Accelerating Machine Learning Reconstruction in CMS

PI: Prof. Sergei Gleyzer, University of Alabama

1 Introduction

The CMS experiment [1] at the Large Hadron Collider (LHC) is exploring the frontier of particle physics with highest energy proton-proton collisions ever recorded in the laboratory. The University of Alabama (UA) has been a member institution of the CMS Collaboration since 2011. The UA CMS group is led by faculty members: Sergei Gleyzer and Paolo Rumerio. In 2021, Professor Gleyzer supervised a postdoctoral researcher, Dr. Davide di Croce, on a US-CMS HL-LHC Software and Computing R&D project titled: *"Accelerating Deep Learning Reconstruction for CMS"*. The focus of this project was on integrating innovative machine learning algorithms into detector reconstruction, while leveraging heterogeneous computing architectures. As Dr. DiCroce recently accepted a more senior position, Professor Gleyzer is currently filling a new position for a postdoctoral researcher with the anticipated starting date of June 2022. The focus of this new position will be on machine learning applications for physics analysis, detector reconstruction and Phase-II upgrades, with a strong emphasis on computational and innovative machine learning aspects of this research. The candidate, supervised by Dr. Gleyzer, will be based at the LPC and make significant contributions to US CMS machine learning operations.

The proposed research program for the postdoctoral researcher, focused on integration of machine learning and hardware acceleration into CMS reconstruction, is closely aligned with the vision and priorities of the CMS Machine Learning group. It closely ties to US-CMS Software and Computing Operations and R&D efforts in the area of heterogeneous computing resources, detector reconstruction and analysis facilities and tools. The PI will enable additional support for the proposed research program at the LPC, through collaboration with existing synergistic LPC machine learning efforts and with the broader CMS Machine Learning community. In this proposal, we extend the original project to additional machine learning reconstruction targets, including applications of **graph neural networks** and **vision transformers** for *end-to-end tau* and *electron reconstruction*, *merged photon* and *tau mass regression*, and further *integration of machine learning* and *hardware acceleration* into CMSSW using containerization. Detailed research plans are described in Sections §2 - §4. Sections §2.1 - §3 describe the results obtained in 2021 and the planned timeline and deliverables are discussed in Sections §4.

2 Machine Learning for New Physics at the LHC and HL-LHC

Our ability to take full advantage of machine learning during HL-LHC crucially depends on **deep integration** of information from various available sources and implementation of powerful algorithms that best exploit *highly-dimensional* and *highly-granular* data. Despite rapid advances, significant challenges still remain for large-scale application of **deep-learning algorithms** beyond the present state of the art. These challenges include model complexity and latency at inference time, **integration of deep learning algorithms in detector reconstruction** and simulation, the *non-euclidean nature* of some input data and *irregular geometry* posing challenges to direct application of computer vision-inspired algorithms to full detector data. These challenges and the program to address them are described in detail in the Community White Paper for Machine Learning in High-Energy Physics [2], co-edited by **PI Gleyzer**.

The LHC will enter its 'High-Luminosity' (HL-LHC) phase in 2029. The center-of-mass energy will be 14 TeV and instantaneous luminosity is projected to reach a peak of up to 7.5×10^{34} cm^{-2}s^{-1}, a factor of four increase beyond Run 2. The goal of HL-LHC is to collect an integrated luminosity of at least 3000 fb^{-1} in ten years of operations [3]. At HL-LHC, the number of pileup collisions is expected to reach 200, a significant increase above the Run 2 mean of ~ 40. To handle this challenge, and to cope with the high radiation environment, the CMS experiment is preparing major detector upgrades for HL-LHC [4]. One of the key HL-LHC challenges is how to best take advantage of these new detectors and additional advantages offered by their expanded *granularity*, *sensitivity* and *timing resolution* to enhance the physics reach and potential of the CMS experiment.

