Liu Group@Purdue Status Report

Mia Liu, Jan Schulte, Dmitry Kondratyev, Lisa Papalaki

A3D3 High-Throughput AI Methods and Infrastructure Workshop
July 10-14 2023
https://indico.cern.ch/event/1282754

OAC-2117997

https://a3d3.ai/

The Team

Mia Liu, PI

Lisa Paspalaki,
PostDoc

Jack Rodgers,
Undergraduate
student

Dmitry Kondratyev,
Research Software
Engineer

Yibo Zhong
PhD student

Jan Schulte,
Research Scientist

Hyeon-Seo Yun,
Master student

Benjamin Simon,
PhD student

+ Yao Yao,
New PostDoc
joining very
soon!

Purdue activities in A3D3

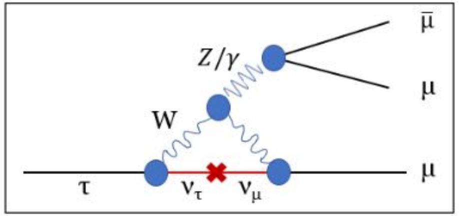
- We are developing **machine learning algorithms** combined with **heterogeneous computing** within each of the three reconstruction tiers of the **CMS** detector, the **L1 Trigger**, the **High Level Trigger**, and **offline reconstruction**.
  - **End-to-end GNN** triggers for rare tau lepton decays
  - Improved object reconstruction offline using GNNs: Semi-Supervised pileup mitigation
  - Close collaboration with Prof. Pan Li's group in ML for science development: i**nterpretable GNN, domain adaptation, end to end efficient GNN**
- Involved in development and maintenance of **software toolkits** that enable the deployment of these algorithms into the existing software and hardware systems of CMS.
  - **HLS4ML**: deployment of GNN on FPGAs for low latency,
  - **SONIC** : deployment and integration in cms distributed computing infrastructure

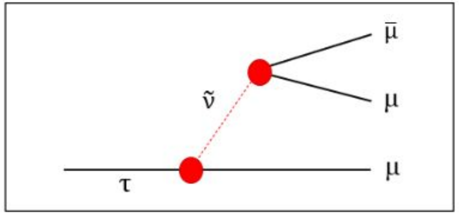
Accelerated AI Algorithms for Data-Driven Discovery

Triggering $\tau\rightarrow3\mu$ Decays with GNNS

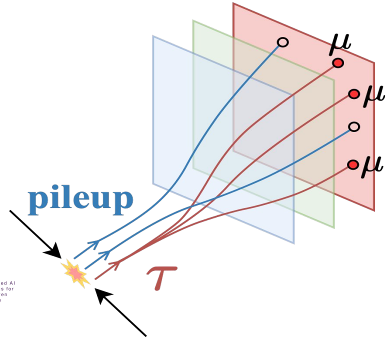
- $\tau\rightarrow3\mu$ decay heavily suppressed in the Standard Model
  - BR ~ $O(10^{-55})$ predicted, current best limits 2.1 x 10^{-8}
  - Many BSM physics models enhance BR($\tau\rightarrow3\mu$)~$O(10^{-8})$
- ~ 1 x 10^{15} τ expected in full HL-HLC dataset
  - **Low transverse momentum**, very forward
  - Very **hard to trigger** with conventional techniques
- Solution: **end-to-end reconstruction** of $\tau\rightarrow3\mu$ topology using GNNs

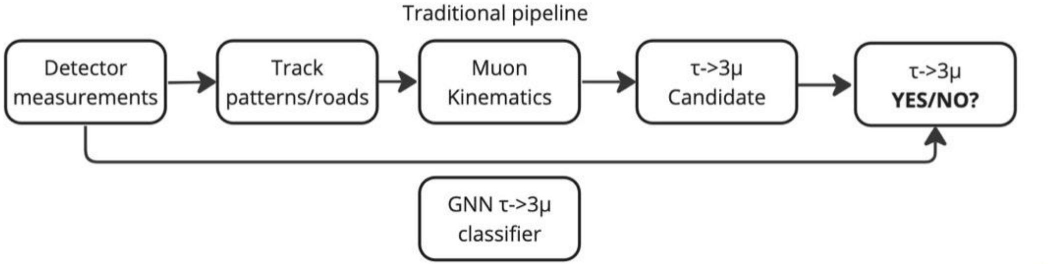
SM w/ Neutrino Osc.
BR ~ $O(10^{-55})$





SUSY w/ R parity violation





B. Simon, H. Yun, Y.Zhong, JS

GNN Graph Construction

CMS muon detectors are arrayed in **4 stations** that muons traverse from the inside-out
Information from detectors within one station are aggregated into track segments we use as nodes for the graph. **First three stations** are used.

