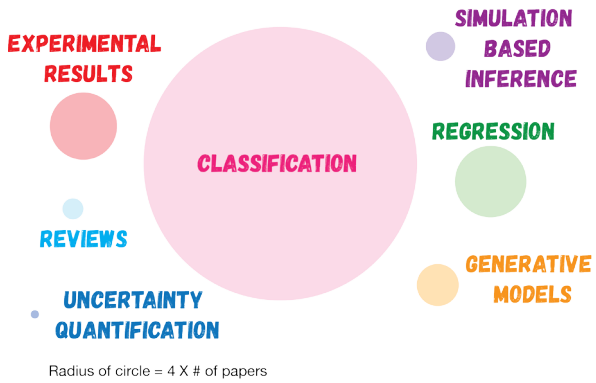

Figure 1. A visual summary of the relative distribution of the types of ML techniques employed in heavy-ion papers, where the radius of the circle corresponds to the number of papers multiplied by 4. Note that the papers in this scheme are allowed to count towards more than one category. Additionally, note that in the *experimental results* category, results using `ROOT TMVA` are included.

In Figure 1, a summary of ML techniques utilized for the analysis of heavy-ion collisions is presented. These results include papers that are related to the study of heavy-ion collisions or represent a technique that is readily applied to heavy-ion collisions. When compared to a similar summary of high energy physics as a whole, the relative distribution of topics remains the same, while the total number of papers is much less - reflective of the smaller

[1] See https://www.nobelprize.org/prizes/physics/2024/press-release/
[2] See https://iml-wg.github.io/HEPML-LivingReview/

community. In these proceedings, we will discuss a selection of these ML applications, grouped by technique. The remainder of this overview will be dedicated to the many ways in which ML can be used. However, also useful is understanding the ways that ML should *not* be used. Firstly, ML should not be used as a replacement for domain knowledge. That is, both the input features and training data must be carefully selected based on expert knowledge of the problem or context. Secondly, ML should never be used as a causation tool. When a ML algorithm is applied to input data, it can provide useful information on the correlation between observables. However, it cannot provide any information on why those correlations exist and should not be used as a means to justify those correlations. Successful applications of ML for physics use cases avoid these common pitfalls.

As shown in Figure 1, the application of ML to signal classification problems, where the algorithm seeks to identify one or more classes of signal candidates amongst the background, are the most numerous. In heavy-ion applications, signal classification is most commonly performed using boosted decision trees (BDTs) [4], where multiple weaker learners are combined in a series where each additional decision tree seeks to minimize the error of the previous network. These applications typically use the software package `XGBoost` [5] and in some cases the dedicated heavy-ion ML package `hipe4ml` [3]. One example relevant to these proceedings is a BDT implemented in `XGBoost` using `hipe4ml` to perform a preliminary first measurement by the ALICE experiment of Λ_c^+ elliptic flow, shown in Figure 2 [4].

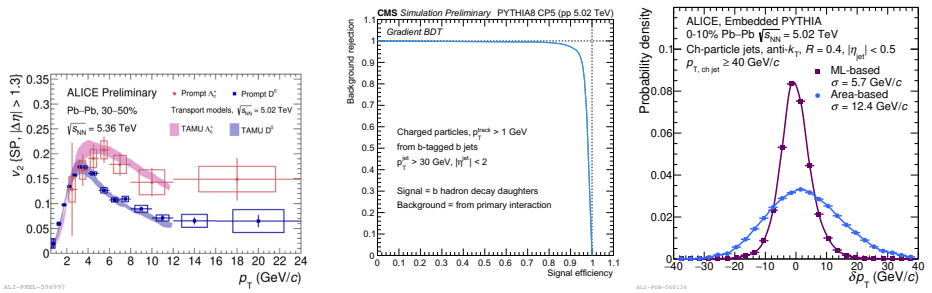

Figure 2. Left: Elliptic flow (v_2) as a function of p_T for prompt Λ_c^+ and D^0 compared to transport models. Middle: Performance of the gradient BDT to identify b hadron decay daughters via the background rejection as a function of the signal efficiency, originally appearing in Ref. [6]. Right: δp_T distributions defined as the difference between the reconstructed and true jet p_T for the ML-based and area-based background subtraction methods for $R = 0.4$ jets, originally appearing in Ref. [7].