2.1 End-to-end Deep Learning for LHC and HL-LHC

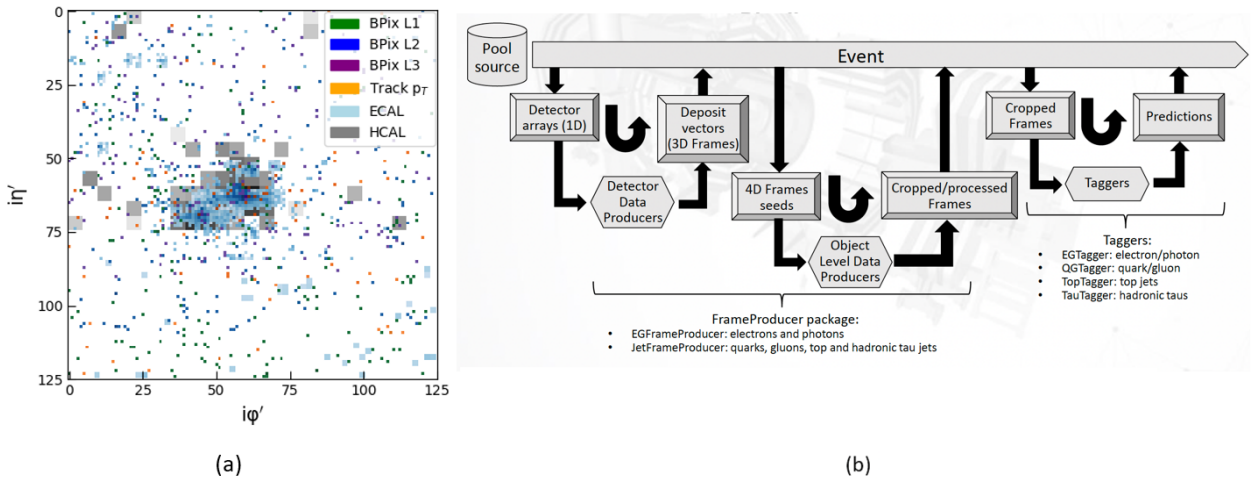
For the past decade, the CMS experiment uses **particle flow** (PF) algorithms to reconstruct the LHC collision events [5]. The CMS PF algorithm attempts to *reconstruct all stable particles* in an event including muons, electrons, photons and hadrons and uses that information to build jets. The algorithm **combines information from all CMS subdetectors** to construct PF candidates from overlapping tracks and calorimeter clusters. Currently, *particle flow* candidates are used as input for many **machine learning applications** at the LHC [6–8]. Alternatively, and crucially for future progress, *machine learning algorithms* can be **directly applied to *low-level data*** to improve the performance of particle flow algorithms, leading to a possible next-generation LHC and HL-LHC reconstruction, enhanced by machine learning techniques.

PI Gleyzer leads the development of a novel deep learning reconstruction technique suitable for LHC and HL-LHC - **end-to-end deep learning** (E2E) [9–17]. *End-to-end deep learning* leverages a combination of *deep learning* and *low-level detector representation* to efficiently identify particles and perform event-level reconstruction. This approach has achieved current *state-of-the-art performance* in identifying *electrons*, *photons*, *jets* and *boosted objects* [9–12] and has been effectively applied in physics analysis [13,14]. End-to-end deep learning takes advantage of various deep learning architectures including *convolutional neural networks* such as ResNet [18], and *graph neural networks* that show excellent *scalability* and *generalization* to particle identification tasks [9, 20–22, 27]. Recently, PI Gleyzer has developed **graph neural network** applications for CMS Phase 2 reconstruction and *end-to-end* deep learning reconstruction and simulation [20–22]. Additional end-to-end developments include incorporation of *low-level tracking* information for improved jet and boosted jet identification, *muon* and *tau identification*, *mass regression* and *event-level anomaly detection*. Beyond standalone particle and event identification tasks, end-to-end deep learning forms the basis of **the first end-to-end deep learning-based CMS search** for *light pseudoscalar decays* of the Higgs boson into four photons [13], a physics analysis that demonstrates the ability of *innovative end-to-end deep learning* algorithms to push the experimental sensitivity in **challenging reconstruction** conditions such as fully-merged particles. This innovative physics result based on the E2E deep learning technique, including a novel application of end-to-end mass regression [14] paves the way for a whole family of analyses that can take full advantage of *end-to-end deep learning* in areas where traditional reconstruction is limited. Machine learning algorithms developed as part of the E2E project have additional applications to novel physics searches and advanced data quality monitoring applications, that the E2E group plans to expand on during 2022.

