A3D3 High-Throughput AI Methods and Infrastructure Workshop

Monday 10 July 2023 - Friday 14 July 2023

University of Washington

Book of Abstracts

Contents

1

Workshop Introduction & Overview

Corresponding Author: melissa.quinnan@cern.ch

2

Tips for an Effective Research Poster Presentation

A3D3 Town Hall / 3

Public Lecture

4

Internal A3D3 NSF Visit

please see https://indico.cern.ch/event/1263918/

5

Internal A3D3 NSF site visit

please see https://indico.cern.ch/event/1263918/

6

Workshop Conclusion

Corresponding Author: melissa.quinnan@cern.ch

7

Public Lecture

Working dinner / **9**

STNDT: Modeling Neural Population Activity with Spatiotemporal Transformers

Author: Trung Le[None]

Co-author: Eli Shlizerman

Corresponding Authors: tle45@uw.edu, shlizee@uw.edu

Modeling neural population dynamics underlying noisy single-trial spiking activities is essential for relating neural observation and behavior. A recent non-recurrent method - Neural Data Transformers (NDT) - has shown great success in capturing neural dynamics with low inference latency without an explicit dynamical model. However, NDT focuses on modeling the temporal evolution of the population activity while neglecting the rich covariation between individual neurons. In this paper we introduce SpatioTemporal Neural Data Transformer (STNDT), an NDT-based architecture that explicitly models responses of individual neurons in the population across time and space to uncover their underlying firing rates. In addition, we propose a contrastive learning loss that works in accordance with mask modeling objective to further improve the predictive performance. We show that our model achieves state-of-the-art performance on ensemble level in estimating neural activities across four neural datasets, demonstrating its capability to capture autonomous and non-autonomous dynamics spanning different cortical regions while being completely agnostic to the specific behaviors at hand. Furthermore, STNDT spatial attention mechanism reveals consistently important subsets of neurons that play a vital role in driving the response of the entire population, providing interpretability and key insights into how the population of neurons performs computation.

Working dinner / **10**

Machine learning for HEP simulations

Authors: Javier Mauricio Duarte[1]; Raghav Kansal[1]

[1] *Univ. of California San Diego (US)*

Corresponding Authors: raghav.kansal@cern.ch, javier.mauricio.duarte@cern.ch

Fast, accurate detector simulations are necessary to keep up with the data collected in the coming years in HEP. Due to their stochastic nature, ML-based generative models are natural opportunities for fast, differentiable simulations. We present two such graph- and attention-based models for generating LHC-like data using sparse and efficient point cloud representations, with state-of-the-art results. We measure a three-orders-of-magnitude improvement in latency compared to LHC full simulations, and also discuss recent work on evaluation metrics for validating such ML-based fast simulations.

Working dinner / **11**

Structural Re-weighting Improves Graph Domain Adaptation

Author: Shikun Liu[None]

Co-authors: Tianchun Li ; Yongbin Feng [1]; Nhan Tran [1]; Han Zhao [2]; Qiu Qiang [3]; Pan Li

[1] *Fermi National Accelerator Lab. (US)*

[2] *UIUC*

[3] *Purdue University*

Corresponding Authors: hanzhao@illinois.edu, ntran@fnal.gov, shikun.liu@gatech.edu, panli@gatech.edu, yongbin.feng@cern.ch, qqiu@purdue.edu, li2657@purdue.edu

In many real-world applications, graph-structured data used for training and testing have differences in distribution, such as in high energy physics (HEP) where simulation data used for training may not match real experiments. Graph domain adaptation (GDA) is a method used to address these differences. However, current GDA primarily works by aligning the distributions of node representations output by a single graph neural network encoder shared across the training and testing domains, which may often yield sub-optimal solutions. This work examines different impacts of distribution shifts caused by either graph structure or node attributes and identifies a new type of shift, named conditional structure shift (CSS), which current GDA approaches are provably sub-optimal to deal with. A novel approach, called structural reweighting (StruRW), is proposed to address this issue and is tested on synthetic graphs, four benchmark datasets, and a new application in HEP. StruRW has shown significant performance improvement over the baselines in the settings with large graph structure shifts and reasonable performance improvement when node attribute shift dominates.

Working dinner / **12**

Lyapunov-Guided Embedding for Hyperparameter Selection in Recurrent Neural Networks

Authors: Eli Shlizerman[None]; Ryan Vogt[1]; YANG ZHENG[None]

[1] *University of Washington*

Corresponding Authors: shlizee@uw.edu, ravogt95@uw.edu, zheng94@uw.edu

Recurrent Neural Networks (RNN) are ubiquitous computing systems for sequences and multivariate time series data. While several robust architectures of RNN are known, it is unclear how to relate RNN initialization, architecture, and other hyperparameters with accuracy for a given task. In this work, we propose to treat RNN as dynamical systems and to correlate hyperparameters with accuracy through Lyapunov spectral analysis, a methodology specifically designed for nonlinear dynamical systems. To address the fact that RNN features go beyond the existing Lyapunov spectral analysis, we propose to infer relevant features from the Lyapunov spectrum with an Autoencoder and an embedding of its latent representation (AeLLE). Our studies of various RNN architectures show that AeLLE successfully correlates RNN Lyapunov spectrum with accuracy. Furthermore, the latent representation learned by AeLLE is generalizable to novel inputs from the same task and is formed early in the process of RNN training. The latter property allows for the prediction of the accuracy to which RNN would converge when training is complete. We conclude that representation of RNN through Lyapunov spectrum along with AeLLE, and assists with hyperparameter selection of RNN, provides a novel method for organization and interpretation of variants of RNN architectures.

Research Subgroup AI Tools and Developments Talks / **13**

Eli Shlizerman Group (UW)

Corresponding Author: jingyli6@uw.edu

Research Subgroup AI Tools and Developments Talks / **14**

Amy Orsborn Group (UW)

Corresponding Authors: lnpeter4@uw.edu, sijia66@uw.edu

Research Subgroup AI Tools and Developments Talks / 15

Maria Makin Group (Purdue)

Corresponding Authors: park1377@purdue.edu, liptonm@purdue.edu

Research Subgroup AI Tools and Developments Talks / 16

Michael Coughlin Group (UMN)

Corresponding Author: healyb@umn.edu

Research Subgroup AI Tools and Developments Talks / 17

Matthew Graham Group (Caltech)

Corresponding Author: mjg@caltech.edu

Research Subgroup AI Tools and Developments Talks / 18

Erik Katsavounidis Group (MIT)

Corresponding Authors: emarx@mit.edu, eric.anton.moreno@cern.ch, alecg@mit.edu

Research Subgroup AI Tools and Developments Talks / 19

Kate Scholberg Group (Duke)

Corresponding Author: janina.hakenmuller@duke.edu

Research Subgroup AI Tools and Developments Talks / 20

Kael Hanson Group (Wisconsin)

Corresponding Author: josh.peterson@icecube.wisc.edu

Research Subgroup AI Tools and Developments Talks / 21

Javier Duarte Group (UCSD, CMS)

Corresponding Authors: javier.mauricio.duarte@cern.ch, rmarroquinsolares@ucsd.edu, daniel.cipriano.diaz@cern.ch, melissa.quinnan@cern.ch

Research Subgroup AI Tools and Developments Talks / 22

Miaoyuan Liu Group (Purdue, CMS)

Corresponding Authors: miaoyuan.liu@cern.ch, dmitry.kondratyev@cern.ch, jan-frederik.schulte@cern.ch

Research Subgroup AI Tools and Developments Talks / 23

Philip Harris Group (MIT, CMS)

Corresponding Author: william.patrick.mc.cormack.iii@cern.ch

Research Subgroup AI Tools and Developments Talks / 24

Shih-Chieh Hsu Group (UW, ATLAS)

Corresponding Author: elham.e.khoda@cern.ch

Research Subgroup AI Tools and Developments Talks / 25

Mark Neubauer Group (UIUC, ATLAS)

Corresponding Author: dewen.zhong@cern.ch

Research Subgroup AI Tools and Developments Talks / 26

Scott Hauck Group (UW)

Corresponding Author: xliu1626@uw.edu

Research Subgroup AI Tools and Developments Talks / 27

Pan Li Group (Georgia Tech)

Corresponding Author: shikun.liu@gatech.edu

Research Subgroup AI Tools and Developments Talks / **28**

Deming Chen Group (UIUC)

Corresponding Author: jz23@illinois.edu

Working dinner / **29**

Predicting Pulsed-Laser Deposition SrTiO3 Homoepitaxy Growth Dynamics using High-Speed Reflection High-Energy Electron Diffraction

Authors: Yichen Guo[None]; Peter Meisenheimer[None]

