Transformer-based End-to-End Reconstruction in the CMS Phase-2 High-Granularity Calorimeter

Lead-PI: Matteo Cremonesi (CMU)
Co-PI: Manfred Paulini (CMU)
Co-PI: Lindsey Gray (FNAL)
Research project term: July 1, 2026 - June 30, 2027
Salary support: \$45,000 (total), \$36,000 (salary) \$10,000 (fringe)
Travel funds: \$5,000

Contents

1 Proposed Research

The high-granularity calorimeter (HGCAL) [1] is the most important upgrade of the Compact Muon Solenoid (CMS) detector, designed to cope with the challenging conditions expected at the high luminosity (HL) LHC. The HGCAL will have approximately six million silicon sensors in the front section and four hundred thousand scintillator tiles in the rear. It will also have unprecedented transverse and longitudinal segmentation, providing measurements of energy deposited by particles interacting with its active material finely segmented in both space and time. At the same time, it poses an exceptional challenge for those algorithms that reconstruct particle features. The execution times of traditional pattern recognition algorithms scale exponentially with the growing granularity, and it is unclear whether they can satisfy computing constraints while providing sufficient high quality physics information that uses the improved detector technology. Furthermore, the level of pile-up (PU) expected at the HL-LHC will increase the likelihood of error in the reconstruction of particle properties, further degrading the performance of the feature extraction. Classical cluster reconstruction algorithms use a sequential approach. They start by identifying a set of seed hits and then proceed to associate the remaining hits iteratively. Not only their exponential scale with granularity and PU can lead to suboptimal usage of computing resources, but they also struggle to distinguish showers that overlap, like the ones that are expected in the crowded HL-LHC environment.

An alternative to sequential reconstruction is a machine learning (ML) approach that uses raw detector hit information to reconstruct particles in parallel, intrinsically accounting for energy deposits from overlapping showers. ML algorithms are highly parallelizable, exhibiting a significant gain in execution speed. Recently, collaborators from MIT and Fermilab, including scientist Lindsey Gray, started investigating transformer architectures to improve the HGCAL reconstruction performed online by the HGCAL backend system. In the context of this proposal, similar **transformer architectures will be studied for the offline HGCAL reconstruction**. This is a natural evolution of the prior award, which focused on the development of graph neural networks (GNNs) to accomplish the same task. The results and milestones of the prior award, together with a more general description of the approach, are presented in Sec. 2.

Transformers have revolutionized natural language processing and computer vision tasks, offering a scalable and efficient architecture for processing sequential data. Their usage in the context of experimental particle physics has become prominent in recent years, due to the gain in performance they provide [2]. The focus of this proposal is on MaskFormers [3]. They are transformer-based architectures originally developed for image object detection. When compared to the GNN approach developed in the context of the prior award, several improvements are expected from the usage of MaskFormers, justified by the following a priori arguments: (i) a transformer with full self-attention corresponds to a fully connected message-passing GNN, in which each node can attend to all others through adaptive, learned weights, thereby offering greater expressivity than the locally defined graph connectivity used in GNNs; (ii) Ref. [2] suggests that MaskFormers have a theoretical advantage for clustering tasks over models based on object condensation [4], the graph-segmentation technique employed in the GNN approach studied in the prior award (and further described in Sec. 2); (iii) the same paper also introduces a way to mitigate the quadratic complexity of transformers through masked attention, which had been the main limiting factor relative to GNN-based architectures; (iv) the estimation of cluster features could be enhanced relative to the purely-clustering GNN approach by directly regressing them from the transformer's latent representation of the detector hits. MaskFormers have already shown promise when used for complex reconstruction tasks in particle physics, such as particle tracking [2] and particle-flow [5]. In the context of this project, similar architectures as those used for these tasks will be adapted to the HGCAL use case.