The **end-to-end** deep learning project has evolved into a major (US) CMS machine learning effort led by University of Alabama, with participating US institutions: University of Alabama, Carnegie Mellon and Brown. **PI Gleyzer leads the overall E2E project**, its *algorithmic development* and *applications to reconstruction and physics analysis*, including e/γ and tau identification, anomaly detection, trigger applications and BSM physics searches in boosted topologies.

3 Accelerating Deep Learning Reconstruction in CMS

One of the key steps in **integrating machine learning algorithms** into CMS reconstruction is *integration of promising particle identification* algorithms such as end-to-end deep learning into the CMS Software Framework. The UA group has focused on the integration of the end-to-end framework into CMSSW and the **inference and scaling performance** of the end-to-end algorithm described in Section §2.1. One important aspect of this work is the *understanding and optimization* of **inference time** of the E2E algorithm and its scaling with the use of heterogeneous computing and *hardware acceleration devices*, such as GPUs and FPGAs.



(a) (b)

Figure 1: (a) A 6-channel composited end-to-end image of a simulated top quark jet. (b) Outline of the E2E pipeline for classification

During the previous US-CMS SC *R&D* project, three distinct E2E benchmarks were developed: e/*gamma* classifier using information from the ECAL subdetector, quark-gluon classifier that relies on ECAL, HCAL and track p_T information, and a top jet classifier that uses hits from barrel pixel layers, track p_T, ECAL, and HCAL information (Figure 1a). All three end-to-end classifiers use ResNet CNNs. In another phase of the project, we developed the first **end-to-end tau classification** benchmark and an **end-to-end graph neural network classifier** for top and tau jets (Figure 2).

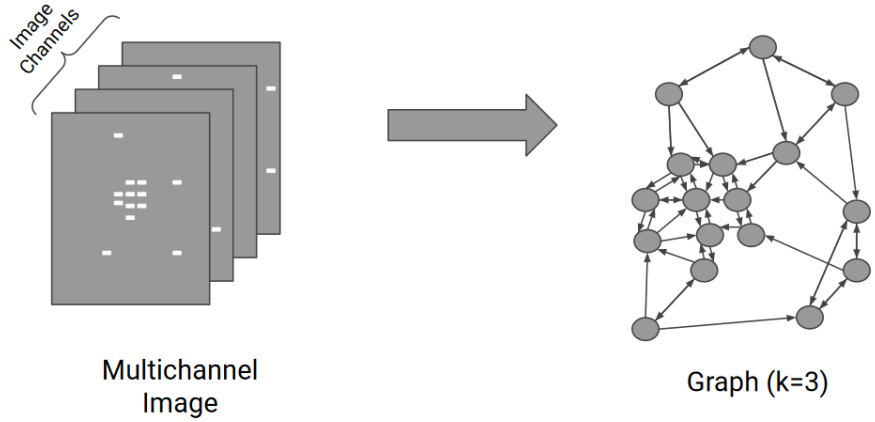


Figure 2: E2E Graph Neural Network

Table 1: E2E Tau Classification with Convolutional and Graph Neural Networks

E2E Layer Combination	AUC(CNN)	AUC(GraphNN)
Track Pt+d0+dz	0.776	0.864
Track Pt+d0+dz+ECAL+HCAL	0.798	0.870
Track Pt+d0+dz+ECAL+HCAL+BPIX+TIB/TOB1-2	0.864	0.887

As can be seen in Table 1, newly developed E2E graph neural network architectures perform very well on the tau classification task for all layer combinations, and show improvements with inclusion of additional layer information. We plan to add more input information and enhance and further customize the models for the low momentum regime.