Edge Formation:

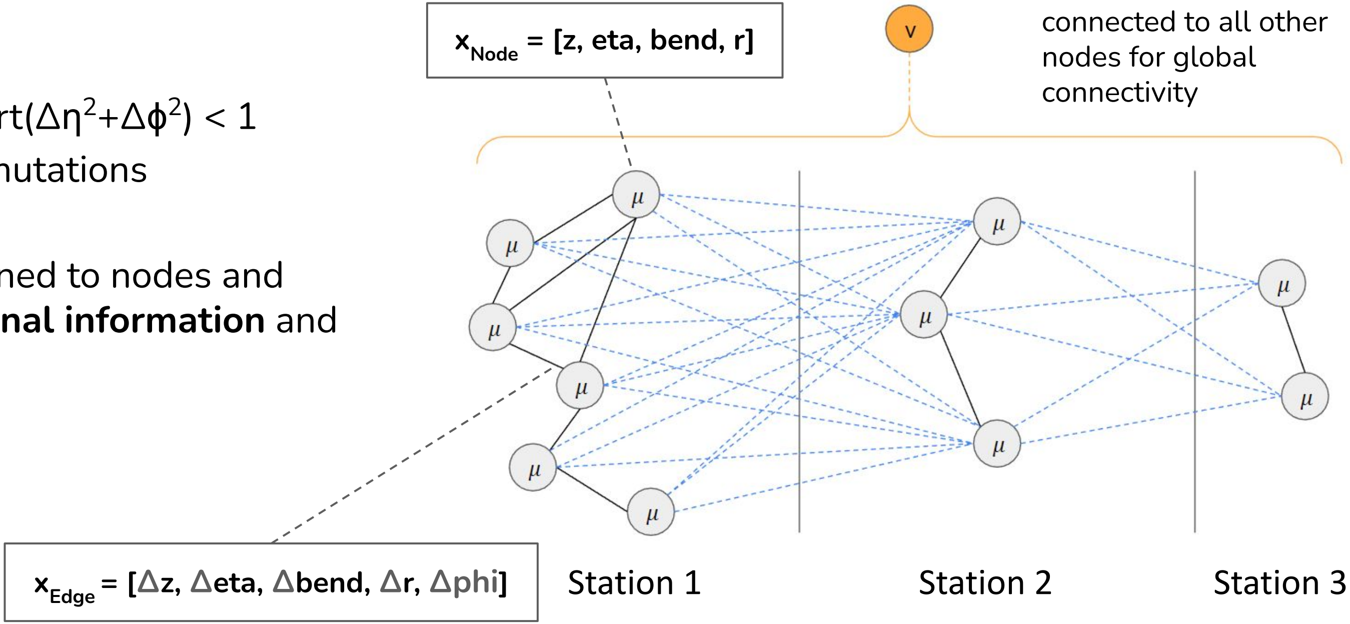
Intra-station: dR = sqrt($\Delta\eta^2+\Delta\phi^2$) < 1

Inter-station: All permutations

Features:

