Semi Supervised Graph Neural Network for Pileup Noise Removal

Shikun Liu, Tianchun Li, **Garyfallia Paspalaki**, Pan Li, Miaoyuan Liu (Purdue University)

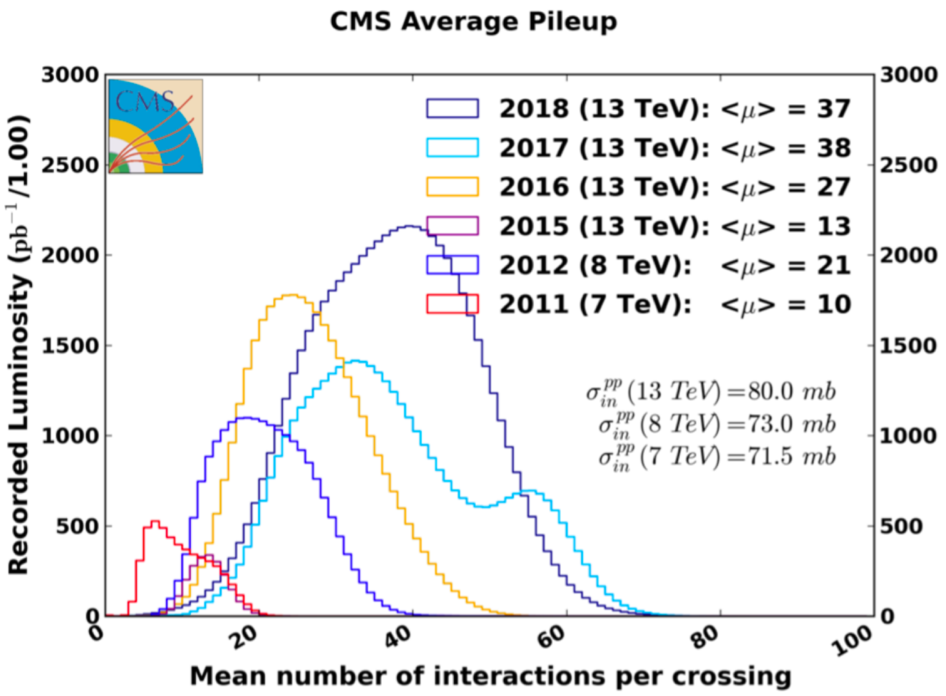
Yongbin Feng, Nhan Tran (Fermilab)

5th Inter-Experiment Machine Learning Workshop

9-13th of May 2022

CERN

pileup -Motivation for a semi-supervised approach



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

The current ML algorithms to model Pileup require a prior knowledge **of ground truth information**: whether the particle is produced from PU or Leading Vertex (LV).

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

Training and Performance per particle

- DELPHES (Z(νν)+jets and H(bb)+jets) samples that contain neutral labels so we can compare SL and SSL

- Construct a graph in η-ϕ space and connect the particles in a ΔR cone:

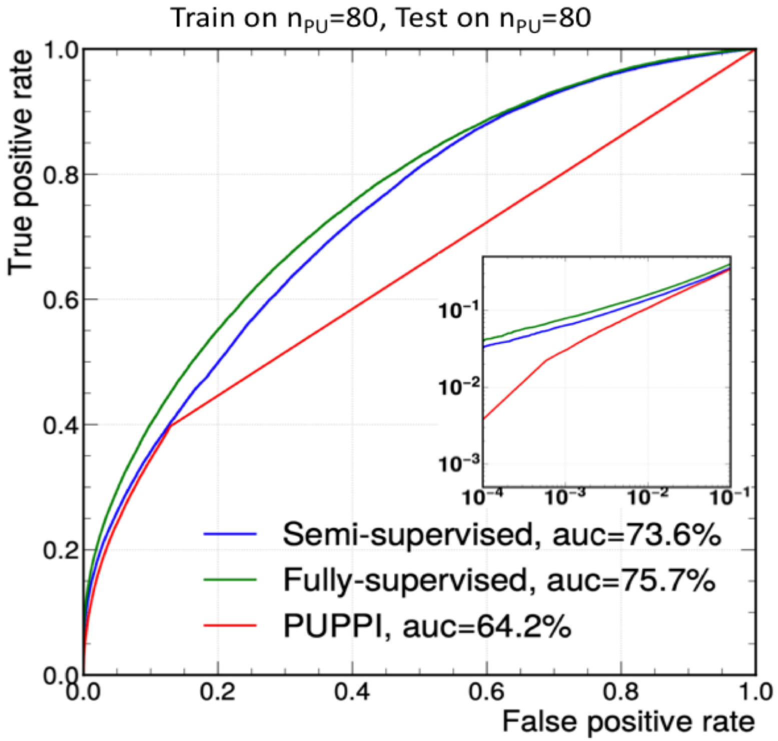
node features: P_T , charge, PUPPI weight, **edge features**: Δη, Δϕ, ΔR between the particles