Co-authors: Shuyu Qin ; Xinqiao Zhang ; Julian Goddy ; Lane Martin ; Ramamoorthy Ramesh ; Joshua Agar [1]

[1] *Drexel University*

Corresponding Authors: jca92@drexel.edu, yig319@lehigh.edu

Pulsed-laser deposition (PLD) is a powerful technique to grow complex oxides with controlled stoichiometry. To understand growth dynamics, it is common to leverage in situ spectroscopies such as reflection high energy electron diffraction (RHEED) to monitor surface crystallinity. Most commercial systems rely on video-rate cameras operating at 60-120 Hz that lack sufficient temporal resolution to capture growth dynamics at practical deposition frequencies. Here, we implement a high-speed platform to record in situ dynamics via RHEED at >500 Hz. We design an open-source analysis package to fit diffraction spots to 2D Gaussians, allowing single-pulse surface reconstruction kinetics extraction. Using homoepitaxially deposited (001)-oriented SrTiO3 as a model system, we demonstrate how high-speed RHEED can provide real time insight into growth processes obscured by slower acquisition systems. By fitting the single-pulse intensity to a set of exponential functions, we observe changes in the characteristic decay time and mechanism correlated to the substrate step width and surface termination. Specifically, we observe exponential decay in per pulse intensity when depositing on lower energy TiO2-terminated surfaces. Conversely, exponential stabilization is observed when surfaces are SrO or mixed terminated. Similarly, the extracted characteristic time decreases with an increase in the density of bonding sites associated with mixed termination and narrower step widths. Ultimately, this work shows how increasing RHEED temporal resolution can uncover new insight into growth processes. This experimental platform provides new capabilities to enable data-driven machine learning analysis and autonomous control systems to enhance the complexity and fecundity of PLD.

Working dinner / **30**

Graph Neural Network Triggers for □ -> 3□ Events at the HL-LHC

Author: Benjamin Simon[1]

Co-authors: Daniel Guerrero Ibarra [2]; Jacobo Konigsberg [3]; Jan-Frederik Schulte [4]; Miaoyuan Liu [4]; Pan Li [5]; Siqi Miao [5]

[1] *Purdue University (US)*

[2] *Fermilab*

[3] *University of Florida*

[4] *Purdue University*

[5] *Georgia Institute of Technology*

Corresponding Authors: simon73@purdue.edu, schul105@purdue.edu, miaoyuan.liu@cern.ch, daniel.guerrero@cern.ch, jacobo.konigsberg@cern.ch, siqim@gmail.com, panli@gatech.edu

A graph neutral network (GNN) was constructed to identify charged lepton flavor violating decays of a tau particle into three muons in proton-proton collisions recorded with the CMS detector of the Large Hadron Collider. The muons from this decay are expected to have very low momentum, making them hard to detect in the high pileup environment expected at the high luminosity LHC (HL-LHC). We therefore propose the use of a GNN to select signal candidate events for readout and storage in the Level-1 trigger system during the run of the HL-LHC. Current standard model calculations indicate that the -> 3 decay is extremely improbable with a branching ratio of ˜O(10-55), but some beyond-the-standard-model (BSM) physics models predict a much larger branching ratio of ˜O(10-8). A large trigger acceptance for these events is crucial to maximize sensitivity to this potential signal of BSM physics. For this purpose, a GNN trigger was developed. The trigger's performance was evaluated by determining the projected yield of accepted signal events at various trigger rates in two phase spaces of interest and comparing to current CMS algorithms. Over the lifespan of the HL-LHC, the GNN trigger is projected to accept ˜80k signal events at a trigger rate of 77kHz, greatly improving the current trigger's projected yield of ˜16k events at the same rate. The graph neural network's high performance makes it a strong candidate for use in the CMS trigger system to enhance the discovery potential for BSM physics.

Research Subgroup AI Tools and Developments Talks / **31**

Song Han Group (MIT)

Corresponding Author: wweichen@mit.edu

Working dinner / **32**

Decoding Upsampled Limb Trajectories of a Running Mouse from 2-Photon Calcium Imaging Using a Recurrent Neural Network Encoder-Decoder

Authors: Seungbin Park[None]; Megan Hope Lipton[None]; Maria Makin[1]

[1] *Purdue University*

Corresponding Authors: liptonm@purdue.edu, mdadarla@purdue.edu, park1377@purdue.edu

Neural decoding is a critical task for understanding the function of the brain and providing solutions for neurological injury and disease. Two-photon calcium imaging has been a promising recording technique to observe a large population of neurons; however, decoding from two-photon calcium images is challenging because of the indirect and nonlinear representation of neural activity, low sampling rates, and slow kinematics. Here, we present the approach of using a recurrent neural network encoder-decoder to decode the limb positions of a running mouse from two-photon calcium images. The neural network could decode limb coordinates sampled at 30 Hz from two-photon calcium images sampled below 8 Hz with the root mean squared errors of 25.35 pixels (3.80 mm). Information about all four limbs (contralateral and ipsilateral front and hind limbs) could be decoded from a single cortical hemisphere. A fraction of the most informative neurons yielded higher decoding accuracy than randomly-sampled neurons. Nevertheless, overall accuracy was directly proportional

to the number of neurons used to decode. This study validates the feasibility of using calcium imaging to decode continuous behavior variables with a higher sampling rate for understanding brain function and brain-machine interfaces.

Working dinner / **33**

MCUNetV1 & V2: On-Device Inference of Tiny Deep Learning on IoT Devices

Authors: Ji Lin[1]; Wei-Ming Chen[1]; Yujun Lin[1]; Han Cai[1]; Wei-Chen Wang[1]; John Cohn[2]; Chuang Gan[2]; Song Han[1]

[1] *MIT*

[2] *MIT-IBM Watson AI Lab*

Corresponding Author: wweichen@mit.edu

Machine learning on tiny IoT devices based on microcontroller units (MCU) is appealing but challenging: the memory of microcontrollers is 2-3 orders of magnitude smaller even than mobile phones. We propose MCUNet, a framework that jointly designs the efficient neural architecture (TinyNAS) and the lightweight inference engine (TinyEngine), enabling ImageNet-scale inference on microcontrollers. TinyNAS adopts a two-stage neural architecture search approach that first optimizes the search space to fit the resource constraints, then specializes the network architecture in the optimized search space. TinyNAS can automatically handle diverse constraints (i.e. device, latency, energy, memory) under low search costs. TinyNAS is co-designed with TinyEngine, a memory-efficient inference library to expand the search space and fit a larger model. TinyEngine adapts the memory scheduling according to the overall network topology rather than layer-wise optimization, reducing the memory usage by 3.4x, and accelerating the inference by 1.7-3.3x compared to TF-Lite Micro and CMSIS-NN. MCUNet is the first to achieve>70% ImageNet top1 accuracy on an off-the-shelf commercial microcontroller, using 3.5x less SRAM and 5.7x less Flash compared to quantized MobileNetV2 and ResNet-18. On visual&audio wake words tasks, MCUNet achieves state-of-the-art accuracy and runs 2.4-3.4x faster than MobileNetV2 and ProxylessNAS-based solutions with 3.7-4.1x smaller peak SRAM. Our study suggests that the era of always-on tiny machine learning on IoT devices has arrived.

Working dinner / **34**

MCUNetV3: On-Device Training Under 256KB Memory

Authors: Ji Lin[1]; Ligeng Zhu[1]; Wei-Ming Chen[1]; Wei-Chen Wang[1]; Chuang Gan[2]; Song Han[1]

[1] *MIT*

[2] *MIT-IBM Watson AI Lab*

Corresponding Author: wweichen@mit.edu

On-device training enables the model to adapt to new data collected from the sensors by fine-tuning a pre-trained model. However, the training memory consumption is prohibitive for IoT devices that have tiny memory resources. We propose an algorithm-system co-design framework to make on-device training possible with only 256KB of memory. On-device training faces two unique challenges: (1) the quantized graphs of neural networks are hard to optimize due to mixed bit-precision and the lack of normalization; (2) the limited hardware resource (memory and computation) does not allow full backward computation. To cope with the optimization difficulty, we propose Quantization-Aware Scaling to calibrate the gradient scales and stabilize quantized training. To reduce the memory footprint, we propose Sparse Update to skip the gradient computation of less important layers and

sub-tensors. The algorithm innovation is implemented by a lightweight training system, Tiny Training Engine, which prunes the backward computation graph to support sparse updates and offload the runtime auto-differentiation to compile time. Our framework is the first practical solution for on-device transfer learning of visual recognition on tiny IoT devices (e.g., a microcontroller with only 256KB SRAM), using less than 1/1000 of the memory of existing frameworks while matching the accuracy of cloud training+edge deployment for the tinyML application VWW. Our study enables IoT devices to not only perform inference but also continuously adapt to new data for on-device lifelong learning.