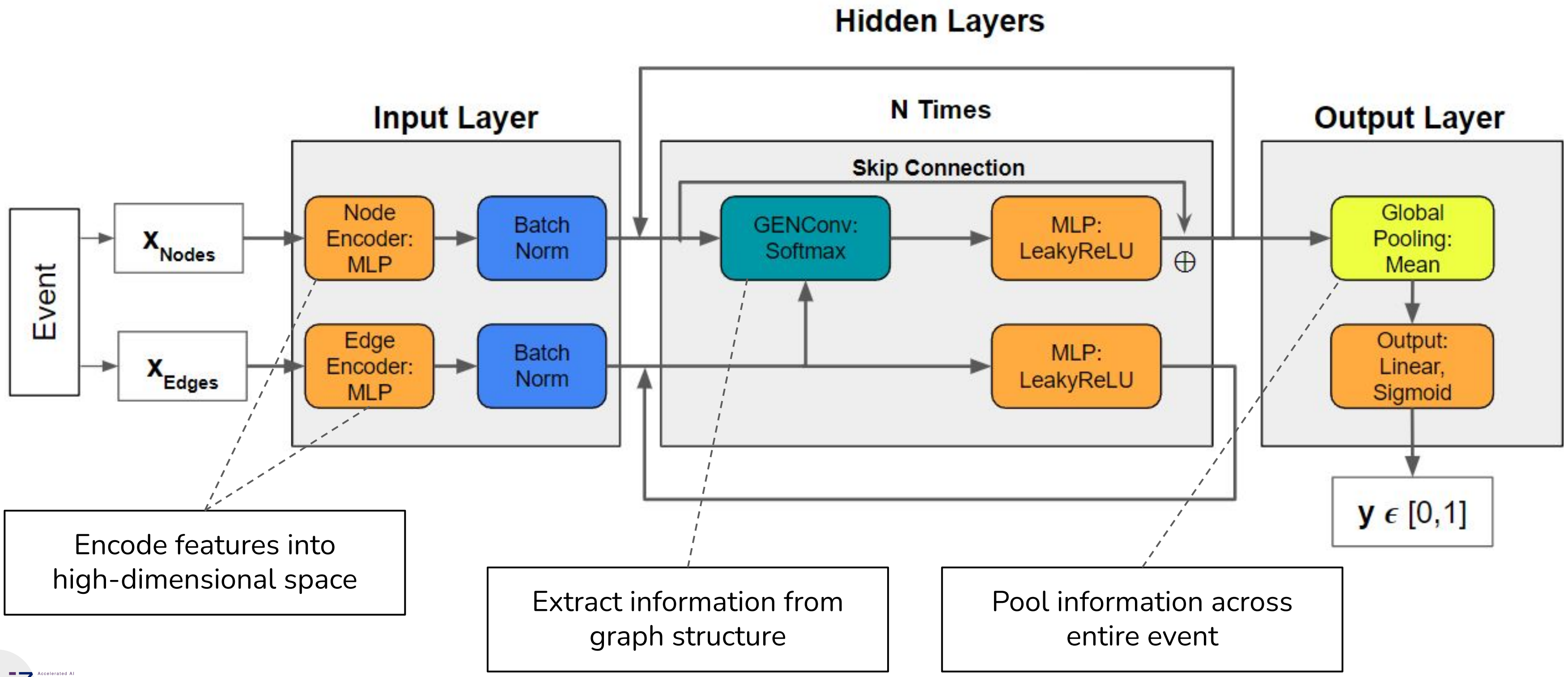
Feature vectors assigned to nodes and edges encode **positional information** and **bending angle**