In another part of the project, we integrated the E2E framework on top of CMSSW base classes, with a highly modular design to support customizable E2E workflows to support ML training and inference [15]. The current version can handle *fully connected*, *convolutional* and *graph neural network* architectures, while future versions will use other models as well. The framework consists of three main package categories: DataFormats, FrameProducers, and Taggers, as shown in Figure 1b. The DataFormats package consists of all the objects and classes needed for running the E2E modules and for storing any output back into EDM-format files.

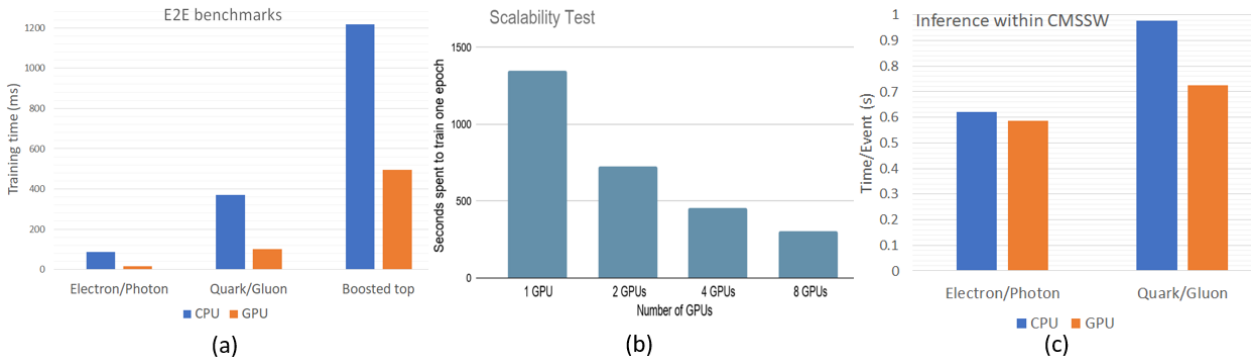


Figure 3: (a) Comparison of training time for different E2E benchmarks. (b) Scaling multi-GPU training for the standalone boosted top jet benchmark. (c) Comparison of the inference time for electron/photon and quark/gluon benchmarks using CPU and GPU architectures within CMSSW.

We also trained the E2E benchmarks on CPUs and GPUs, as shown in Figure 3a. We leveraged the Nvidia Horovod framework to train E2E benchmarks on multiple GPUs and studied its scaling performance. With this procedure different layers of the deep learning model are trained on different devices, taking advantage of the inter-GPU and inter-node communication methods such as NCCL and MPI to distribute the deep learning model parameters between various workers and aggregate them. Our results show good scaling performance (Fig. 3b.). Finally, with the CMSSW TensorFlow C++ API, we performed inference on CPU and GPU architectures within CMSSW (Figure 3.c).

The end-to-end deep learning application is an excellent test case for machine learning integration as it uses state-of-the-art machine learning algorithms and significant input space and data complexity. The postdoctoral associate can additionally focus on the graph neural network HGCAL reconstruction algorithm developed by the PI. This prioritization can be discussed before the project begins. Longer-term vision includes integrating end-to-end graph neural networks and transformer architectures for HGCAL reconstruction, due to their natural ability to incorporate irregular HGCAL geometry and timing inputs for a deep spatio-temporal combination with the eye towards addressing Phase-2 HL-LHC reconstruction challenges.

4 Research Agenda: Timeline and Milestones

Leveraging the state-of-the-art jet and tau identification results achieved by the *end-to-end* deep learning algorithm that exploits low-level tracking [10–12], we propose to further extend and customize the E2E algorithm to **electron and tau identification for the full momentum range** in 2022. The goal of this work is to **significantly improve identification of electrons and taus** in the low momentum regime. Present deep learning techniques used for tau identification are used more traditionally with combinations of hand-selected high-level features and particle flow inputs. The *end-to-end* approach uses full detector information and deep learning algorithms applied at a much earlier reconstruction stage, and can potentially recover information losses in tau reconstruction. This advantage should be even more apparent in the low-momentum regime.

- **Specific Reconstruction Aim 1**. Extend the **end-to-end deep learning** reconstruction to electron and tau identification in 2022 with a focus on the low momentum regime. Integrate the electron and tau benchmarks into CMSSW E2E prototype and perform scaling studies for training and inference on heterogeneous hardware platforms relying on **containerization**.