Working dinner / **35**

Using convex feature selection to improve offline feature decoding

Author: Lauren Peterson[1]

Co-authors: Si Jia Li [2]; Leo Scholl [3]; Pavithra Rajeswaran [2]; Lydia Smith [4]; Amy Orsborn

[1] *University of Washington*

[2] *UW Bioengineering*

[3] *UW Electrical & Computer Engineering*

[4] *WaNPRC*

Corresponding Authors: lydias3@uw.edu, pavir@uw.edu, aorsborn@uw.edu, lscholl@uw.edu, sijia66@uw.edu, lnpeter4@uw.edu

Brain-computer interfaces use the electrical activity of the brain to control an external device, but decoding complex neural signals requires large amounts of computational power and time. We use a novel convex optimization algorithm to do real-time feature selection based on relevance, sparsity, and smoothness. We demonstrate that the algorithm can reduce the feature set while maintaining decoding accuracy.

Working dinner / **36**

Accelerating CNNs on FPGAs for Particle Energy Reconstruction

Authors: Alexander Joseph Schuy[1]; Bo-Cheng Lai[None]; ChiJui Chen[None]; Dylan Ranklin[2]; Ling Chi Yang[3]; Philip Coleman Harris[4]; Scott Hauck[None]; Shih-Chieh Hsu[5]; Yan Lun Huang[3]; Ziang Yin[1]

[1] *University of Washington (US)*

[2] *University of Pennsylvania*

[3] *National Yang Ming Chiao Tung University*

[4] *Massachusetts Inst. of Technology (US)*

[5] *University of Washington Seattle (US)*

Corresponding Authors: lostecho@uw.edu, hauck@uw.edu, dsrankin@sas.upenn.edu, kugelblitz.ee05@gmail.com, hisky1256@gmail.com, alexander.joseph.schuy@cern.ch, philip.coleman.harris@cern.ch, bclai@nycu.edu.tw, schsu@uw.edu, yanlun172@gmail.com

Given the recent advances of machine learning techniques, the Large Hadron Collider (LHC) at CERN is incorporating deep learning (DL) models, such as DeepCalo, to enhance the quality of data analysis of particle experiments. However, the need for in-time inference to keep up with data generation rates, as well as the dynamics of the experiments, require that the data processing feature short processing latency as well as flexibility to quickly implement different DL models. The LHC plans to

use FPGAs (Field Programmable Gate Arrays) to provide timely data analysis via the highly parallel dataflow-based processing and short latency enabled by customized logic. A high level synthesis tool, hls4ml, is also adopted to facilitate design and synthesis of the fully on-chip dataflow architecture which avoids long-latency DRAM accesses. However, the current hls4ml framework has limited support for very large CNN models due to suboptimal data streaming schemes and inefficient processing architectures. The dataflow architecture also requires proper data quantization to efficiently utilize the limited resources within the FPGA. In this paper, we present the first automated design and optimization workflow based on hls4ml to implement DeepCalo models on FPGAs. The current DeepCalo framework is extended and integrated with QKeras layers to perform quantization-aware training to minimize resource consumption while retaining good model quality. A comprehensive exploration is performed on various key design factors, and observations have been summarized as useful design guidelines for future applications. With the proposed workflow, we have shown that the design on a Xilinx Alveo U50 FPGA can significantly outperform the implementations on Ryzen-5600H CPUs and Tesla V100 GPUs by up to 14.1x and 7.9x respectively, and meet the latency requirement of the HLT (High Level Trigger) within the particle
experiment.

Working dinner / **37**

Accelerating Hadronic Calorimetry with Sparse Point-Voxel Convolutional Neural Networks

Authors: Alexander Joseph Schuy[1]; Haoran Zhao[1]; Haotian Tang[2]; Jeffrey Krupa[2]; Philip Coleman Harris[3]; Scott Hauck[None]; Shih-Chieh Hsu[4]; Song Han[5]; William Patrick Mccormack[3]; Zhijian Liu[2]

[1] *University of Washington (US)*

[2] *Massachusetts Institute of Technology*

[3] *Massachusetts Inst. of Technology (US)*

[4] *University of Washington Seattle (US)*

[5] *MIT*

Corresponding Authors: william.patrick.mc.cormack.iii@cern.ch, haoran.zhao@cern.ch, alexander.joseph.schuy@cern.ch, zhijian@mit.edu, philip.coleman.harris@cern.ch, hauck@uw.edu, jeffrey.krupa@cern.ch, schsu@uw.edu

In this study, we demonstrate the potential of sparse point-voxel convolutional neural networks (SPVCNN) for hadronic calorimetry tasks using HCAL and HGCAL datasets. By employing a modified object condensation loss, we train the network to group cell deposits into clusters while filtering out noise. We show that SPVCNN performs comparably to generic topological cluster-based methods in both pileup and no pileup scenarios, with the added advantage of acceleration using GPUs. This type of acceleration, as part of heterogeneous computing frameworks, will be crucial for the High-Luminosity Large Hadron Collider (HL-LHC). Our findings indicate that SPVCNN can provide efficient and accurate calorimetry solutions, particularly for high level trigger (HLT) applications with latency on the order of milliseconds.

38

Internal A3D3 discussion

please see https://indico.cern.ch/event/1263918/

39

Poster Award Presentation

Corresponding Authors: melissa.quinnan@cern.ch, daniel.cipriano.diaz@cern.ch

Working dinner / **41**

Sleep Spindle (LFADs) Project Abstract

Author: Xiaohan Liu[None]

Corresponding Author: xliu1626@uw.edu

A specific type of Electroencephalography (EEG) signals, sleep spindle, is believed to contribute to neuronal plasticity and memory consolidation. In this project, we proposed a system that is based on ultra-low latency and power FPGA to detect and interact with the sleep spindles to further understand the mechanism behind the theory. The proposed system will have a programmed FPGA that connects with a headstage. The headstage will record the subject's brain signals and the FPGA will process the signals to detect and interact with the sleep spindles.

Latent Factor Analysis via Dynamical Systems (LFADs) is the baseline deep learning model for this project. It is an RNN variational autoencoder for analyzing spiking neural data. LFADs follows the encoder-decoder structure. The input spiking data will be sent to a bidirectional GRU, Gaussian sampling, a unidirectional GRU, and several dense layers to produce the final outputs. LFADs will generate two vital outputs. One is a set of low-dimensional temporal factors which contains the information of the input spiking data. The other is the log firing rate, which can be translated to the firing rate that can generate the input spiking data.

We removed the Gaussian sampling from LFADs in this current project and will add the sampling layer back later. In this case, we have successfully deployed the no-sampling LFADs onto FPGA by implementing the unsupported layers in HLS4ML. The FPGA we used for the current deployment is Xilinx Alveo U50. By running LFADs on this board, we have significantly decreased the latency while maintaining reasonable model performance. We are currently working on optimizing the model by applying quantization aware training (QAT) on LFADs and managing to deploy multiple LFADs onto FPGA to enable high throughput processing. The high throughput processing can be used for analyzing largescale neural recordings and the low latency processing will allow neuroscientists to develop closed-loop technology to analyze the neural signal in real time.

Working dinner / **42**

Symmetry Informed Autoencoder for Domain Classification of BaTiO3 Brightfield Images

Author: Xinqiao Zhang[None]

Corresponding Author: zhang.xinqiao314@gmail.com

Ferroelectrics, characterized by spontaneous polarization and reversible switching, play a crucial role in various applications such as non-volatile FeRAM, ferro-TFET, and catalysis. However, the influence of environmental factors on ferroelectric domain dynamics remains poorly characterized. This work aims to investigate the impact of temperature and background gas on the domain mapping of BTO, considering the challenges posed by sample warping (˜150nm thickness) and reduced signal-to-noise ratio for linear analysis methods like PCA.

To address these challenges, deep learning techniques are employed to capture noise and non-linearities in the data. Previous research in our group has utilized autoencoders to learn domain structures from STEM images, and affine transforms to extract symmetry information. In this study, we extend the approach by simulating diffraction patterns through windowing and Fourier Transform, commonly used in electron microscopy. An autoencoder augmented with an affine grid is trained to learn the symmetries and periodicities of the FFT windows. The methodology is applied to characterize

phases, sample warping, and contamination in an ideal ultra-high vacuum sample. Transfer learning is then employed to analyze lower-quality scans with background gas, leveraging the knowledge gained from the trained model.