Virtual Node: connected to all other nodes for global connectivity



B. Simon, H. Yun, Y.Zhong, JS

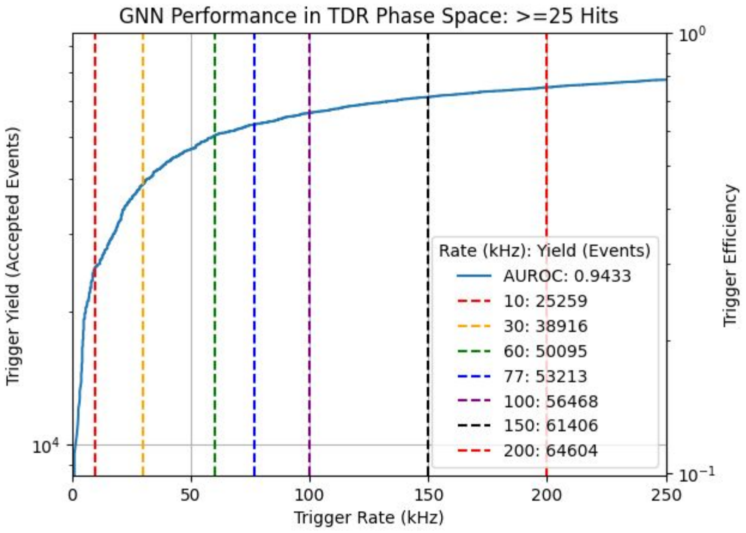
GNN Method: Model Architecture

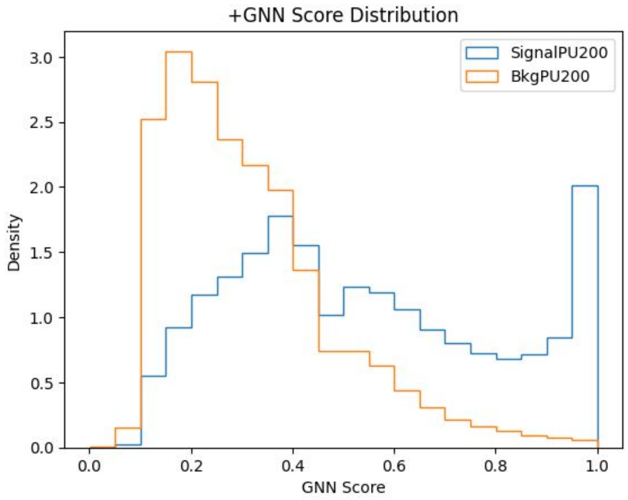


B. Simon, H. Yun, Y.Zhong, JS

Current model performance

- GNN is able to confidently **separate** part of the **signal** phase space **from background**
- Signal events with **very few nodes** almost **indistinguishable** from background





- **Preselection on the number of nodes** allows to allocate **trigger bandwidth** only to events where we are confident in the trigger decision
- Signal acceptance up to x10 larger than expected for traditional techniques: likely surpass Bell 2 that has current world best limit projection.

- For details, see **Ben's poster** tonight!
- Future plans: Study **more signals**, **anomaly detection**, implement model on **L1 demonstrator** at Purdue

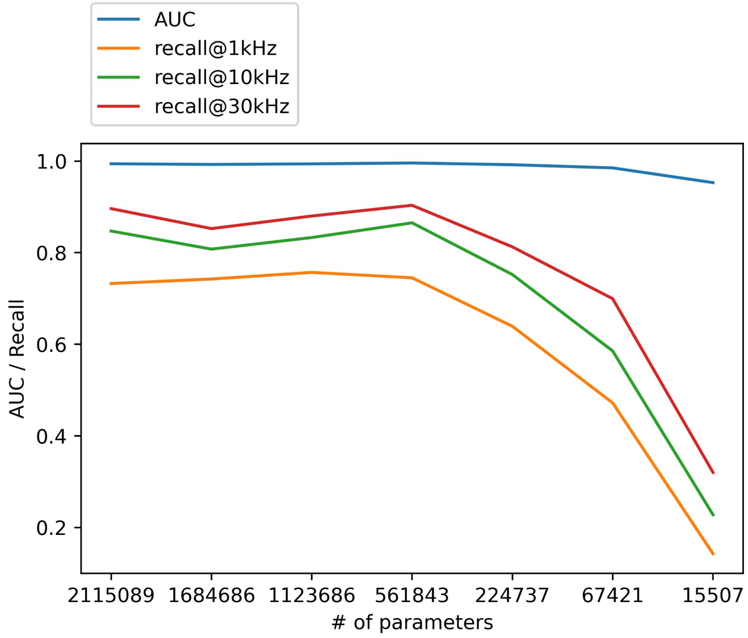
B. Simon, H. Yun, Y.Zhong, JS

FPGA implementation

- Model architecture optimized for best performance
  - In the **trigger** it will have to run on **FPGA** with tight latency constraints

- Current model is simply **too large**, have to reduce complexity
  - Investigating **pruning** and **quantization**
  - Found we can **prune ~70% of internal nodes**
  - First tests with quantization aware training in **Brevitas** give promising results:
    92% of AUROC when going from **32bit float** to **8bit ap-fixed** precision

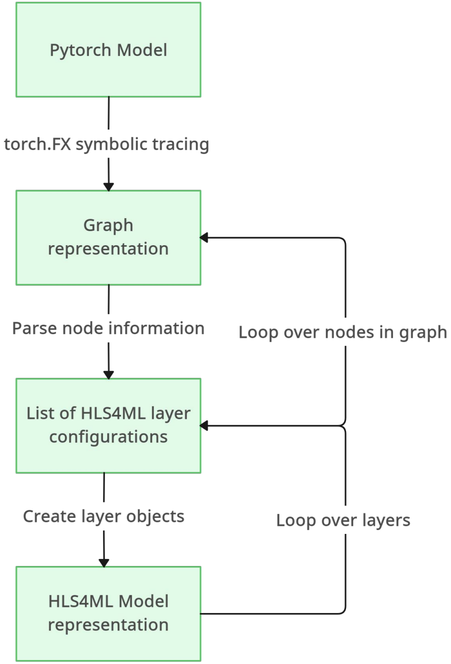
- **HLS4ML** package does currently not support **Pytorch Geometric** models (Pytorch support also limited) -> **GNNs not supported** in general
  - (Model-specific private implementations exist, e.g. by Javier's group)
  - Working with other developers to implement **GNN** support in a more **generalized** way