Another common classification problem aims to select jets originating from a heavy quark (herein referred to as HF jets). The standard approach for tagging HF jets is to select jets with displaced decay vertices and including large impact parameter tracks, whereas the ML-based approach utilizes these low-level features in a supervised approach using BDTs or more complex algorithms like Graph NNs (GNNs). For example, the CMS collaboration recently used a gradient BDT to tag jets originating from a bottom quark (b-jets) [6]. Here, low-level parameters such as reconstructed tracks and the associated secondary vertex information are used to train the model. This method yields excellent performance quantified by the background rejection as a function of signal efficiency, shown in Figure 2. The gradient BDT was then utilized to perform the first measurement of the dead cone effect in b-jets, where vacuum emissions are suppressed within the jet inside a cone whose opening angle

[3] https://zenodo.org/records/7014886
[4] See talk by Chuntai Wu for more details

is related to $\sim m/E$. Another approach by the sPHENIX collaboration uses a long short-term memory network that allows information to be retained over long sequences [5]. This method demonstrated 2-3 times improvement over the traditional approach along with a 40-50% purity improvement, which will be further refined with additional parameter tuning.

Regression techniques have been widely applied to heavy-ion analysis over the last years. One approach utilizes a shallow NN trained in a supervised manner using PYTHIA embedded into a heavy-ion background to perform a regression of the true jet p_{T} [8]. The ALICE experiment applied this method, where the performance is given by δp_{T} distribution as shown in the right panel of Figure 2. Here, the ML-based method yields a significantly improved performance, which enabled a measurement of larger-radius $R = 0.6$ jets for the first time in ALICE [7]. Still under investigation is whether generative AI in the form of a so-called cycle-GAN [9] could be applied to the problem of the jet background subtraction. In this proposed procedure [6], two generator/discriminator pairs are used to perform and unpaired image-to-image translation between two domains while ensuring cyclic closure. Here, one domain would contain vacuum and background images (PYTHIA and HIJING separately) and the other domain would be a single combined image of the signal and background (PYTHIA and HIJING together). This preliminary work is potentially useful for jet background subtraction as the unpaired nature of the images makes this approach unsupervised, something that may reduce the dependence on the simulation used in training.

ML can also greatly improve unfolding procedures that seek to correct measured distributions for detector smearing or background contributions. Traditionally, unfolding is applied to a binned distribution and must be repeated for each observable. One ML-based solution to this problem is referred to as `OmniFold` [10]. Here, unfolding is implemented as a NN that calculates reweighting factors to translate between measured and true distributions. These reweighting factors are determined on an event-by-event basis before the choice of binning or the observable, unfolding the whole phase space all at once. This is a valuable tool, especially for high-dimensional analyses where unfolding can be difficult and time consuming.`OmniFold` has been applied to a number of different experimental applications [11–13]. This technique was originally developed for pp collisions and has recently been applied in heavy-ion collisions [14] to unfold PYTHIA and HERWIG samples embedded in a heavy-ion background. No explicit background correction procedure or fake correction was applied in this case; rather, these components were directly built into the unfolding procedure. The performance is found to be equivalent or better than iterative Bayesian unfolding, illustrating the potentially transformative nature of this technique in heavy-ion collisions.

In the near-term future, there are a number of open questions that should be addressed to improve and standardize the use of ML as a technique for physics analysis. Firstly, how should the systematic uncertainty be evaluated for ML-based approaches? This question is particularly relevant for heavy-ion collisions where a large uncertainty arises from the variation of potential models used in training. Secondly, how can more interpretable models be constructed? In physics approaches, interpretability is needed to guarantee that the ML is sensitive to physics information as opposed to transient information in the training data. Thirdly, how can reproducible ML-applications be reconstructed? Often ML applications are trained using the specific data format or detector simulations of an experiment, which then poses a problem for validation or cross checks of results. Finally, to what extent do ML-based methods need be standardized between experiments? Recent years have shown progress towards these aims, but community standards for ML-based approaches, particularly amongst the heavy-ion community, are largely lacking.

[5] See poster by Zhiwan Xu for more details.
[6] See poster by Yeonju Go for more details.

3 What does the future hold?

In the decades to come, new and improved facilities such as the high-luminosity LHC and the EIC will take large volumes of data at incredibly fast rates. Presently, data rates are already growing with the development of new streaming readout capabilities at experiments such as sPHENIX at RHIC, where data collected in the beginning of operations is already a large fraction of the total data stored from all experiments at Brookhaven National Lab's High Performance Storage System [7]. Given these dramatic increases, new ML techniques will be increasingly more important to fully exploit the large volumes of available data.

Dramatic recent improvements in the readout capabilities of the experiments have lead to a larger volume of data available for analysis. However, this poses an additional challenge due to limited storage and processing power, often preventing the storage of all raw data. To address this, trigger systems perform initial selection of events to be read out for further analysis using simplified cut-based algorithms on low-level detector outputs. However, as the data volume grows, it is desirable to have more complex trigger algorithms that more efficiently select desired events by using information from multiple variables simultaneously. One natural solution is to replace simple triggering algorithms with ML-based approaches. However, integrating ML-algorithms on the firmware via field programmable gating arrays while meeting resource and latency requirements can be challenging. In recent years there has been significant work to develop new software tools for efficient firmware implementations of ML algorithms such as `hls4ml` [15, 16]. A large number of experimental applications employ these techniques [17, 18] [8]. One additional use case of this technology is to perform a selection of rare decay topologies, such as heavy flavour decay topologies, to enhance statistics and/or simplify data processing procedures. Such capabilities are currently being demonstrated using the sPHENIX experiment as a test case [19], but may also be used at the future EIC. Additionally, due to the larger volume of data available, larger simulation samples are also needed, which can be expensive due to the detector simulation. Recent investigations show that denoising diffusion probabilistic models and general adversarial networks can be used to dramatically speed up full event simulations simulations at RHIC.