The use of symmetry informed autoencoders enables more efficient analysis compared to manual phase mapping, removing human bias and providing a pathway for real-time analysis of brightfield images.

Working dinner / **43**

Progress towards an improved particle-flow algorithm at CMS with machine learning

Authors: Eric Wulff[1]; Farouk Mokhtar[2]; Javier Mauricio Duarte[2]; Joosep Pata[3]; Michael Zhang[None]

[1] *CERN*

[2] *Univ. of California San Diego (US)*

[3] *National Institute of Chemical Physics and Biophysics (EE)*

Corresponding Authors: javier.mauricio.duarte@cern.ch, joosep.pata@cern.ch, eric.wulff@cern.ch, fmokhtar@ucsd.edu, mezhang@ucsd.edu

The particle-flow (PF) algorithm, which infers particles based on tracks and calorimeter clusters, is of central importance to event reconstruction in the CMS experiment at the CERN LHC, and has been a focus of development in light of planned Phase-2 running conditions with an increased pileup and detector granularity. In recent years, the machine learned particle-flow (MLPF) algorithm, a graph neural network that performs PF reconstruction, has been explored in CMS, with the possible advantages of directly optimizing for the physical quantities of interest, being highly reconfigurable to new conditions, and being a natural fit for deployment to heterogeneous accelerators. We discuss progress in CMS towards an improved implementation of the MLPF reconstruction, now optimized using generator/simulation-level particle information as the target for the first time. This paves the way to potentially improving the detector response in terms of physical quantities of interest. We describe the simulation-based training target, progress and studies on event-based loss terms, details on the model hyperparameter tuning, as well as physics validation with respect to the current PF algorithm in terms of high-level physical quantities such as the jet and missing transverse momentum resolutions. We find that the MLPF algorithm, trained on a generator/simulator level particle information for the first time, results in broadly compatible particle and jet reconstruction performance with the baseline PF, setting the stage for improving the physics performance by additional training statistics and model tuning.

Working dinner / **44**

Self-Supervised Learning for Jet Tagging

Authors: Carlos Pareja[1]; Farouk Mokhtar[2]; Javier Mauricio Duarte[2]; Raghav Kansal[2]; Zihan Zhao[2]

[1] *University of California, San Diego*

[2] *Univ. of California San Diego (US)*

Corresponding Authors: z.zhao@cern.ch, raghav.kansal@cern.ch, fmokhtar@ucsd.edu, cpareja@ucsd.edu, javier.mauricio.duarte@c

Limited by the lack of truth labels on real data, fully supervised ML algorithms are constrained to training only with simulated samples. With self-supervised learning, we can leverage vast amounts of unlabeled real data to facilitate training. We investigate the application of VICReg, a contrastive

learning model, on a classification task: discriminating signal jets (e.g. $H \to b\bar{b}$ jets) from background jets (e.g. QCD jets). We also explore the use of jet augmentations in contrastive learning.

Working dinner / **45**

Denoising Autoencoder for LArTPC Detectors

Author: Van Tha Bik Lian[1]

[1] *Duke University*

Corresponding Author: vanthabik.lian@duke.edu

We present a denoising autoencoder for extracting low-energy signals in Liquid Argon Time Projection Chamber (LArTPC) detectors. In particular, we are interested in neutrinos originating from core-collapse supernova events, and the detection of these neutrinos can help improve our knowledge of the physics of core-collapse supernova events 1. Additionally, if we can detect them fast enough, we can provide an alert via the SuperNova Early Warning System, so that other observatories may direct their telescopes and detectors at the supernova. However, these neutrinos can have energies on the order of 10MeV, which makes their detection challenging because the electronic signals resulting from their interaction in liquid argon are close to noise levels. To address this, we apply an autoencoder consisting of convolutional layers to denoise and extract these low-energy signals. We show that the autoencoder can detect the presence of low-energy signals better than a threshold-based method, and we present the model's ability to denoise these signals.

Working dinner / **46**

Searching Better, Faster: Detecting Binary Black Hole Mergers with Deep Learning Networks

Authors: Alec Gunny[None]; Ethan Marx[1]; Will Benoit[None]

Co-authors: Deep Chatterjee ; Dylan Sheldon Rankin [2]; Eric Anton Moreno [3]; Erik Katsavounidis [1]; Michael Coughlin [4]; Muhammed Saleem Cholayil [4]; Philip Coleman Harris [5]; Rafia Omer [4]; Ryan Raikman

[1] *MIT*

[2] *University of Pennsylvania (US)*

[3] *Massachusetts Institute of Technology (US)*

[4] *University of Minnesota*

[5] *Massachusetts Inst. of Technology (US)*

Corresponding Authors: mcholayi@umn.edu, deep1018@mit.edu, kats@mit.edu, philip.coleman.harris@cern.ch, benoi090@umn.edu, eric.anton.moreno@cern.ch, emarx@mit.edu, michael.w.coughlin@gmail.com, alecg@mit.edu, dylan.sheldon.rankin@cern.ch

As the global network of gravitational wave detectors grows in both size and sensitivity, the traditional matched filtering method for detecting signals from compact object mergers becomes computationally prohibitive. Machine learning algorithms are a compelling alternative approach to this problem due to their ability to shift the computational cost to the model training process, enabling efficient use of computational resources at search time. Here, we present an end-to-end binary black hole search pipeline, aframe, capable of low-latency identification of binary black hole mergers in LIGO data. Using simulated binary black hole signals and real LIGO noise and glitches, a 1-dimensional convolutional neural network is trained to perform a binary classification of detector strain timeseries data. Further, we highlight the novel infrastructure development steps that have

been taken to improve the robustness of our model and establish a paradigm for applying machine learning solutions to gravitational wave problems.

Working dinner / **47**

Parameter Estimation of Unmodeled Burst Gravitational Waves Using Likelihood-free Inference

Authors: Deep Chatterjee[None]; Ethan Marx[1]

Co-authors: Alec Gunny ; Dylan Sheldon Rankin [2]; Eric Anton Moreno [3]; Erik Katsavounidis [1]; Katya Govorkova [4]; Michael Coughlin [5]; Muhammed Saleem Cholayil [5]; Philip Coleman Harris [4]; Rafia Omer [5]; Ryan Raikman ; Will Benoit

[1] *MIT*

[2] *University of Pennsylvania (US)*

[3] *Massachusetts Institute of Technology (US)*

[4] *Massachusetts Inst. of Technology (US)*

[5] *University of Minnesota*

Corresponding Authors: philip.coleman.harris@cern.ch, michael.w.coughlin@gmail.com, dylan.sheldon.rankin@cern.ch, eric.anton.moreno@cern.ch, ekaterina.govorkova@cern.ch, benoi090@umn.edu, kats@mit.edu, deep1018@mit.edu, mcholayi@umn.edu, alecg@mit.edu, emarx@mit.edu

The observed events from the LIGO-Virgo-Kagra collaboration (LVK) have been modeled sources called compact binary coalescences (CBCs). Un-modeled transients, for example, from core-collapse supernovae or pulsar glitches remain undiscovered. In this work, we demonstrate the use of likelihood-free inference using normalizing flows for parameter estimation of un-modeled burst-type gravitational-wave signals. Our framework is designed for real-time parameter estimation for generic Sine-Gaussian morphology that make minimal assumptions about the source. We show the ability of our model to accurately recover sky localization and intrinsic parameters like frequency and quality from sources embedded in real data from the LVK third observing run. The time of inference is significantly faster, and comparable in accuracy to stochastic sampling techniques, like MCMC and nested sampling, used traditionally. This is crucial in the light of increasing sensitivity of the LIGO-Virgo-KAGRA instruments requiring higher throughput especially in real-time setting.

Working dinner / **48**

UNRAVELING GRAVITATIONAL RIPPLES: NEURAL NETWORK CLASSIFICATION

Authors: Daniel G Fredin[1]; Cole Welch[1]

[1] *University of Washington*

Corresponding Authors: colewelch151@gmail.com, dfredin@uw.edu

The Laser Interferometer Gravitational Wave-Observatory (LIGO) has accumulated more than 4.5 petabytes (Pb) of data in its quest to detect gravitational waves. Furthermore, it is anticipated that the total data accrued will increase by approximately 0.8 petabytes per year. The processing and analysis of the extensive volume of data from LIGO necessitates a tremendous amount of computational resources and time. Among the most computationally demanding stages are the initial steps of signal extraction and classification, which pose significant challenges. There is a critical need for more efficient detection and classification algorithms that can overcome these current challenges. In this study, we introduce a machine learning methodology utilizing a binary classifier to differentiate and categorize the data. More specifically, we train a convolutional neural network (CNN) using

approximately 100,000 simulated time series data samples of gravitational waves encompassing four distinct signal categories: glitches, background noise, binary black hole, and sine-gaussian. To facilitate the recognition and classification of data, our approach involves encoding the time series data into images using Gramian Angular Summation Fields (GASF). By employing this encoding technique, we enable our convolutional neural network (CNN) to identify and categorize the data. Our primary objective is to classify the GASF images into one of the following two groups: noise/glitch signals or transient sine-gaussian/binary black hole signals. Our model achieved a testing accuracy of 97%, demonstrating a high effectiveness when compared to other approaches to classification in the literature. However, there is a need for further investigation into the viability of the Gramian Angular Summation Field approach in converting real LIGO strain data into images and how our approach compares to the current standard, which consists of converting signal data to spectrograms by taking fast Fourier transforms. This exploration aims to evaluate the relative effectiveness of the GASF method and determine its potential for practical implementation.