JS, H. Yun

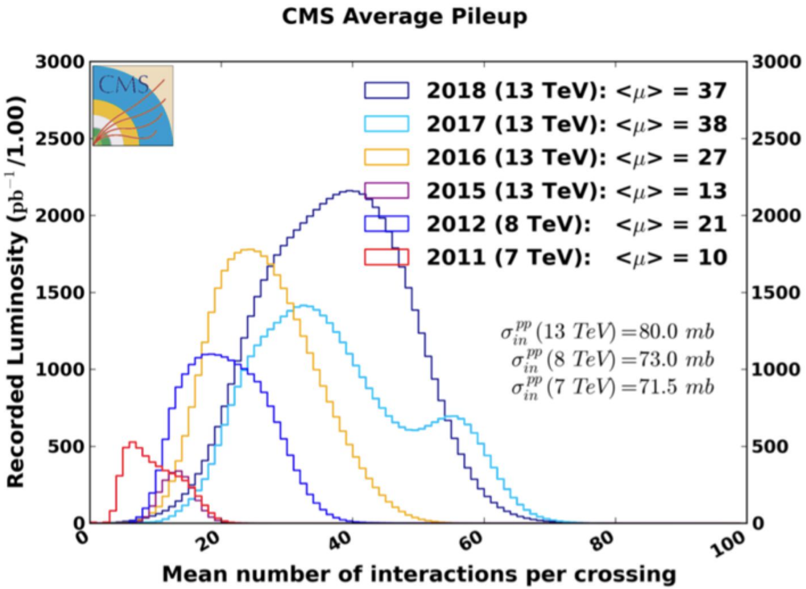
GNN support in HLS4ML

- GNNs typically implemented using **Pytorch Geometric**
  - Limited Pytorch support in HLS4ML prevented implementation of a generalized parser for PyG models
- **Re-implemented parsing of Pytorch models in HLS4ML** using torch.FX symbolic tracing functionality
  - Converts model into a graph with individual layers as node. Can then traverse the graph and pick up layer configurations from the nodes
- Significantly improves **ease of use** and **types of layers supported**. New parser will be part of next major HLS4ML release (v8.0) in Q2(-ish)
- **Extending** this parser **to PyG models**, in collaboration with Vladimir Loncar
  - **Parsing of PyG models**
  - Support for **message passing** operations (hard to parse with symbolic tracing because of nested structure)
  - Support for **new operations** like scatter_add in HLS4ML
- First full prototype will be available at the **end of summer**



JS

Pile Up-Motivation for a semi supervised network



pileup (PU): multiple proton interactions in the same bunch-crossing affect many variables: jet mass, jet pt, missing transverse momentum (MET)

Can reject **charged particles** from PU based on **track information**. Real problem are the neutrals. Current best approach in CMS (**PUPPI**) weights neutrals based on neighboring charged particles