Graph neural networks provide a **natural representation** for encoding *relational information* of physical systems. Proposed by Scarcelli [23] and further developed to learn across *graph nodes* and *edges* [24], *graph representation learning* can handle irregular grids with non-Euclidean geometry [25], encode physics knowledge via graph construction [26] and introduce *relational inductive bias* into data-driven learning systems [27]. **Graph representation learning** has shown early promise for LHC applications and elsewhere [7, 19, 28–30]. As Table 1 shows, progress obtained with graph models highlights the potential of **end-to-end deep learning algorithms with graph neural networks** for improving LHC reconstruction with upgraded CMS detectors. Additionally, **PI Gleyzer** has recently developed end-to-end graph generative and attention models to successfully learn end-to-end detector representation with graph models [19–22]. We propose to additionally develop transformer-based models for computer vision (*vision transformers* [31]) and *graph transformer* [32] architectures and evaluate their performance on newly established benchmarks.

- **Specific Reconstruction Aim 2**. Further extend the **end-to-end deep learning** tau reconstruction with graph neural networks and vision and graph transformers for combination of low-level tracker and calorimeter inputs and integrate them into CMSSW. Additionally, extend to E2E regression tasks. Perform inference and training scaling tests with heterogeneous hardware relying on **containerization**.

The *end-to-end deep learning* application is an excellent test case for machine learning integration as it uses state-of-the-art machine learning algorithms and significant input space and data complexity. We will continue to build upon E2E CMSSW integration in 2022 by incorporating new **reconstruction targets including mass regression**, as in recent E2E analyses [13, 14] and make it flexible enough to handle **graph inputs** at training and inference stage. Tau reconstruction offers an excellent benchmark to study both training and inference of deep learning models within CMSSW. We will leverage **containerization** as we work towards CMSSW GPU inference and continue expanding deep learning with E2E benchmarks on heterogeneous hardware, with a greater focus on inference and additional hardware architectures (FPGAs). As part of project we will also use and evaluate the new Fermilab *elastic analysis cluster facility*. This project will provide useful feedback and a realistic test of the cluster capabilities and limitations for CMS ML needs. Longer-term vision includes integrating *end-to-end graph neural networks* for HGCAL tau reconstruction, due to their natural ability to incorporate irregular HGCAL geometry and timing inputs

for a *deep spatio-temporal combination* with the eye towards addressing further Phase-2 HL-LHC reconstruction challenges with end-to-end deep learning.

Period	Deliverables
8/22-10/22	Develop standalone E2E CNN e benchmark (BM) on single CPU and GPU Develop single-GPU inference on E2E CNN e BM Develop vision transformer (VIT) application for E2E τ BMs
11/21-1/22	Extend graph neural network (GNN) application for E2E τ BMs Perform scaling multi-GPU studies on standalone e BMs Develop vision transformer (VIT) application for E2E e BMs
2/22-4/22	Develop single-GPU inference on E2E VIT e BM within CMSSW Develop E2E mass regression BM Integrate E2E VIT e and τ BMs into CMSSW Benchmark E2E VIT/GNN e and τ CPU inference within CMSSW
5/22-8/22	Integrate E2E mass regression into CMSSW Develop CMSSW multi-GPU inference on E2E VIT/G(T)NN e and τ BMs Study E2E e and τ inference within CMSSW on GPU and FPGA

Milestones:

- Q3 2022. Single CPU/GPU inference and classification performance of standalone E2E e benchmark (BM).
- Q4 2022. GNN application for E2E on the τ BM. Multi-GPU training and inference of standalone E2E e BM. VIT application for E2E on the e BM.
- Q1 2023. Inference of E2E VIT e and τ BM. Inference of E2E GNN τ BM within CMSSW. E2E mass regression application BM. Benchmarking and performance plots of CPU inference on E2E VIT/GNN e and τ BM within CMSSW.
- Q2 2023. Benchmarking and performance plots of single-GPU and FPGA inference on E2E VIT/GNN τ BM within CMSSW. Full integration of E2E mass regression in CMSSW. Multi-GPU inference on E2E VIT/GNN e and τ BM within CMSSW.