Working dinner / **49**

Testing the Supernova Pointing Resolution of DUNE with ICEBERG

Author: Joshua Queen[None]

Corresponding Author: joshua.queen@duke.edu

The Deep Underground Neutrino Experiment (DUNE), a 40 kt fiducial mass liquid argon time projection chamber (LArTPC), will be unique among supernova (SN) neutrino detectors due to its ability to measure the electron neutrino flavor component of a SN burst. Crucial to achieving a good pointing resolution is the ability to discriminate the directionality of primary electron tracks via a process known as daughter flipping. The daughter flipping algorithm takes the vertex of an electron track and determines whether this vertex is the “head” or “tail” of the track via the angle between it and daughter particles produced, such as ionization tracks from bremsstrahlung gammas. Studies are ongoing to understand DUNE’s SN pointing ability with daughter flipping, however, the daughter flipping reconstruction algorithms have only been used on simulated data. The ICEBERG detector, a small ˜1-ton LArTPC located at Fermi National Accelerator Laboratory, is equipped with the DUNE DAQ system and thus permits the testing of reconstruction algorithms on data. Due to their similar energy scale compared to SN neutrinos, Michel electrons can be used as a proxy to study the reconstruction algorithms of SN neutrino produced electrons. Therefore, Michel electrons will be simulated in the ICEBERG detector along with detector responses to produce selection criteria for Michel candidates in the ICEBERG data. Once Michel electron candidates are identified, the daughter flipping algorithm can be applied to the data to test its performance. A data driven noise model will also be developed in ICEBERG, permitting testing of the daughter flipping algorithm as a function of noise. By characterizing the daughter flipping algorithm through data, DUNE’s SN pointing performance can be better understood. This poster will describe the concept and current status of this study.

Working dinner / **50**

SpectroGW: A computer vision model for Binary Black Hole merger classification

Author: Arif Chu[1]

[1] *University of Washington Seattle*

Corresponding Author: arifchu@gmail.com

Spectrograms are frequently used to provide qualitative insights into the types of noise and signals present in audio data. Similarly, we can use them to gain insights from data such as real gravitational

wave from gravitational wave detectors. Simply by eye, we can see characteristic chirp signals from gravitational waves due to the physics of the black holes' inspiral. Designing a novel machine learning model named SpectroGW, which is based on a convolutional neural network, we can determine if spectrograms containing gravitational wave signals does in fact possess qualitative characteristics. The model can be modified to predict the masses of the merging bodies. We can utilize fast Fourier transforms and amplitude spectral density can be used to characterise the changes in the data across time in the frequency domain to bring out the feature of the binary black hole merger's chirp. Additionally, we can identify the features and types of noise present in the gravitational wave data.

Working dinner / **51**

Developments in Digital Optical Module Waveform Processing for the IceCube Neutrino Observatory

Author: Josh Peterson[None]

Co-authors: Benedikt Riedel [1]; Kael Hanson [1]

[1] *University of Wisconsin - Madison*

Corresponding Authors: kael.hanson@wisc.edu, josh.peterson@icecube.wisc.edu, briedel@icecube.wisc.edu

The IceCube Neutrino Observatory is a neutrino telescope located at the South Pole designed to detect Cherenkov radiation produced when neutrinos interact with the ice. It consists of 86 strings of digital optical modules, each containing a photomultiplier tube, embedded deep in the Antarctic ice. Each photon that encounters a photomultiplier tube produces a voltage waveform, and the photon information must be recovered from those waveforms. Currently, we utilize CPUs for this processing. However, if we need to reprocess this data it is very time consuming, power intensive, and expensive. Thus, it makes sense to accelerate this process with GPUs or FPGAs. Neural networks are highly compatible to both GPUs and FPGAs, so we are developing a neural network for PMT voltage waveform unfolding. In this poster presentation we present a simple compact neural network that is trained to find photon hits in simulated voltage waveforms. We find that the neural network is able to find single photon hit times nearly as reliably as the algorithms that are currently used. In the future, we plan to modify the neural network to find photon charge information as well, and we plan on implementing the CPU-based algorithm on GPU / FPGA.

Working dinner / **52**

Pointing to a supernova with the DUNE experiment

Author: Janina Hakenmüller[1]

[1] *Duke University*

Corresponding Author: janina.hakenmuller@duke.edu

The detection of a supernova burst is a unique opportunity to derive insights on astro and particle physics especially neutrinos. Neutrinos are the first hint of a supernova occuring to arrive on Earth due to their very low interaction cross section. They can provide extremely valuable information on the direction of burst enabling to point optical detection systems there in a multi messenger approach.
The Deep Underground Neutrino Experiment (DUNE) aims at detecting these neutrinos with time projection chambers (TPCs) containing up to 40 ktons of liquid argon, located under an overburden of about 1500 m. This technology has an excellent 3D imaging capability.
On my poster, I will show the pointing resolution achievable with the current framework. Ultimately, an online pointing analysis is required for the multi messenger approach. I will present an outline

on the conversion of the existing code to a fast online version highlighting where machine learning can improve the speed and precision of the result.

Working dinner / **53**

Data for Low-Latency Electromagnetic Training

Author: Abigail Gray[None]

Corresponding Author: agray@umn.edu

The current and upcoming Gravitational Wave (GW) observing runs by LIGO/Virgo/KAGRA detectors will result in significantly more detections than previous runs. Preparation to follow up associated electromagnetic signals promptly and accurately from binary neutron star (BNS) and neutron star black hole (NSBH) detections now depends heavily on real-time ML implementation at the detectors. The public alert data products from IGWN are at the center of real-time go-no-go triggering of EM telescopes, such as ZTF. We develop a comprehensive low-latency data set that incorporates all available data provided by IGWN alert stream. This dataset is built on GW localizations from observing scenarios simulations and draws of additional products are derived from associated injections for BNS and NSBH sources. While many ML models focused on electromagnetic follow-up are trained on select data products, this dataset will be the first to provide a complete set of alert products drawn from a realistic end-to-end GW-EM simulation. This data will be used to train classifiers and autonomous agents focused on optimizing low-latency follow-up strategies. By optimizing alert-to-trigger decision-making we positively impact resource allocation and technical capabilities as well as the collective scientific return of these highly valuable observing runs.

Working dinner / **54**

Hyperparameter Tuning for Semi-Supervised Graph Neural Network for Pileup Mitigation

Authors: Garyfallia Paspalaki[None]; Miaoyuan Liu[None]; Nhan Tran[None]; Pan Li[None]; Shikun Liu[None]; Tianchun Li[None]; Yongbin Feng[None]

Co-authors: Jack Rodgers ; Yuji Li

Corresponding Authors: garyfallia.paspalaki@cern.ch, panli@gatech.edu, jprodger@purdue.edu, li2657@purdue.edu, miaoyuan.liu@cern.ch, shikun.liu@gatech.edu, yuji.li@cern.ch, ntran@fnal.gov, yfeng@fnal.gov

In the Large Hadron Collider (LHC) at CERN, protons collide more than a million times per second. Pileup, which are interactions in the same or nearby proton bunch crossings in the accelerator, can be thought of as noise which affects many reconstructed physics variables such as the Jet Mass, Jet PT, and missing transverse momentum. This noise also results in worse resolution and as a consequence lower physics reconstruction performance. Furthermore, pileup is expected to increase by more than a hundred times in the timespan that the energy in the LHC is increased until 2029. Currently, an algorithm called PUPPI (Pileup Per Particle Identification) exists to mitigate pileup. However, recent machine learning developments can provide a power to be able to more effectively remove this noise. A proof of concept study that uses a semi-supervised graph neural network for particle level pileup mitigation was previously tested using CMS fast simulation data. The idea is to connect the training samples (labeled) and testing samples (unlabeled) as nodes in a graph using tracking and physics information. In addition, a dedicated masking technique was applied to reduce bias. The model was retrained using CMS full simulation which introduced more complexity in geometry and consequently graph neighbor construction. This increase in the complexity of geometry necessitated a more complex masking technique of particle labels. To use hyperparameter tuning in conjunction with these masking parameters, we introduce a bayesian optimization framework that aims to minimize the performance metric: $\sigma_{\frac{1-|\mu|}{}} for validation datasets. An improved performance of the reconstructed physics variables is obtained. The current results from the Semi-Supervised Graph Neural Network outperform the baseline pileup mitigational gorithm PUPPI.$

Working dinner / **55**

PyLog-HLS4ML Integration: Introducing higher level of automatic design in HLS4ML

Author: Jialiang Zhang[None]

Corresponding Author: jz23@illinois.edu

HLS4ML is an influential Python package that creates firmware implementations of machine learning algorithms uses high-level synthesis (HLS) technique. While most of the templates are handwritten in HLS4ML, we want to further automate this manual design process by introducing PyLog, an algorithm-centric Python-based FPGA programming and synthesis flow, into the current HLS4ML flow and therefore providing a more efficient design pathway as well as initial design space explorations while maintaining the same level of performance compared to the original manual design. We hope this effort could help improve the scalability as well as the usability of HLS4ML.