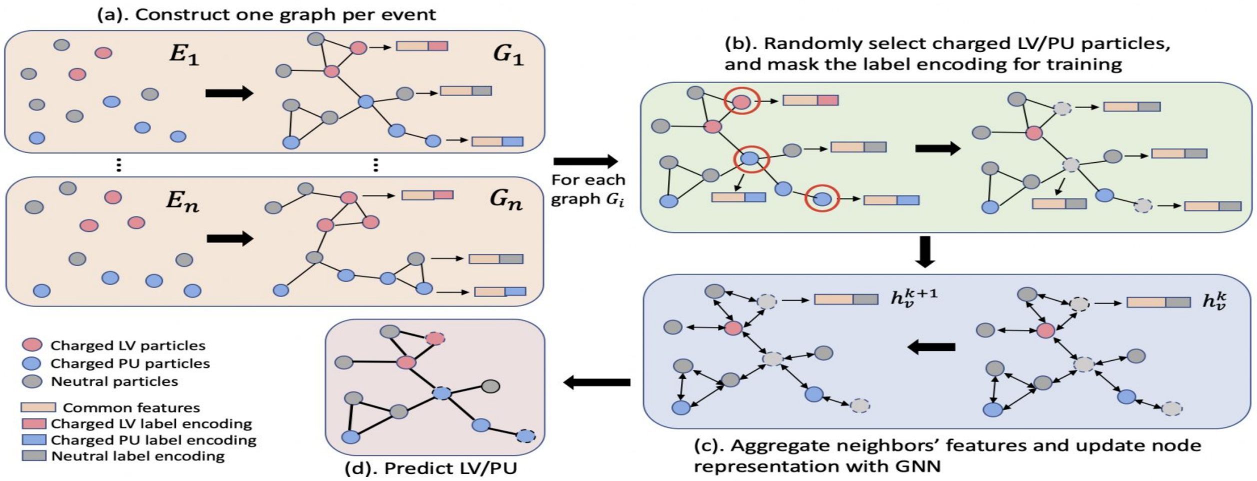
new approach: Graphed based semi-supervised (SSL)

train directly on real data/full simulation, without worrying about the labels for the ground truth information

towards to a new direction of fully data-driven pileup mitigation technique

L. Paspalaki, J. Rodgers

The network



- **Semi-supervised** approach aims to develop an NN for PU reduction
- The Semi-supervision enables the possibility of t**raining on data**
- Graph architecture builds on the rich graph algorithms already shown
  - Means that acceleration of graphs leads to a fast algorithm here

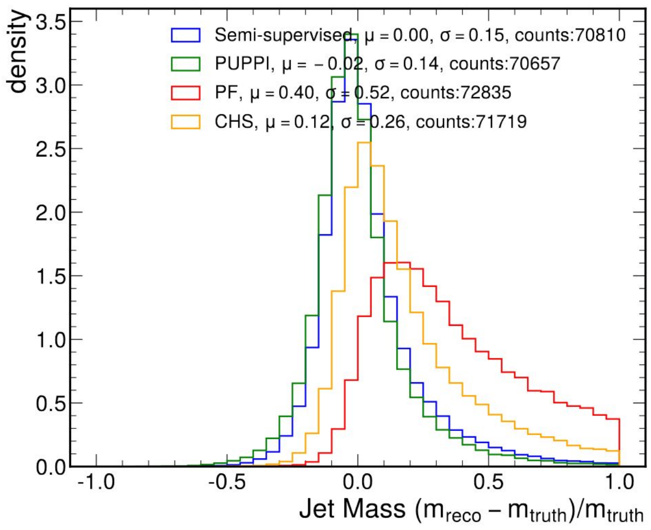
First results on CMS fast simulation: [2203.15823](2203.15823)

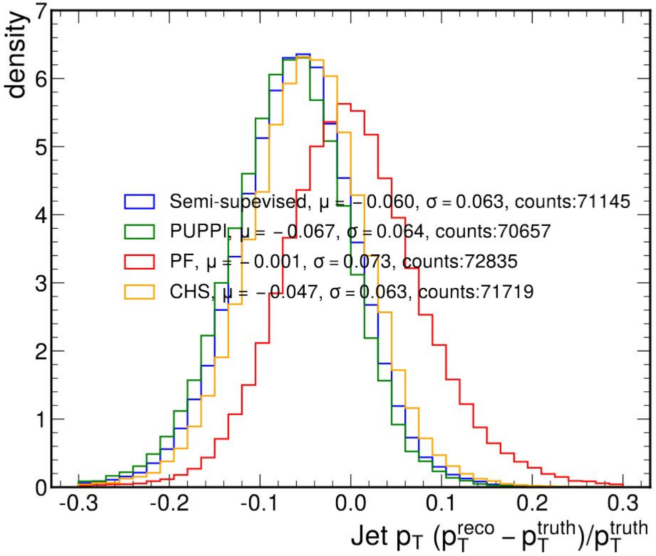
The network is now tested and trained in CMS full simulation

L. Paspalaki, J. Rodgers

Performance on physics variables

- A **Bayesian optimization** framework was developed to optimize the physics performance
  - $\boldsymbol{\sigma}/(1-\boldsymbol{\mu})$ for the jet mass as figure of merit