Project Synergy within US-CMS and CMS Machine Learning Group. The proposed research program is closely aligned with the vision and priorities of the CMS ML group. **PI Gleyzer** will foster involvement of other interested CMS groups in this key machine learning reconstruction activity, with University of Alabama providing a key push. The proposed project is also closely aligned with US-CMS Software and Computing Operations and *R&D* efforts in the area of *heterogeneous computing resources*, *detector reconstruction* and *analysis facilities and tools*. The PI will enable additional support for the proposed research program at the LPC, through collaboration with existing synergistic LPC machine learning efforts and with the broader CMS Machine Learning community. **PI Gleyzer** taught a graduate ML course at the LPC, training a number of PhD students who can potentially further contribute to the machine learning reconstruction effort.

In addition to the postdoctoral researcher, UA CMS group can commit additional resources: PhD students Ana Maria Slivar, Colin Crovella and Bhim Bam to this long-term broad machine learning reconstruction integration effort. Due to the maturity of the E2E effort and its relevance to CMS Machine Learning group activities, we foresee that the project will have a strong support structure to ensure long-term project success both in 2022-2023 and beyond.

5 Summary

In summary, we propose to work on integrating end-to-end deep learning algorithms into CMS reconstruction, including graph neural networks and vision and graph transformers, with a focus on accelerating deep learning-enabled reconstruction algorithms for LHC Run 3 and HL-LHC, enabled by heterogeneous computing.

References

[1] CMS collaboration, *The CMS experiment at the CERN LHC*, *JINST* **3** (2008) S08004.

[2] K. Albertsson et al., *Machine Learning in High Energy Physics Community White Paper*, *arXiv e-prints* (2018) [1807.02876].

[3] G. Apollinari, I. Béjar Alonso, O. Brüning, M. Lamont and L. Rossi, *High-Luminosity Large Hadron Collider (HL-LHC) : Preliminary Design Report*, .

[4] CMS Collaboration, *Technical Proposal for the Phase-II Upgrade of the CMS Detector*, Tech. Rep. CERN-LHCC-2015-010.

[5] CMS collaboration, *Particle-flow reconstruction and global event description with the CMS detector*, *JINST* **12** (2017) P1003.

[6] G. Louppe, K. Cho, C. Becot and K. Cranmer, *QCD-Aware Recursive Neural Networks for Jet Physics*, *arXiv e-prints* (2017) [1702.00748].

[7] P. W. Battaglia, R. Pascanu, M. Lai, D. Rezende and K. Kavukcuoglu, *Interaction Networks for Learning about Objects, Relations and Physics*, *ArXiv e-prints* (2016) [1612.00222].

[8] M. Stoye, J. Kieseler, H. Qu, L. Gouskos and M. Verzetti, "DeepJet: Generic physics object based jet multiclass classification for LHC experiments." NIPS Workshop on Deep Learning for Physical Sciences(2017).

[9] M. Andrews, M. Paulini, S. Gleyzer and B. Poczos, *End-to-End Physics Event Classification with the CMS Open Data: Applying Image-based Deep Learning on Detector Data to Directly Classify Collision Events at the LHC*, *Computing and Software for Big Science 1* (2020) [1807.11916].

[10] M. Andrews, J. Alison, S. An, P. Bryant, B. Burkle, S. Gleyzer et al., *End-to-End Jet Classification of Quarks and Gluons with the CMS Open Data*, *Nuclear Instruments and Methods A 977* (2020) [1902.08276].

[11] M. Andrews, B. Burkle, Y. Chen, D. DiCroce, D, S. Gleyzer et al., *End-to-end Jet Classification of Boosted Top Quarks with the CMS Open Data*, *Phys. Rev. D 105, 052008* (2022) [2104.14659].