Working dinner / **56**

Analog-Domain Implementation of Neural Networks for Energy-Efficient High Energy Physics Applications

Author: Dewen Zhong[1]

[1] *Univ. Illinois at Urbana Champaign (US)*

Corresponding Author: dewen.zhong@cern.ch

Efficient computational strategies are paramount for devices in resource-limited settings, particularly within high-energy physics experiments. To address this, we propose research primarily focused on improved energy efficiency and reduced latency inherent to AI algorithms implemented with analog circuits such as memristive crossbar arrays that perform in-memory matrix-vector multiply computations for Artificial Neural Networks (ANN) as compared with digital Al (e.g. on FPGAs). This study proposes the customization of a small ANN model and within the HLS4ML framework to transition quantized NN models into the analog domain. We have created an 8-T unit Cell that employs a compute-in-memory SRAM cell for multi-bit precision interference, featuring memory-integrated data conversion and multiplication-free operations. This investigation will offer comprehensive insights into the performance of the Analog AI model and opens up the possibility of extending the analog AI model to more complex and larger models. We are also exploring HEP applications for this technology, including current and future colliders and experiments.

Special Event / **57**

Affiliate Flash Talks: Xiangyang Ju

Corresponding Author: xiangyang.ju@cern.ch

Special Event / **58**

Affiliate Flash Talks: Ben Carlson

Corresponding Author: bcarlson@cern.ch

Special Event / **59**

Affiliate Flash Talks: Ari Sravan

Corresponding Author: niharika.sravan@gmail.com

Special Event / **60**

Affiliate Flash Talks: Dylan Rankin

Corresponding Author: dylan.sheldon.rankin@cern.ch

Special Event / **61**

Affiliate Flash Talks: Bo-Cheng Lai

Corresponding Author: bclai@nycu.edu.tw

Special Event / **62**

CENPA nuclear physics lab tour (Prof. Alejandro Garcia)

Corresponding Author: agarcia3@uw.edu

Lab tour - CENPA nuclear physics lab tour (Prof. Alejandro Garcia)

Working dinner / **63**

Multi-objective Bayesian Optimization for High-resolution Electron Ptychography

Author: Desheng Ma[None]

Co-author: David Muller [1]

[1] *Cornell University*

Corresponding Authors: dm852@cornell.edu, dm24@cornell.edu

Electron ptychography enables deep sub-angstrom spatial resolution of atomic structures by solving the inverse problem of electron scattering through the sample, provided the complete distribution of transmitted electrons enabled by a new generation of detectors for scanning transmission electron microscopy (STEM). However, in practice, ptychographic reconstructions are computationally intensive and require a delicate selection of both experimental and algorithmic parameters, heavily relying on the trials and errors of user experiences. Here we propose an automatic parameter selection scheme based on multi-objective Bayesian optimization. We show that introducing Fourier ring correlation (FRC) as an additional objective in addition to the Fourier reconstruction error captures the subtle differences in resolution and circumvents unphysical local minima. Instead of a single

optimum produced by a black box optimization, the multi-objective framework produces a Pareto front of equally feasible solutions that lead to the best reconstruction result.

2023_A3D3_dm852.pdf

Working dinner / **64**

Real-time Fitting and Materials Characterization in Band-Excitation Piezoresponse Force Microscopy

Authors: Alibek Kaliyev[1]; Amir Gholami[None]; Joshua Agar[None]; Martin Takáč[None]; Michael Mahoney[None]; Nhan Tran[2]; Pedro Sales[3]; Philip Coleman Harris[4]; Rama Vasudevan[5]; Ryan Forelli[1]; Seda Memik[6]; Shuyu Qin[None]; Stephen Jesse[None]; Veronica Obute [7]; Yael Passy[None]; Yichen Guo[None]

[1] *Lehigh University*

[2] *Fermi National Accelerator Lab. (US)*

[3] *Massachusetts Institute of Technology*

[4] *Massachusetts Inst. of Technology (US)*

[5] *Oak Ridge National Laboratory*

[6] *Northwestern University*

[7] *Drexel University*

Corresponding Authors: jca92@drexel.edu, yig319@lehigh.edu, shq219@lehigh.edu, mo623@drexel.edu, ntran@fnal.gov, philip.coleman.harris@cern.ch, rama87@gmail.com, psales@mit.edu, alk224@lehigh.edu, rff224@lehigh.edu, seda@northwestern.edu

Increased development and utilization of multimodal scanning probe microscopy (SPM) and spectroscopy techniques have led to an orders-of-magnitude increase in the volume, velocity, and variety of collected data. While larger datasets have certain advantages, practical challenges arise from their increased complexity including the extraction and analysis of actionable scientific information. In recent years, there has been an increase in the application of machine and deep learning techniques that use batching and stochastic methods to regularize statistical models to execute functions or aid in scientific discovery and interpretation. While this powerful method has been applied in a variety of imaging systems (e.g., SPM, electron microscopy, etc.), simplistic analysis alone takes on the order of weeks to months due to scheduling and IO overhead imposed by GPU and CPU based systems which limits streaming inference rates to speeds above 50ms. This latency precludes the possibility of real-time analysis in SPM techniques such as band-excitation piezoresponse force spectroscopy (BE PFM), where typical measurements of cantilever resonance occur at 64Hz.
One method to accelerate machine learning inference is to bring computational resources as close to the data acquisition source as possible to minimize latencies associated with I/O and scheduling. Therefore, we leverage the National Instruments PXI platform to establish a direct, peer-to-peer channel over PCIe between an analog-to-digital converter and a Xilinx field programmable gate array (FPGA). Through the LabVIEW FPGA design suite, we develop this FPGA-based pipeline using cantilever resonances acquired in BE PFM to conduct real-time prediction of the simple harmonic oscillator (SHO) fit. To accomplish this, we use hls4ml to compile a high-level synthesis (HLS) representation of the neural network. Once this HLS model is synthesized to a register transfer level description (RTL), we implement the design on the FPGAs programmable logic. The parallelizable nature of FPGAs allows for heavily pipelined neural network implementations to achieve latencies on the order of microseconds. We currently benchmark our implementation at 36us per inference with a fourier transformation accounting for an additional 330us. At the expense of FPGA resources, we overlap data acquisition with computation to enable continuous acquisition and processing of response data. Overall, this work provides a foundation for deploying on-sensor neural networks using specialty hardware for real-time analysis and control of materials imaging systems.