Puppi-GNN outperforms the baseline PUPPI algorithm

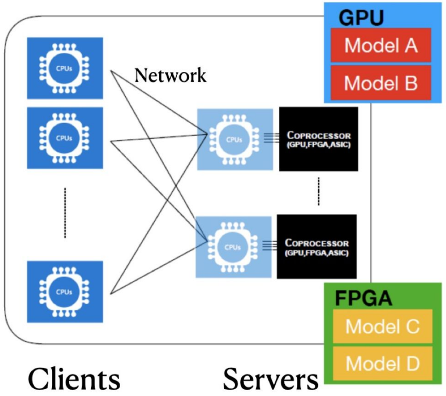
L. Paspalaki, J. Rodgers

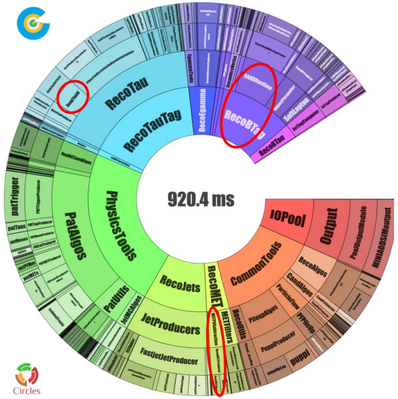
Summary and next steps

- **Puppi-GNN** is **trained and tested** in CMS full simulation
- Bayesian optimization techniques were used to improve the network's performance
- Puppi-GNN **outperforms the baseline** PUPPI algorithm
- **Now:** Domain Adaptation techniques are considered to further improve the performance
- Check out **Jack's poster** tonight for more information!

Future steps-goals: integrate the network in CMSSW and commissioning using Run 3 data

Heterogeneous computing as-a-service





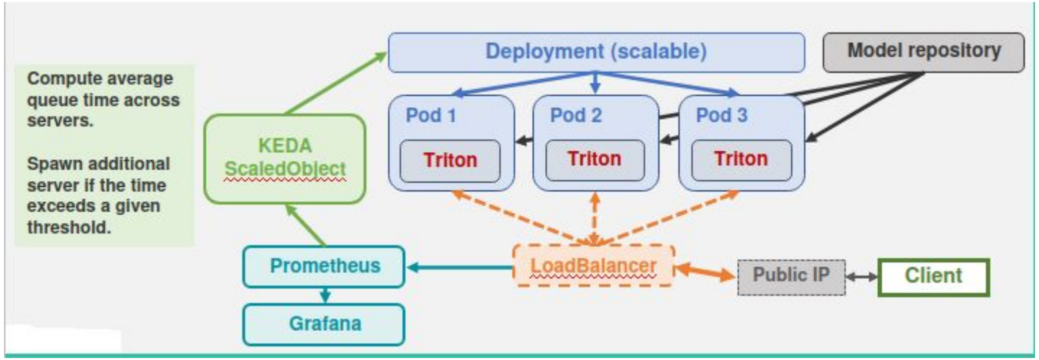
- Providing access to different **accelerators as a service** allows scalable/flexible/modular software stacks
- **SONIC** uses Nvidia **Triton** servers to provide **GPU** resources to CMS software workflows
  - **Developed** and **tested** a miniAOD (one step in CMS data processing) workflow that offloads 3 ML inferences to SONIC.
  - Performance measurements produced on **Purdue computing resources**, CMS paper in preparation
  - Challenge: Have to create interface to SONIC for each ML model or algorithm separately
  - New group members (Yibo, Yao) will join Ben in investigating **automated "sonification"** of workflows

D. Kondratyev

Sonic/Triton infrastructure at Purdue Tier-2 center

- **CMS software** is run on many computing centers worldwide that have to provide **Triton servers** to enable **SONIC** in CMS workflows. Triton servers could also be utilized to enable GPU access for local users.
- Ongoing development efforts: **load balancer**, dynamic **creation / destruction of servers**, service to advertise **available servers** to jobs, treatment of ML model versions / CMS software **versions**
- As an alternative to traditional Tier-2 cluster, we are developing a setup in a **Kubernetes** cluster
  - Run Triton servers as **containerized** applications.
  - Utilize **industry-grade** solutions for our development challenges.
  - Example: **automatic load balancing** setup based on Triton performance metrics adopted from Nvidia's implementation.
  - Ongoing studies: access to **remote GPUs** outside of the Kubernetes cluster, load-balancing across different **types of GPUs**, **model repository** solutions (filesystem mount vs. cloud object storage)