Interface MaskFormers to the HGCAL data: As a first deliverable, Marchegiani will implement an interface between the HGCAL datasets, already used for the training of the GNN, and the MaskFormer model, using its publicly available code implementation [6]. In doing so, different approaches could be attempted to exploit the full capabilities of a transformer while considering the computational constraints given by the size of an event with the PU profile expected at the HL-LHC, where the total number of hits in the HGCAL could rise to more than $N = 100$k per event. Passing directly all detector hits to the transformer would require the computation of a self-attention matrix of dimension $N \times N$. Given the magnitude of N, this may prove prohibitive even with the use of masked attention. Instead, an approach similar to that followed in Ref. [5] could be attempted. Single calorimeter hits are combined into aggregated objects, defined as "topoclusters", which gather the information from neighboring hits by grouping them with certain geometric criteria. Topoclusters are then used as inputs to the MaskFormer, thus reducing the dimensionality of the attention matrix and allowing for a more efficient computation. Marchegiani will adapt the implementation of topoclusters developed for particle reconstruction at the Compact Linear Collider (CLIC) [7] to the HGCAL use case. This first milestone will be achieved before the end of Q1.

Perform training with increasing PU: Once a dedicated strategy for interfacing the HGCAL detector hits information with the MaskFormer model is determined, a full training on a 0 PU dataset can be performed to showcase the feasibility of the MaskFormer model for the reconstruction of energy clusters. The same dataset derived in the context of the prior award will be used. Similarly, profiting from the results of the prior award, the MaskFormer model will be trained with high-multiplicity simulated events including PU collisions. Following a progressive strategy, a first training with 30 PU can be performed. Marchegiani will evaluate potential criticalities caused by the increased PU conditions and study possible solutions and adaptations to the MaskFormer model. Once the model trained with PU is thoroughly tested and fully functional, a more realistic training can be performed on simulated events with 200 PU. Further adaptations to the model are foreseen in these extreme conditions. The number of hits that are input to the MaskFormer will be closely monitored to ensure that memory consumption and computation time fit the budget of the CMS offline processing system in Phase-2. This task will be delivered by the end of Q3.

Study physics performance metrics: In the context of the prior award, several performance metrics inspired by physics motivated observables have been developed for the benchmarking of the GNN-based model. The first metric to study is the energy resolution, which is a crucial property for assessing the precision of the HGCAL in measuring the energy of particles of various nature. Another important parameter to measure is the energy response, which is used to assess the accuracy of the energy measurement. The position and time resolution are also evaluated, since the precision of these quantities affect the overall quality of particle reconstruction, especially considering that the measurement of position and time in the HGCAL will eventually be combined with measurements performed by other sub-detector systems by the particle-flow algorithm. The performance of the MaskFormer model will be evaluated as a function of the number of PU interactions (0, 30, or 200 PU), to assess the potential degrade in resolution due to the presence of overlapping PU collisions. Marchegiani will use the same code infrastructure developed in the context of the prior award to estimate the aforementioned performance metrics by the end of Q4.

The timeline and milestones of the proposed research are summarized in Table 1.

2 Results from Prior Award

In the context of the prior award, Marchegiani led to completion several tasks to make GNNs a viable option for end-to-end HGCAL reconstruction. The GNN model is built with various "GravConv" [8] layers and it uses the object condensation segmentation technique. Its first version was originally trained using a dataset of simulated di-tau production events with 0 PU, but with outdated detector geometry. The first task that Marchegiani undertook was to reproduce the

Interface MaskFormers to the HGCAL data
Perform training with increasing PU
Training with 0 PU dataset
Training with 30 PU dataset
Training with full 200 PU dataset
Study physics performance metrics

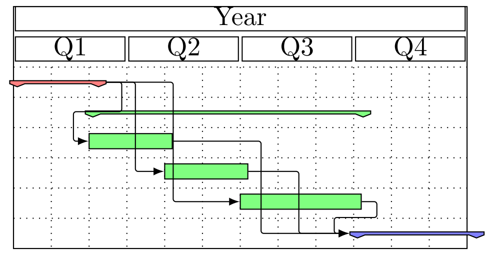

Table 1: Quarter by quarter outline of the proposal.

input data with updated detector geometry. This required to port the entire production pipeline to the most recent version of the CMS software (CMSSW). Most of the work was dedicated to adapting "FineCalo", a Geant4 patch that implements a more finely grained simulation of the calorimeter, to the new CMSSW version. Several changes have been made to the simulation of the HGCAL readout and geometry that required core modifications to FineCalo to make it compatible. The completion of this task allowed Marchegiani to generate massive training datasets spanning a variety of physics topologies, including the previously produced di-tau events, as well as events with single photons or pions. Then, Marchegiani successfully retrained the GNN with the new data.