Working dinner / **65**

Machine learning evaluation in the Global Event Processor FPGA for the ATLAS Phase 2 Level 0 trigger upgrade

Authors: Ben Carlson[1]; Boping Chen[2]; Bowen Zuo[3]; Jeff Eastlack[4]; Liron Barak[5]; Santosh Parajuli[6]; Scott Hauck[None]; Shih-Chieh Hsu[7]; Zhixing "Ethan" Jiang[3]

Co-author: Elham E Khoda [3]

[1] *Westmont College*

[2] *Tel Aviv University (IL)*

[3] *University of Washington (US)*

[4] *Michigan State University (US)*

[5] *Tel Aviv University*

[6] *Southern Methodist University (US)*

[7] *University of Washington Seattle (US)*

Corresponding Authors: zhixing.jiang@cern.ch, schsu@uw.edu, hauck@uw.edu, elham.e.khoda@cern.ch, jeff.eastlack@cern.ch, bowen.zuo@cern.ch, santosh.parajuli@cern.ch, boping.chen@cern.ch, lironbarak83@gmail.com, bcarlson@cern.ch

During the next update of the High-Luminosity Large Hadron Collider (HL-LHC) of ATLAS, a new global trigger subsystem will be installed into the L0 Trigger. New and improved hardware and algorithms will be deployed during the upgrade to increase the performance of the trigger system. The global trigger subsystem consists of various components, including the FPGA-based Global Event Processor (GEP), which processes the data through the trigger algorithm. Within the GEP, data will be pipelining through different Algorithm Processing Units (APU), which handle individual subtasks of the overall trigger. We present our work in creating an APU specification and sample APU as a guide to future APU developers. We also present a redesign of the APU interface to follow the AXI-stream protocol, which allows streaming computations that overlap operations at multiple pipeline levels, potentially improving overall throughput. Finally, we present our work deploying HLS4ML and FwX (two high-level synthesis tools for machine learning) into the APU development. HLS4ML is a design tool for generating a deep neural network (DNN) algorithm model with ultra-low delays, and has been developed specifically to support the needs of high-energy physics experiments, while FwX is another tool for generating the boost decision tree models. Our goal is to demonstrate that the application of HLS4ML to APU development is practical, and we have already implemented Gluon Tagger model using convolution neural network (CNN) for the APU. The performance of the Gluon Tagger APU is tested using a test vehicle and a sandbox provided by the ATLAS developers. The next step is developing another new algorithm using a deep neural network.

Working dinner / 66

Graph Neural Network-based particle tracking as a Service

Authors: Andrew Naylor[1]; Dylan Sheldon Rankin[2]; Elham E Khoda[3]; Paolo Calafiura[4]; Shih-Chieh Hsu[5]; Steven Farrell[6]; Xiangyang Ju[4]

[1] *Lawrence Berkeley National Lab*

[2] *University of Pennsylvania (US)*

[3] *University of Washington (US)*

[4] *Lawrence Berkeley National Lab. (US)*

[5] *University of Washington Seattle (US)*

[6] *Lawrence Berkeley National Laboratory*

Corresponding Authors: pcalafiura@lbl.gov, xiangyang.ju@cern.ch, elham.e.khoda@cern.ch, anaylor@lbl.gov, dylan.sheldon.rankin@cern.ch, sfarrell@lbl.gov, schsu@uw.edu

Recent studies on the ITk data showed that the Graph Neural Network (GNN) -based track finding can provide not only satisfied track efficiency but also reasonable track resolutions. However, the

GNN-based track finding is computationally slow in CPUs, demanding the usage of coprocessors like GPUs to speed up the inference time. The large graph size, normally 300k nodes and 1M edges, necessitates significant GPU memory for feasible computation. Not all ATLAS computing sites are harnessed with high-end GPUs like A100s. These challenges have to be addressed in order to deploy the GNN-based track finding into production. We propose to address these challenges by establishing the GNN-based track-finding algorithm as a service hosted either in clouds or high-performance computing centers.

In this poster, we will describe the implementation of the GNN-based track-finding workflow as a service using the Nvidia Triton inference server. The pipeline contains three discrete deep-learning models and two CUDA-based algorithms. Because of the heterogeneity in the workflow, we explore different server settings to maximize the throughput of track finding. At the same time, we study the scalability of the inference server using the Perlmutter supercomputer at NERSC and cloud resources like AWS and Google Cloud. We will present the studies performed with the stand-alone algorithm. Integration and optimization of the workflows into ACTS and Athena are in progress.

Working dinner / **67**

Interaction Networks for Anomaly Detection at the CMS Level-1 Trigger

Author: Andrew Skivington[1]

[1] *University of California-San Diego, Duarte Lab*

Corresponding Author: askivington@ucsd.edu

At the LHC proton bunches are collided at a rate of 40MHz. The Compact Muon Superconducting Solenoid (CMS) detector's Level-1 (L1) trigger system is responsible for reducing this data rate to about 100kHz so that approximately 1% of these events can be saved for offline physics analyses. The task is to develop algorithms to determine what data to keep and what to discard. Traditionally, trigger algorithms are physics inspired data selections based on corresponding hard-coded high-level features. These methods require a large amount of data preprocessing and potentially bias us to discard interesting physics. Hence, there is a need for theory independent anomaly detection (AD) algorithms. AD algorithms have the potetial to detect "unknown, unknowns."In the context of high energy particle physics, these anomalies could come in the form of undefined intermediate decay states, or even something as simple as a detector flaw. Either way, AD algorithms can increase the physics we collect from our detector. This is critical for the coming era of the high luminosity LHC (HL-LHC), so that we can further probe the standard model (SM) and increase the trigger sensitivity for beyond the standard model (BSM) physics. This study explores two different graph neural networks (GNNs), specifically interaction networks (INs). The IN architecture was adapted to both autoencoder (AE) and variational autoencoder (VAE) models. The models are known as INAE and INVAE models, respectfully. Each model was trained and tested on the respective LHC ADC2021 datasets. The results were compared to the current benchmark DNN (Deep Neural Network) AE and VAE architectures for the task of anomaly detection at the L1 trigger. The goal of the study was to determine if INs can be applied as future AD algorithms at the trigger level. Various performance metrics, such as L1 physics reconstruction tasks for representative SM and BSM signals at several different trigger thresholds were used in the analysis of the IN's viability for anomaly detection and are further discussed in the results.

Working dinner / **68**

Benchmarking HLS4ML vs. SystemVerilog

Authors: Caroline Johnson[None]; Waiz Khan[None]

Corresponding Authors: wkhan@uw.edu, cjj9@uw.edu

High-level synthesis (HLS) offers the promise of simpler and easier hardware development, but at a cost. We consider the application of high-level synthesis to machine learning applications, seeking to quantify the resource and performance costs of this technique within the widely used HLS4ML framework. By creating carefully optimized SystemVerilog versions of identical HLS4ML designs, we demonstrate that the HLS designs are very competitive with hand-optimization techniques. We also identify weaknesses in the existing tools, and develop work-arounds to help provide significant quality improvements.

Working dinner / **69**

Jet Tagging Algorithm for Long-Lived Particles at the CMS Level-1 Trigger

Authors: Anthony Vizcaino Aportela[1]; Russell Denilson Marroquin Solares[1]

[1] *Univ. of California San Diego (US)*

Corresponding Authors: rmarroquinsolares@ucsd.edu, aaportel@ucsd.edu

This poster presents an exploration into the realm of Beyond the Standard Model (BSM) Long-Lived Particles (LLPs) with a focus on the integration and development of a jet-tagging algorithm for the Compact Muon Solenoid (CMS) experiment's Level 1 (L1) Trigger system, in the context of the forthcoming upgrade to the High Luminosity Large Hadron Collider (HL-LHC). In spite of the challenges posed by traditional particle collision data analyses at the LHC, involving variable thresholds such as pT, HT, and displaced vertices, which have proven insufficiently sensitive to LLP signatures, we propose a pathway beyond these limitations. The conventional L1 triggers, currently designed to select tracks from events decaying near the collision vertex, face difficulties arising from pileup interactions, hence impeding discovery of novel BSM signals. The solution we present involves a Deep-Neural Network (DNN) based jet-tagging algorithm specifically designed to amplify the tagging efficiency of LLP signatures and considerably attenuate pileup effects. As we progress into an era characterized by increased collision rates and voluminous data due to the HL-LHC upgrade, the need for faster and more efficient real-time event selection algorithms becomes imperative. Drawing lessons from prior studies on b-tagging and tau-tagging neural networks, we endeavor to optimize this jet-tagging algorithm using tools such as the hls4ml Python library. Our work also includes application of optimization techniques to fulfill stringent timing requisites and the validation of the jet-tagging algorithm's performance using simulated collision data.

Working dinner / **70**

A3D3 Equity & Career Activities

Authors: Anthony Vizcaino Aportela[1]; Elham E Khoda[2]; Janina Hakenmüller[3]; Javier Mauricio Duarte[1]; Mark Neubauer[4]; Melissa Quinnan[1]; Noah Paladino[5]

[1] *Univ. of California San Diego (US)*

[2] *University of Washington (US)*

[3] *Duke University*

[4] *Univ. Illinois at Urbana Champaign (US)*

[5] *Massachusetts Inst. of Technology (US)*

Corresponding Authors: aaportel@ucsd.edu, javier.mauricio.duarte@cern.ch, msn@illinois.edu, melissa.quinnan@cern.ch, elham.e.khoda@cern.ch, janina.hakenmuller@duke.edu, npaladin@mit.edu

The Accelerated AI Algorithms for Data-Driven Discovery (A3D3) Institute, funded by the National Science Foundation (NSF), under the Harnessing the Data Revolution (HDR) program, is a multi-disciplinary and geographically distributed entity with the primary mission to lead a paradigm shift

in the application of real-time artificial intelligence (AI) at scale to advance scientific knowledge and accelerate discovery in particle physics, astrophysics, biology, and neuroscience.

