US CMS Software and Computing R&D Initiative

PI: Matt LeBlanc (Brown University)
PDRA: Lauren Hay (University at Buffalo → Brown University[1])

We propose a set of activities within CMS that aim to significantly reduce the tape and disk storage needs of the CMS Collaboration by incorporating our generic R&D and prototyping activities within the Experiment's workflows. Our research aims to **reduce tape and disk storage needs** by developing slimmer (compressed) data formats, and improving the efficiency of MC generators by mitigating negative event weights**.** The PI has recruited Dr. Lauren Hay as a new Postdoctoral Research Associate (PDRA), effective 1 January 2026, supported by an award from the DOE Computational High Energy Physics program (DE-SC0026285) for the theoretical development of these tools. **An award from this US CMS initiative will enable Lauren to dedicate 50% of her effort to CMS computing work, including the integration, validation and maintenance of these tools within the CMS software stack.** In addition to direct contributions, both the PDRA and PI will also supervise members of the PI's research group (including doctoral, master's and undergraduate students) in carrying out the proposed work, creating a pathway for graduate students to increase their future involvement in computing operations.

1. Importance of Project and Impact on HL-LHC S&C

Both the CMS and ATLAS Phase-2 computing roadmaps [3,4] indicate a risk of resource shortfall in terms of both computational and storage resources, and the LHCC has recently requested that the Collaborations assess the physics impact of future scenarios where MC production is reduced by up to a factor of 2. This motivates an aggressive research & development strategy empowered by Artificial Intelligence & Machine Learning (AI/ML) methods to reduce resource demand. **We propose two areas of activity that will reduce the footprint of CMS samples on disk & tape: (1) the mitigation of negative MC event weights and (2) studying the performance of lossy compression algorithms on CMS datasets.**

Nearly every analysis performed at the LHC relies on theoretical predictions, typically in the form of Monte Carlo (MC) event generators. The accuracy of these predictions is often the limit on the precision of an experimental result. Realizing the highest levels of precision requires adopting the most sophisticated predictions, which drastically increases computational and storage costs. Simulations of quantum chromodynamics (QCD) beyond leading order generate weighted events, including **negatively weighted events** that must be included to obtain a physical cross-section. These events degrade the sample's statistical power: a sample with 37% negative weights, typical of a MadGraph 5 sample generated at next-to-leading-order (NLO) in QCD, would require generating ~14 times more events to to reach the same statistical power as a sample with positive-definite weights [1,2].

Storage shortfalls will be exacerbated in-part by the development of AI/ML classifiers using low-level (particle/track-level) physics objects, which are data-intensive but also the most performant. Cutting-edge foundation models for particle physics have been trained using event

[1] Lauren will begin her position as a Postdoctoral Research Associate at Brown on 1 January, 2026.

samples with up to one billion jets [5], although no studies within experimental collaborations have been performed at this scale. **Beyond utility as a potential solution to the wider storage crisis, our group is interested in the performance of lossy compression algorithms on particle physics datasets, particularly those used to train AI/ML algorithms.** We believe that reliance on external groups to train foundation models poses a strategic risk to CMS. To maintain CMS's leadership in ML/AI applications for particle physics, the groundwork should be laid to facilitate the development of such models 'in-house'. Maintaining samples for training jet taggers at such large scales would be prohibitive in current CMS event formats, and so this initiative would also provide a near-term motivation for drastically reducing the per-event footprint of samples with low-level information, resulting in synergistic benefits that can be applied across the Experiment to reduce the footprint of samples on disk more widely.

2. Proposed Plan of Work

We will integrate generic algorithms that we are developing for **negative weight mitigation** and **lossy compression** directly into the CMS software ecosystem.

2.1 Negative weight mitigation

The PI's research group is developing efficient, Infrared- and Collinear-safe cell resampling algorithms that can be run alongside MC event generation, dynamically reweighting events using Optimal Transport and manifold learning techniques. This work solves several pathological behaviours with existing cell resampling algorithms in literature (e.g. [1,2]). The initial theoretical phase of this work is nearing completion. The PDRA, working with the PI (the L3 Physics Modelling & Validation Convener in the Physics Generator group), will incorporate this algorithm into CMSSW as an optional 'afterburner' algorithm that may be run following event generation (GEN). This will allow the cell resampling approach to be validated within a large experimental collaboration for the first time, using standard CMS MC validation workflows. A main focus of this work beyond validating the approach will be in improving the runtime efficiency of cell resampling algorithms to be tractable in a large-scale production environment. If successful, even partial mitigation of negative weights in CMS MC samples would result in a meaningful reduction in both computational and storage requirements.