[12] M. Andrews, B. Burkle, D. DiCroce, S. Gleyzer, U. Heintz, M Narain, M. Paulini, E. Usai, *End-to-End Jet Classification of Boosted Top Quarks with the CMS Open Data*, in *Proceedings, 25th International Conference on Computing in High Energy and Nuclear Physics (CHEP2021), EPJ Web of Conferences 251, 04030, 2021*

[13] CMS Collaboration, *Search for Exotic Higgs Boson Decays H to AA to 4 photons with Events Containing Two Merged Photons in Proton-Proton Collisions at sqrt(s) = 13 TeV*, CMS-PAS-HIG-21-016, 2022

[14] CMS Collaboration, *Using End-to-end Deep Learning with Domain Continuation to Reconstruct Merged Particle Decays to Diphotons in the CMS detector*, AN-20-149/EGM-20-001, 2022

[15] M. Andrews, B. Burkle, S. Chaudhary, D. DiCroce, S. Gleyzer, U. Heintz, M Narain, M. Paulini, E. Usai, *Accelerating End-to-End Deep Learning for Particle Reconstruction using CMS open data*, in *Proceedings, 25th International Conference on Computing in High Energy and Nuclear Physics (CHEP2021), EPJ Web of Conferences 251, 03057, 2021*

[16] CMS Collaboration, *Exploring End-to-end Deep Learning Applications for Event Classification at CMS*, in *Proceedings, 23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2018): Sofia, Bulgaria, July 9-13, 2018*, vol. 214, p. 06031, 2019, DOI.

[17] J. Alison, S. An, P. Bryant, B. Burkle, S. Gleyzer, M. Narain et al., *End-to-end particle and event identification at the Large Hadron Collider with CMS Open Data*, in *Meeting of the Division of Particles and Fields of the American Physical Society (DPF2019) Boston, Massachusetts, July 29-August 2, 2019*, 2019, [1910.07029].

[18] K. He, X. Zhang, S. Ren and J. Sun, *Deep residual learning for image recognition*, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, 2016, [DOI].

[19] A. Hariri, D. Dyachkova, S. Gleyzer, *Graph Generative Models for Fast Detector Simulations in High Energy Physics*, 2021, [2104.01725].

[20] A. Hariri, S. Gleyzer,D. Dyachkova, M. Awad, D. Morozova, *Graph Generative Models for Fast Detector Simulations in Particle Physics*, in *Third Machine Learning for Physical Sciences Workshop at NeurIPS2020*, 2020.

[21] A. Hariri, D. Dyachkova, S. Gleyzer, *Graph Generative Neural Networks for Fast Detector Simulations in High-Energy Physics*, in *Proceedings, 25th International Conference on Computing in High Energy and Nuclear Physics (CHEP2021), EPJ Web of Conferences 251, 03051, 2021*

[22] A. Hariri,D. Dyachkova, S. Gleyzer, *Scaling Graph Generative Models for Fast Detector Simulations in High Energy Physics*, *to appear in Nvidia GPU Technology Conference (GTC2021), April 11-16, 2021*

[23] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, *The graph neural network model*, *IEEE Transactions on Neural Networks* **20** (2009) 61.

[24] Y. Li, D. Tarlow, M. Brockschmidt and R. Zemel, *Gated Graph Sequence Neural Networks*, *arXiv e-prints* (2015) [1511.05493].

[25] J. Bruna, W. Zaremba, A. Szlam and Y. LeCun, *Spectral networks and locally connected networks on graphs*, [1312.6203].

[26] D. Zheng, V. Luo, J. Wu and J. B. Tenenbaum, *Unsupervised learning of latent physical properties using perception-prediction networks*, [1807.09244].

[27] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski et al., *Relational inductive biases, deep learning, and graph networks*, [1806.01261].

[28] X. Ju et al., *Graph Neural Networks for Particle Reconstruction in High Energy Physics detectors*, [2003.11603]

[29] N. Choma et al, *Track Seeding and Labelling with Embedded-space Graph Neural Networks*, [2007.00149]

[30] A. Heintz et al, *Accelerated Charged Particle Tracking with Graph Neural Networks on FPGAs* [2012.01563]

[31] A. Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* [2010.11929]

[32] S. Yun et al., *Graph Transformer Networks* [1911.06455]