We will describe the activities of the A3D3 Equity & Career Committee, including the A3D3 post-baccalaureate research fellowship aimed at increasing participation in research from traditionally underrepresented groups in STEM, the mentoring program, the code of conduct, workshops, and other activities.

Research Subgroup AI Tools and Developments Talks / **71**

Kate Scholberg Group (Duke)

Corresponding Author: janina.hakenmuller@duke.edu

Janina's flight got delayed. So, she will present later in the afternoon

Hackathon / **72**

Introduction & Overview

Corresponding Author: melissa.quinnan@cern.ch

hackathon is introduced and helpers are introduced

Hackathon / **73**

Topic 3: HEP Trigger Anomaly Detection

Corresponding Authors: elham.e.khoda@cern.ch, melissa.quinnan@cern.ch

Hackathon / **74**

Topic 2: MMA telescope (ZTF) source classification

Corresponding Author: healyb@umn.edu

Hackathon / **75**

Topic 1: Neuroscience mouse touch stimulus brain signals

Corresponding Authors: liptonm@purdue.edu, park1377@purdue.edu

Hackathon / **76**

Group Formation/Go to breakout rooms

Based on participant desires and group needs people are split into 3 projects

Hackathon / **77**

Helper Introductions

people who have volunteered to help out with all hackathon groups introduce themselves and their areas of expertise

Hackathon / **78**

HEP breakout session

Hackathon / **79**

MMA breakout session

Hackathon / **80**

Neuro breakout session

Working dinner / **81**

GWAK: Gravitational-Wave Anomalous Knowledge

Authors: Alec Gunny[None]; Deep Chatterjee[None]; Dylan Sheldon Rankin[1]; Eric Anton Moreno[2]; Erik Katsavounidis[3]; Ethan Marx[3]; Katya Govorkova[4]; Michael Coughlin[5]; Philip Coleman Harris[4]; Ryan Raikman[None]; Will Benoit[None]

[1] *University of Pennsylvania (US)*

[2] *Massachusetts Institute of Technology (US)*

[3] *MIT*

[4] *Massachusetts Inst. of Technology (US)*

[5] *University of Minnesota*

Corresponding Authors: eric.anton.moreno@cern.ch, alecg@mit.edu, michael.w.coughlin@gmail.com, ekaterina.govorkova@cern.ch, emarx@mit.edu, kats@mit.edu, deep1018@mit.edu, dylan.sheldon.rankin@cern.ch, benoi090@umn.edu, philip.coleman.harris@cern.ch

Matched-filtering detection techniques for gravitational-wave (GW) signals in ground-based interferometers rely on having well-modeled templates of the GW emission.

However, interesting science cases aside from compact mergers do not yet have accurate enough modeling to make matched filtering possible, including core-collapse supernovae and other stochastic sources. Therefore the development of techniques to identify sources of these types is of significant interest. We present a method of anomaly detection techniques based on deep recurrent autoencoders to enhance the search region to unmodeled transients. We use a semi-supervised strategy named Gravitational Wave Anomalous Knowledge (GWAK). While the semi-supervised nature of the problem comes with a cost in terms of accuracy as compared to supervised techniques, there is a qualitative advantage in generalizing experimental sensitivity beyond pre-computed signal templates. We construct a low-dimensional embedded space using the GWAK method, capturing the physical signatures of distinct signals on each axis of the space. By introducing alternative signal priors that capture some of the salient features of gravitational-wave signals, we allow for the recovery of sensitivity even when an unmodeled anomaly is encountered. We show that regions of the GWAK space can identify compact binary coalescences, detector glitches and also a variety of unmodeled astrophysical sources.

Working dinner / **82**

Real-Time AI for the Particle Flow and PUPPI Algorithm in the CMS Level-1 Trigger Upgrade

Authors: Aidan Chambers[1]; Duc Minh Hoang[2]; Noah Paladino[1]; Orion Foo[None]; Philip Coleman Harris[1]

[1] *Massachusetts Inst. of Technology (US)*

[2] *MIT*

Corresponding Authors: ofoo@mit.edu, npaladin@mit.edu, aidandc@mit.edu, dhoang@mit.edu, philip.coleman.harris@cern.ch

The Particle Flow algorithm has proven highly effective in the offline reconstruction of events in the CMS detector. Combined with Pile-Up Per Particle Identification (PUPPI), the two algorithms provide the necessary basis for the construction of higher-level physics options, such as jets and taus. With the upcoming High Luminosity upgrade of the Large Hadron Collider (HL-LHC), implementing the PF and PUPPI algorithms in the Level-1 (L1) trigger has become a way to significantly improve trigger performance. The integration of these elements in the L1 trigger allows for the implementation of machine learning algorithms, such as b-tagging and tau-tagging, which can be implemented on the FPGAs using the hls4ml framework. This allows for greater sensitivity to multiple signals, including di-Higgs events, which are essential for measuring the Higgs self-coupling, by resulting in a sharper and earlier trigger turn-on as well as increased signal acceptance.

Working dinner / **83**

Multi-block RNN Autoencoders Enable Broadband ECoG Signal Reconstruction

Authors: Michael Nolan[None]; Bijan Pesaran[1]; JINGYUAN LI[None]; Eli Shlizerman[None]; Amy Orsborn[None]

[1] *New York University*

Corresponding Authors: manolan@uw.edu, aorsborn@uw.edu, bp31@nyu.edu, shlizee@uw.edu, jingyli6@uw.edu

Neural dynamical models reconstruct neural data using
dynamical systems. These models enable direct reconstruction and estimation of neural time-series data as well as estimation of neural latent states. Nonlinear neural dynamical models using recurrent neural networks in an encoder-decoder architecture have recently enabled accurate single-trial reconstructions of neural activity for neuronal spiking data. While these models have been applied to neural field potential data, they have only so far been applied to signal feature reconstruction (e.g. frequency band power), and have not yet produced direct reconstructions of broadband timeseries

data preserving signal phase and temporal resolution. Approach. Here we present two encoder-decoder model architectures - the RNN autoencoder (RAE) and multi-block RAE (MRAE) for direct time-series reconstruction of broadband neural
data. We trained and tested models on multi-channel microElectricorticography (μECoG) recordings from non-human primate motor corticies during unconstrained behavior. Main Results. We show that RAE reconstructs micro-electrocorticography recordings, but has reconstruction accuracy that is band-limited to model scale. The MRAE architecture overcomes these time-bandwidth restrictions, yielding broadband (0-100 Hz), accurate reconstructions of μECoG data. Significance. RAE and MRAE reconstruct broadband μECoG data through multiblock dynamical modeling. The MRAE overcomes time-bandwitdh restrictions to provide improved accuracy for long time duration signals. The reconstruction capabilities provided by these models for broadband neural signals like μECoG may enable the development of improved tools and analysis for basic scientific research and applications like brain-computer interfaces.

Working dinner / **84**

Graph Neural Networks for Electron and Photon Reconstruction at CMS

Author: Simon Rothman[1]

[1] *Massachusetts Inst. of Technology (US)*

Corresponding Author: srothman@mit.edu

The Compact Muon Solenoid (CMS) detector is one of two general-purpose detectors at
the CERN LHC. Products of proton-proton collisions at a center of mass energy of 13 TeV are reconstructed in the CMS detector to probe the standard model of particle physics and to search for processes beyond the standard model. The development
of precision algorithms for this reconstruction is therefore a key objective in optimizing the precision of all physics results at CMS. While the use of machine learning techniques are now prevalent at CMS for these tasks, they have largely relied on high-level human-engineered input features. However, much of the disruptive impact of machine learning in industry has been realized by bypassing human feature engineering and instead training deep learning algorithms on low-level data. We have
developed a novel machine learning architecture based on dynamic graph neural networks that allows regression directly on low-level detector hits and applied this model to photon energy corrections in CMS. In this work, the performance of our new architecture is demonstrated in the corrections to the the energies of the photons that are used in most analyses at CMS, where we obtain an improvement in energy resolution by a factor of 10\% with respect to the previous state-of-the-art reconstruction method.

Special Event / **85**

Meet outside physics building to walk over to CENPA

Hackathon / **86**

Neuro Hackathon Presentation

Hackathon / **87**

MMA Hackathon Presentation

Corresponding Author: healyb@umn.edu

Hackathon / **88**

HEP Hackathon Presentation