- **gated model** was used for graph convolution

- **training:** for each event randomly select 7 LV charged particles and 126 PU charged particles to train

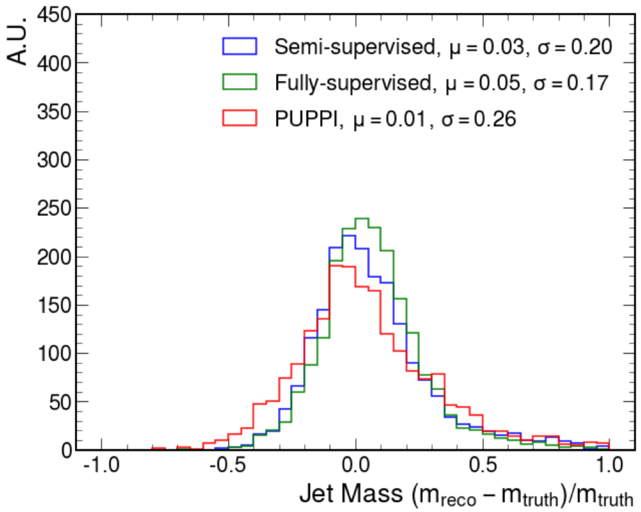
- **masking:** to avoid label leakage and potential bias, perform masking on the charged related features so that the GraphNN treats them as 'neutral'

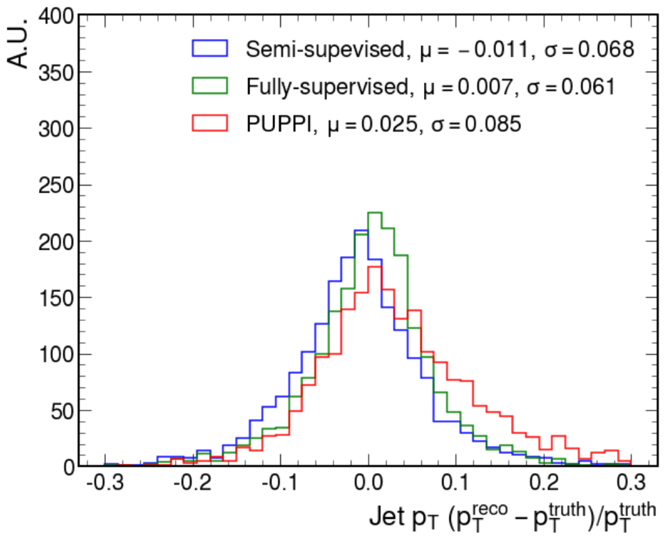
performance at a per-particle level of the LV and PU labels is shown in the ROC curve for PU=80 scenario

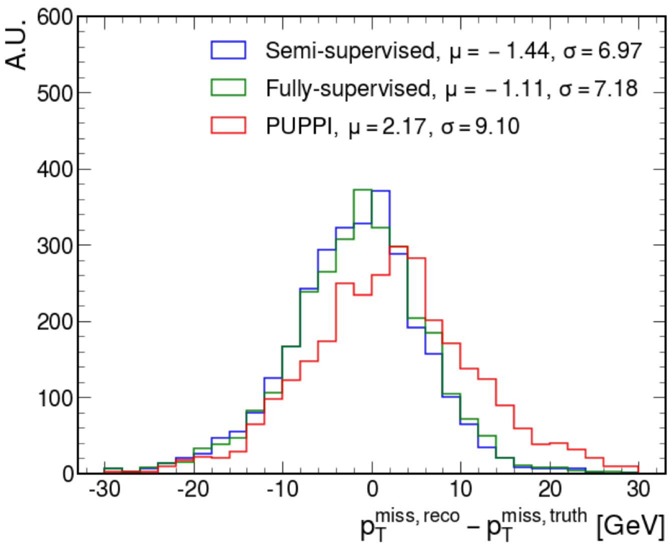


SL and SSL outperform PUPPI by around **10%**, a few percent lower performance from SL to SSL