kubernetes



NVIDIA
TRITON INFERENCE SERVER

a3d3 Accelerated AI Algorithms for Data-Driven Discovery

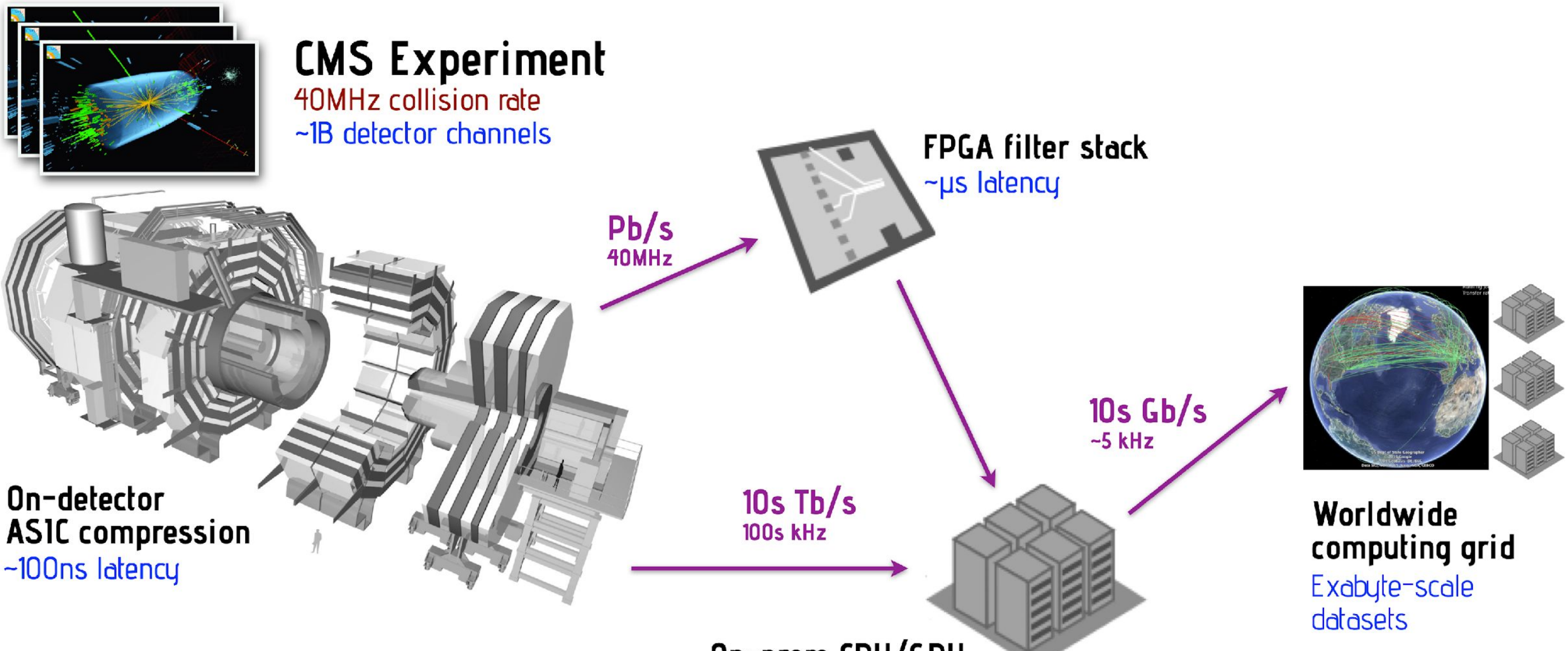
D. Kondratyev

Conclusions

- Our group aim to use ML to improve physics and computational performance at all stages of the data pipeline in CMS, with focus on GNNs
    - End-to-end reconstruction of $\tau \to 3\mu$ decays in the L1 trigger, HLT and offline processing to be studied later
    - Semi-supervised learning to improve pileup mitigation in offline reconstruction workflows
    - GNN support on FPGAs by improving HLS4ML
    - Heterogeneous computing as-a-service for CMS offline workflows using SONIC/TRITON
- Group is continuously expanding, excited to work on many new ideas going forward

Backup

CMS Data Flow



ML in 3 tiers of data processing