2.2 Lossy compression

We will broadly investigate developments to reduce the on-disk footprint of CMS samples, beyond the scope of the generic R&D work on lossy compression algorithm development that is supported by the DOE Award. We propose a staged approach:

1. Optimization. In consultation with S&C and other CMS experts (*e.g.* the cross-POG officers), we will begin this portion of the project by surveying existing CMS samples to identify redundant and/or unused information can be reduced, and occurrences where the use of reduced-precision floating-point datatypes can reduce filesize without compromising physics performance. As NanoAOD is already extremely efficient and utilizes reduced floating point precision for certain quantities, we expect to focus our attention on MiniAOD and "nano-like" formats used for ML algorithm training (*e.g.* BTVNano) that include low-level information like tracks, secondary vertices, *etc.*

2. Compression. We will study the impact of applying algorithms (e.g. variational autoencoders) that trade bit-wise fidelity for a reduced sample footprint in a compressed format. Traditional approaches have been explored and deployed at the LHC [6,7], so-far focusing on derived quantities. In one previous result, the compression of pre-computed jet kinematics was studied [8,9], and in another entire events were compressed but only studied in aggregate, in the context of a dimuon resonance search [10]. In both cases, recovery of physics features of interest is generally achieved without significant distortion, suggesting that exploring compression techniques for CMS samples could provide an opportunity for savings.

The PI has overseen initial studies of autoencoder-based compression in the context of particle-level data representations for a master's student thesis, which indicated that lossy compression may degrade performance during complex tasks like jet tagging, which has also been shown to be sensitive to small perturbations ('adversarial attacks') [11]. The PDRA also has prior experience developing efficient data formats for specialized applications within CMS (*i.e.* JMENano). We therefore propose leveraging this joint experience with a staged approach, where compression is initially studied using samples with 'high-level' information and then with samples including 'low-level' information. **Our metric for performance is not only the compression ratio, but the degree of compression that can be achieved while maintaining indistinguishable physics performance on key benchmarks,** including tasks that are sensitive to multi-particle correlations like tagger development.

3. Timeline for Major Activities, Milestones & Deliverables

Upon the successful funding of this request (Proposed start date in **Spring 2026**), we plan to immediately begin to work with S&C and other experts within CMS to investigate existing NanoAOD and MiniAOD formats for redundant and unused information, and for additional opportunities where reduced-precision floating point datatypes can reduce the footprint of samples on-disk without compromising physics performance. This exercise will ensure that any studies of lossy compression performance on CMS datatypes will be performed on samples that have been optimized for any 'low-hanging fruit' before lossy algorithms are introduced.

We expect that studies of compression in the context of the CMS Experiment will be spread across the duration of this award, in times when the PDRA's focus is not expected to be on implementation of the negative weight mitigation algorithm. Initial work with existing algorithms can be performed immediately with existing NanoAOD samples. Depending on the outcome of these studies, the PDRA's focus will shift to either adjust the algorithms that are being studied (if they are found to be unsuitable for physics applications), or to then study the effect of lossy compression on low-level data representations.

We anticipate that the integration of an afterburner algorithm for cell resampling in CMS MC generation workflows will be performed over the course of **Summer 2026**, after R&D activities on the OT-based algorithm have concluded. Thorough validation of the algorithm and approach would then be performed during the **Fall and Winter of 2026/27**, through to the conclusion of this award.

Negative Weight Mitigation

- **Summer 2026:** The implementation of an optional afterburner algorithm in CMS MC generation workflows for event sample reweighting with OT-based cell resampling algorithms to mitigate negative MC weights.

- **Fall/Winter 2026/27:** The thorough validation of such an algorithm by the CMS Generators Physics Modelling and Validation subgroup, rendering it ready for study in the context of physics analysis.
    - The performance of this algorithm within CMS, once validated, will be documented in a DP note prepared within the Physics Generators Group at the end of this project.

Lossy compression

- **Spring 2026:** A survey of existing CMS sample formats for redundant and unused information, in close consultation with S&C.
- **Spring/Summer 2026:** Study of the impact of available lossy compression algorithms on 'high-level' physics samples (derived quantities such as the jet energy, etc.).
    - **Winter 26/27:** If promising, the development of an afterburner algorithm to optionally compress samples directly following MC generation, *e.g.* for use in AI/ML algorithm training.

4. PDRA's Past Accomplishments & Competency

The PDRA, Dr. Lauren Hay, is a recent graduate from the doctoral program at the University at Buffalo, where she studied under the supervision of Prof. Salvatore Rappoccio. **Over the course of her PhD, Lauren was the main analyzer of a precision differential cross-section measurement of the soft-drop jet mass in Drell-Yan events,** which offers new insights into quark- and gluon-initiated jet fragmentation modelling from the complementary flavour admixture in this process vs. the dijet channels where the soft-drop mass has previously been studied. **Lauren has served as a regular Detector On-Call (DOC) for the CMS Strip Tracking Detector** and developed detector control and safety systems panels for the Tracker.