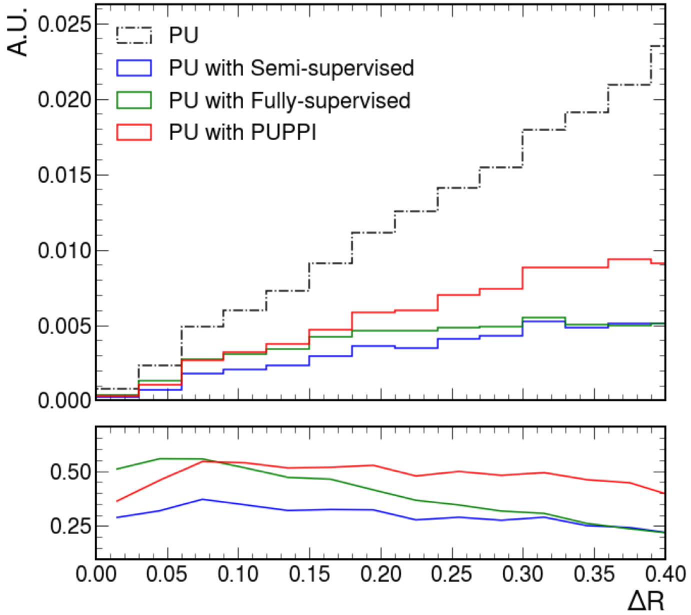
Performance on jet observables (jet mass, jet pt, and MET resolutions, at PU=80)







- **truth jets:** jets reconstructed using particles from the LV at the truth level
- **PUPPI jets and GNN predicted jets:** jets reconstructed using rescaled particle momentum with PUPPI weights/GNN outputs
- The resolutions look to be **~20% better** for jet observables comparing supervised GNN with PUPPI, slightly drop in performance for SL to SSL
- **GNN is more effective at removing PU particles**