4 Summary and Conclusions

In summary, currently more data is being taken and more complex measurements are being performed than ever before. Machine learning has proven to be a useful tool throughout the whole analysis pipeline to lead to new physics insights in a variety of disciplines and contexts. These techniques are especially useful for physics analyses due to their ability to select on multiple variables simulataneously and learn from simulation. In these proceedings, various applications have been highlighted that aim to provide a representative summary of the ways in which ML is currently being used in heavy-ion collisions. In the coming years, ML tools will become necessary for both data-taking and analysis operations of experiments at future facilities such as the HL-LHC and EIC. As a result, the use of ML will undoubtedly maintain or exceed its rapidly growing popularity as a technique for novel physics insights.

References

[1] J.J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Nat. Acad. Sci. **79**, 2554 (1982). 10.1073/pnas.79.8.2554

[7] See https://www.bnl.gov/newsroom/news.php?a=122118 for more details.
[8] https://sse-ml-lhcb.gitlab.io/

[2] D.H. Ackley, G.E. Hinton, T.J. Sejnowski, A learning algorithm for boltzmann machines, Cogn. Sci. **9**, 147 (1985).

[3] J.W. Harris, B. Müller, "QGP Signatures" Revisited, Eur. Phys. J. C **84**, 247 (2024), 2308.05743. 10.1140/epjc/s10052-024-12533-y

[4] Y. Coadou, Boosted decision trees (2022), 2206.09645. 10.1142/9789811234033_0002

[5] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794

[6] Tech. rep., CERN, Geneva (2024), https://cds.cern.ch/record/2909071

[7] S. Acharya et al. (ALICE), Measurement of the radius dependence of charged-particle jet suppression in Pb–Pb collisions at sNN=5.02TeV, Phys. Lett. B **849**, 138412 (2024), 2303.00592. 10.1016/j.physletb.2023.138412

[8] R. Haake, C. Loizides, Machine Learning based jet momentum reconstruction in heavy-ion collisions, Phys. Rev. C **99**, 064904 (2019), 1810.06324. 10.1103/PhysRevC.99.064904

[9] D. Torbunov, Y. Huang, H. Yu, J. Huang, S. Yoo, M. Lin, B. Viren, Y. Ren, Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation (2022), 2203.02557, https://arxiv.org/abs/2203.02557

[10] A. Andreassen, P.T. Komiske, E.M. Metodiev, B. Nachman, J. Thaler, OmniFold: A Method to Simultaneously Unfold All Observables, Phys. Rev. Lett. **124**, 182001 (2020), 1911.09107. 10.1103/PhysRevLett.124.182001

[11] G. Aad et al. (ATLAS), Simultaneous Unbinned Differential Cross-Section Measurement of Twenty-Four Z+jets Kinematic Observables with the ATLAS Detector, Phys. Rev. Lett. **133**, 261803 (2024), 2405.20041. 10.1103/PhysRevLett.133.261803

[12] Y. Song (STAR), Measurement of CollinearDrop jet mass and its correlation with SoftDrop groomed jet substructure observables in $\sqrt{s}$ = 200 GeV pp collisions by STAR (2023), 2307.07718.

[13] T. Pani (STAR), Generalized angularities measurements from STAR at $\sqrt{}$SNN = 200 GeV, EPJ Web Conf. **296**, 11003 (2024), 2403.13921. 10.1051/epjconf/202429611003

[14] A. Falcão, A. Takacs, High-Dimensional Unfolding in Large Backgrounds (2025), 2507.06291.

[15] J. Duarte et al., Fast inference of deep neural networks in FPGAs for particle physics, JINST **13**, P07027 (2018), 1804.06913. 10.1088/1748-0221/13/07/P07027

[16] FastML Team, fastmachinelearning/hls4ml (2023), https://github.com/fastmachinelearning/hls4ml

[17] Tech. rep., CERN, Geneva (2020), final version, https://cds.cern.ch/record/2714892

[18] C. ATLAS, Tech. rep., CERN, Geneva (2022), https://cds.cern.ch/record/2802799

[19] J. Kvapil et al., Intelligent experiments through real-time AI: Fast Data Processing and Autonomous Detector Control for sPHENIX and future EIC detectors, PoS **ICHEP2024**, 1033 (2025), 2501.04845. 10.22323/1.476.1033