In addition to her accomplishments in physics and operations, Lauren served as the CMS Particle Flow Validation and Reco Contact for an extended period, where **she optimized and improved reconstruction and validation software in CMSSW to monitor the impact of changes to core software on Particle Flow performance. Lauren served as the sole contact for multiple years, effectively managing a workload assigned to two people.** She has also held the role of JME Reconstruction and Algorithms Contact, **in which capacity she optimized and improved software used for jet reconstruction by the CMS Experiment, and additionally developed a new data format that is the *de facto* for jet calibration work** (JMENano). She has experience with AI/ML algorithm development outside of the CMS Collaboration, and was the main author on an early study of the application of explainable AI techniques to jet classification algorithms in 2020.

Beyond CMS, Lauren has been recognized by the Buffalo physics department as an outstanding teaching assistant (TA) with a departmental award for her efforts that included acting as the lead TA for their introductory physics sequence and mentoring new TAs. She has a track record of service to the wider community, and has participated in the HEP advocacy trip to Washington, DC and served in various roles including President, Vice-President and Secretary of the University at Buffalo Physics Graduate Student Association Executive Board.

5. Mentorship Plan

The PDRA will benefit from a holistic mentoring plan that is designed to foster their development as an independent scientific leader. Direct oversight will be provided through weekly one-on-one meetings with the PI to discuss the project's status, technical challenges, and recent developments in the field. They will be integrated directly within the PI's interdisciplinary research group, where students from the Brown University physic, computer science, engineering and math departments meet twice weekly for a journal club on topics related to QCD and AI/ML, and for a round-table research update meeting. The PDRA will be connected to nearby centers of expertise to facilitate interdisciplinary discussions that are pertinent to this work, including the Brown University Data Science Institute (DSI) and the NSF Institute for Artificial Intelligence and Fundamental Interactions (IAIFI), where the PI maintains affiliations.

Professional development will focus on both technical competency and career advancement. In addition to the expertise in the local CMS group, the PDRA will have access to regular technical training opportunities run through the Brown Center for Computing and Visualization (CCV) which manages the Ocean State Center for Advanced Research, Brown's local high-performance computing facility. Other aspects of mentorship will focus on the development of written and oral methods of scientific communication and eventual counseling on career paths in both academia and industry, leveraging the PI's growing experience with industry partnerships. The PDRA will have opportunities for involvement with outreach and advocacy activities, including through the US LHC Users' Association (of which the PI is an elected member of the Executive Committee), to build a practice of service to the field of particle physics alongside technical capability.

6. References

1. J. Andersen and A. Maier, *Unbiased elimination of negative weights in MC samples.* Eur.Phys.J.C 82 (2022) 5, 433. arXiv:2109.07851 [hep-ph].
2. J. Andersen *et al.*, *Efficient negative-weight elimination in large high-multiplicity Monte Carlo event samples.* Eur.Phys.J.C 83 (2023) 9, 835. arXiv:2303.15246 [hep-ph].
3. ATLAS, *ATLAS Software and Computing HL-LHC Roadmap*. CERN-LHCC-2022-005, 2022.
4. CMS, *CMS Phase-2 Computing Model: Update Document*. CMS-NOTE-2022-008, 2022.
5. W. Bhimji *et al., OmniLearned: A Foundation Model Framework for All Tasks Involving Jet Physics.* arXiv:2510.24066 [hep-ph].
6. A. Vestbo, *Pattern Recognition and Data Compression for the ALICE High Level Trigger.* PhD thesis, Universitetet i Bergen, 2004. arXiv:physics/0406003 [physics.ins-det].
7. C. Patauner, *Lossy and lossless data compression of data from high energy physics experiments.* PhD thesis, Tech. Univ. Graz, 2011. CERN-THESIS-2011-211.
8. F. Bengtsson *et al., Baler -- Machine Learning Based Compression of Scientific Data.* arXiv:2305.02283 [physics.comp-ph].
9. F. Bengtsson *et al., Baler - Machine Learning Based Compression of Scientific Data. EPJ Web Conf. 295 (2024) 09023.* Proceedings of the 26th International Conference on Computing in High Energy & Nuclear Physics (CHEP 2023).
10. J. H. Collins *et al. Machine-Learning Compression for Particle Physics Discoveries.* arXiv:2210.11489 [hep-ph], SLAC-PUB-17704.

11. B. Nachman and C. Shimmin, *AI Safety for High Energy Physics.* arXiv:1910.08606 [hep-ph].