To assess the performance of the GNN, Marchegiani developed several performance metrics based on physics motivated observables to evaluate the resolution and accuracy of the reconstruction. The cluster energy is reconstructed as the sum of the energy of the detector hits assigned to a given cluster. The energy resolution is extracted from the width of the corresponding energy distribution over thousands of single photons or pions events. Similarly, the position and time resolutions are evaluated considering the distribution of the difference between the predicted and true position and time of the cluster. All performance metrics considered meet the requirements outlined in the HGCAL TDR [1]. These results have been presented in many CMS internal meetings by Marchegiani and by the graduate and undergraduate students he is directly supervising. Among the most recent talks is the one presented by Marchegiani at the 2025 CMS ML Town Hall. The performance in 0 PU of the GNN trained with the latest geometry of the detector will be described in a paper currently under review by the CMS Machine Learning (MLG) group. A CADI line has already been assigned with the goal of publishing in time to present the results at CHEP 2026.

At the time of writing, Marchegiani's main focus is on the generation of a training dataset including PU collisions. Collaborators have previously attempted the generation of a dataset with PU with an old CMSSW version. Marchegiani has adapted the old code base to develop a new production pipeline that works with the latest CMSSW version. He successfully generated minimum-bias events added on top of the main physics process. The current challenge is to ensure that the FineCalo patch works with collisions with high PU. Marchegiani is expected to be able to produce datasets with 30 and 200 PU by the beginning of Spring 2026.

With 200 PU, the number of reconstructed hits in the HGCAL could reach values as high as 200k hits per event. In these conditions, the construction of the underlying graph of the GNN becomes a bottleneck, leading to saturation of the memory of the GPU used for the training. The graph construction is done via a k-nearest-neighbor (kNN) search in a learned latent space, an operation that is repeated at each GravConv layer. To mitigate this issue, Marchegiani worked with ML experts from KIT and CMU to develop a library in Python, called FastGraph [9], that leverages the features of modern GPUs to create graphs with a $40\times$ lower latency with respect to state-of-the-art industry tools, such as Facebook's FAISS library [10, 11]. Marchegiani will integrate the FastGraph library in the training pipeline of the GNN model to circumvent the memory limitation due to PU. He will then proceed to train the algorithm with 30 and 200 PU by the beginning of Summer 2026. This will leave enough time to estimate the physics performance of the new trainings by the end of the prior award (June 30, 2026).

3 Postdoc Accomplishments and Mentoring Plan

Matteo Marchegiani obtained his doctoral degree from ETH Zurich in May 2025 with a thesis on the search for a Higgs boson in association with a top quark-antiquark pair and on the calibration of large-radius jet flavor taggers. He conducted his graduate research as a member of the CMS Collaboration, contributing to the analysis of data recorded during Run 2. His work laid the foundations for the future analysis of Run 3 data. His substantial contributions include the first application of modern transformer architectures to event reconstruction and classification in the $\mathrm{t\bar{t}H}\left(\mathrm{H} \to \mathrm{b\bar{b}}\right)$ analysis, along with a sophisticated background-estimation strategy employing ML techniques to improve the modeling of the irreducible $\mathrm{t\bar{t}} + \mathrm{b\bar{b}}$ background. These tools allowed him to define a new analysis strategy that could achieve an expected significance of 4σ by the end of Run 3, considering only the semileptonic channel. In addition to studying the Higgs boson, Marchegiani has developed a new method to calibrate large-radius jet flavor taggers by considering jets enriched with muons as a proxy for signal jets. Such measurement is of crucial importance for several analyses performed by the CMS Collaboration, including searches for new particles as well as precision Standard Model measurements, which exploit ML based jet flavor taggers to enhance the signal sensitivity. When applied to real data, such algorithms require calibration in order to correct for the difference in tagging efficiency between data and simulation. His work was documented in a dedicated publication of the CMS Collaboration [12] and is the main reference for the calibration of large-radius jet flavor taggers in CMS analyses. As a recognition of his contributions to the developments of calibration methods, Marchegiani was recently nominated Level-3 convener of the Performance and Calibration subgroup within the b-tagging and vertexing (BTV) Physics Object Group (POG) of CMS, showing his ability to take leadership roles within a wide collaboration. Finally, Marchegiani is the main developer of a new configurable analysis framework for fast and efficient analysis based on the the Columnar Object Framework For Effective Analysis (COFFEA) package, called PocketCoffea. This open-source software is endorsed by the Common Analysis Tools (CAT) group of CMS and has been adopted by hundreds of collaborators within CMS, showcasing his ability in managing and maintaining a large-scale project requiring highly technical software engineering skills. During the first half of the prior award period, Marchegiani demonstrated exceptional commitment and initiative in advancing the HGCAL reconstruction project which demanded a comprehensive grasp of the underlying physics as well as sophisticated expertise in software design and ML. Marchegiani is also the leader of ML4RECO, the group within CMS that focuses on ML approaches to particle reconstruction, including the offline HGCAL reconstruction which is the object of this proposal and of the prior award.

CMU is one of the only three HGCAL module assembly centers in the US, responsible for producing about 10k standard modules for the hadronic section of the HGCAL. The local CMS group built at CMU a class 1000 cleanroom and it commissioned all the major pieces of equipment needed for the module production. This provides Marchegiani with the unique opportunity to connect the work done on the reconstruction software to the hardware components of the HGCAL. Furthermore, CMU ranks top 1 in the world in Artificial Intelligence and direct collaboration with members of its School of Computer Science can have a major impact on an AI-based project like the one described in this proposal. Marchegiani has already been supervising undergraduate students from the Machine Learning Department on topics related to the prior award.

Lead-PI Matteo Cremonesi has given significant contributions to CMS computing, such as leading the Computing Operation Production & Reprocessing Group as well as co-founding the COFFEA project. In recent years he has been focusing predominantly on the HGCAL reconstruction, both offline, as the lead-PI of the prior award, and online, promoting the adoption of ML approaches for the trigger primitive generation performed by the HGCAL backend system. His leadership in the area of online HGCAL reconstruction was recognized with an NSF "Accelerating Computing-Enabled Scientific Discovery" (ACED) grant.

Co-PI Manfred Paulini is a senior member of the CMU faculty as well as the leader of the CMS group at CMU. He is one of the leading HGCAL experts in the US-CMS HL-LHC upgrade project, where he serves as Level-2 Manager for HGCAL detector operations. He has also been leading an effort that use ML for particle reconstruction in the electromagnetic calorimeter (ECAL), developing expertise which is directly transferable to the goals of the project described in this proposal.

Co-PI Lindsey Gray has been one of the leaders of the HGCAL reconstruction project. Together with Jan Kieseler, he conducted the development of the first GNN for HGCAL reconstruction, which was the stepping stone toward the research supported by the prior award. More recently he has shifted his focus on online HGCAL reconstruction, where he has been studying the applicability of MaskFormers. His direct experience with the same class of models that will be used for the research described in this proposal is an amazing resource. Working in tight collaboration with him will ensure that Marchegiani can confidently achieve the goals described above.

Furthermore Cremonesi and Gray have a long history of collaborating on a multitude of computing projects. For example, Gray was the other co-founder of COFFEA and the current Project Manager. They also started a new collaboration on online HGCAL reconstruction, which is more directly relevant to this proposal. This will also create the opportunity to leverage synergies between offline and online reconstruction, which are historically separated but that, in an era when AI approaches on hardware are becoming increasingly feasible, are getting closer. Marchegiani will certainly benefit from this well established collaboration that has always resulted in the creation of a lively group dynamic.

The HGCAL reconstruction is reviewed by the CMS ML4RECO group. Marchegiani took over the leadership of the group since he joined CMU, making sure that his own activities in the context of the prior award are properly integrated in the larger ML4RECO community. Besides taking part in official CMS meetings, Marchegiani will meet with the PIs at least on a weekly basis. In these meetings the discussion will also be focused on career planning. At the same time, the PIs will work with Marchegiani in identifying key conferences and meetings where to participate and ensure he receives excellent visibility both locally and across the scientific community, in the US and internationally, which is key to building a successful academic career.

4 Conclusions

The HGCAL is the single most important project for the HL-LHC CMS detector upgrade. The ability to reconstruct particles inside the HGCAL with high precision is key to fully take advantage of the capabilities of the new detector and have an impact on the physics reach. At the same time, with the demanding HL-LHC PU conditions it will be challenging for particle-shower reconstruction algorithms to satisfy computing constraints while exploiting the full physics potential of the improved detector technology. Reconstructing particles with ML can provide the ideal solution. In this proposal, support for 50% of the remuneration of a postdoc is asked in order to develop MaskFormers for the HGCAL reconstruction that work with 200 PU. Postdoc Marchegiani will build on top of the work already done in this direction in the context of the prior award. He will also rely on the guidance of the PIs, who are all experts in ML for particle reconstruction (including HGCAL reconstruction), and on the resources offered by CMU, which is a HGCAL module assembly center and ranks among the best US schools in ML/AI.

References

[1] “The Phase-2 Upgrade of the CMS Endcap Calorimeter,” CERN, Geneva, Tech. Rep., 2017. [Online]. Available: https://cds.cern.ch/record/2293646

[2] S. V. Stroud, P. Duckett, M. Hart, N. Pond, S. Rettie, G. Facini, and T. Scanlon, “Transformers for charged particle track reconstruction in high energy physics,” 2024. [Online]. Available: https://arxiv.org/abs/2411.07149

[3] B. Cheng, A. G. Schwing, and A. Kirillov, “Per-pixel classification is not all you need for semantic segmentation,” 2021. [Online]. Available: https://arxiv.org/abs/2107.06278

[4] K. Lieret, G. DeZoort, D. Chatterjee, J. Park, S. Miao, and P. Li, “High pileup particle tracking with object condensation,” 2023. [Online]. Available: https://arxiv.org/abs/2312.03823

[5] N. Kakati, E. Dreyer, A. Ivina, F. A. D. Bello, L. Heinrich, M. Kado, and E. Gross, “Hgpflow: Extending hypergraph particle flow to collider event reconstruction,” 2025. [Online]. Available: https://arxiv.org/abs/2410.23236

[6] samvanstroud, “hepattn: Unified end-to-end ml reconstruction for particle physics,” GitHub repository, 2025, commit corresponding to version used. [Online]. Available: https://github.com/samvanstroud/hepattn

[7] cocoa hep, “Cocoa: Configurable calorimeter simulation for ai,” GitHub repository, 2025, commit corresponding to version used. [Online]. Available: https://github.com/cocoa-hep/cocoa-hep

[8] S. R. Qasim, K. Long, J. Kieseler, M. Pierini, and R. Nawaz, “Multi-particle reconstruction in the High Granularity Calorimeter using object condensation and graph neural networks,” *EPJ Web Conf.*, vol. 251, p. 03072, 2021.

[9] A. Agarwal, R. He, J. Kieseler, M. Cremonesi, and S. R. Qasim, “Fastgraph: Optimized gpu-enabled algorithms for fast graph building and message passing,” 2025. [Online]. Available: https://arxiv.org/abs/2511.10442

[10] J. Johnson, M. Douze, and H. JÃ©gou, “Billion-scale similarity search with gpus,” 2017. [Online]. Available: https://arxiv.org/abs/1702.08734

[11] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. MazarÃ©, M. Lomeli, L. Hosseini, and H. JÃ©gou, “The faiss library,” 2025. [Online]. Available: https://arxiv.org/abs/2401.08281

[12] C. collaboration, “Performance of heavy-flavour jet identification in lorentz-boosted topologies in proton-proton collisions at $\sqrt{s} = 13$ TeV,” *Journal of Instrumentation*, vol. 20, no. 11, p. P11006, nov 2025. [Online]. Available: https://doi.org/10.1088/1748-0221/20/11/P11006