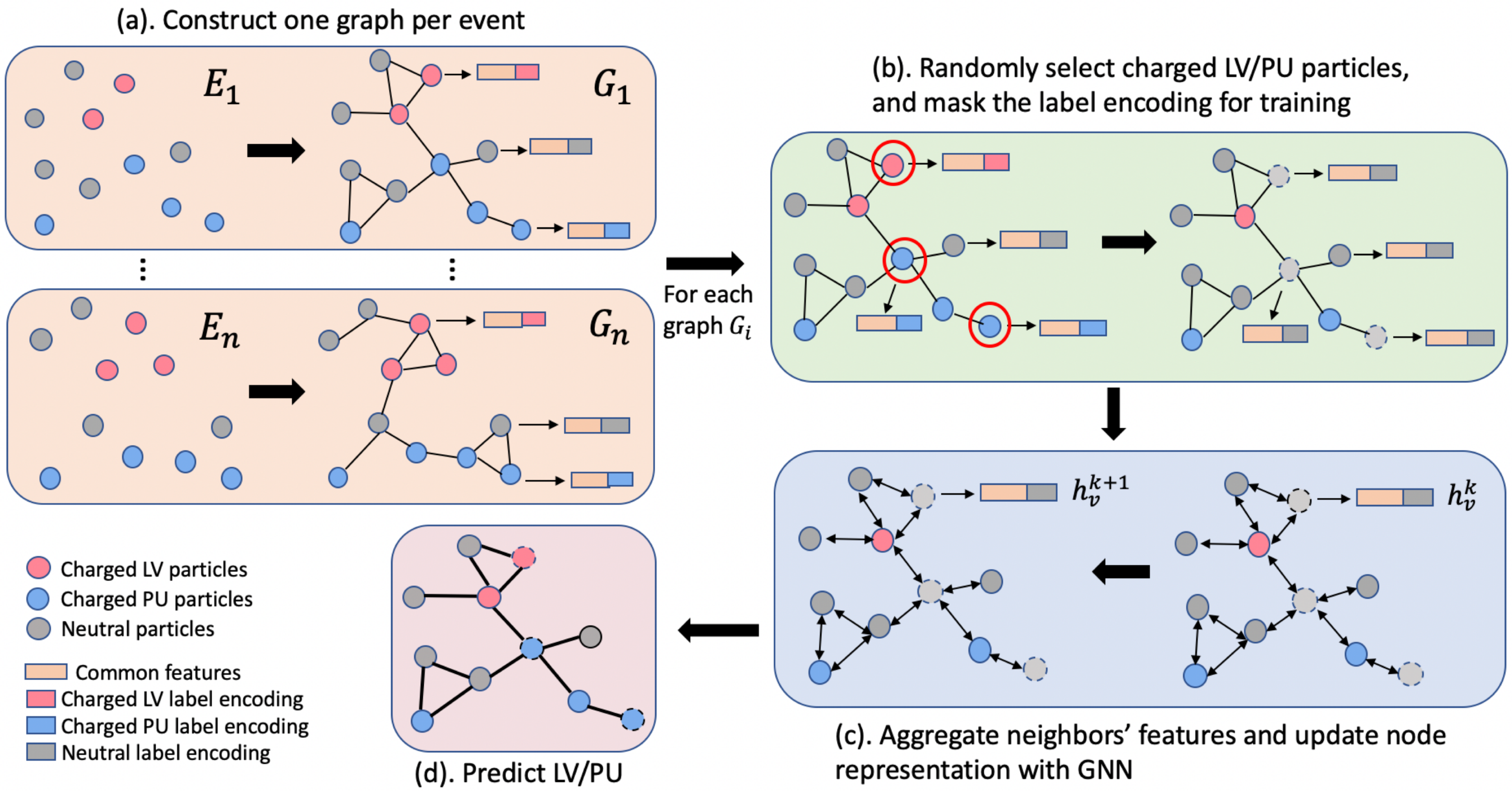


Conclusions-Future Steps

- Presented the study of applying **semi-supervised training** for pileup mitigation with **GraphNN** using fast simulation datasets
- training is performed on charged particles and the inference is on neutral particles
- results look very promising and are submitted to EPJC (arxiv)
- **Ongoing Efforts**: training and testing the performance of this technique on real CMS data/full simulation

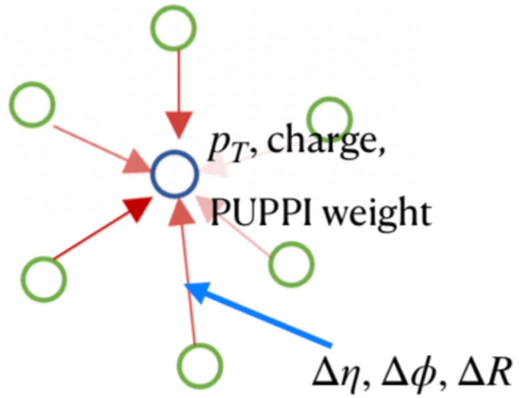
Back up

Illustration of the SSL training flow



Model Architecture

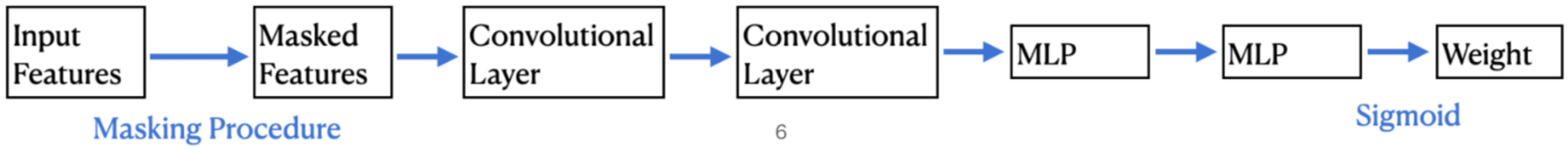
- Build graph in $\eta - \phi$ space. Connect the particles in the $\Delta R = 0.4/0.8$ cone.
- Input features:
  - Node features: p_T , charge, and PUPPI weights for the nodes
  - Edge features: $\Delta\eta$, $\Delta\phi$, and ΔR between particles
- Outputs are a weight between 0 and 1, representing the probability that the particle is produced from the LV
- Model architecture: **gated model**:
  - Build the model in a similar way to the PUPPI metric alpha
  - Node $h_u^{k+1} = G_u^k \cdot h_u^k + (1 - G_u^k) \cdot M_u^k$ where $G_u^k = \mathrm{Sigmoid}(h^k \oplus M_u^k)$ is the gate controlling the node feature updates
  - Message $M_u^k = \frac{1}{N} \sum_v (G_{u,v}^{'k} \cdot M_{u,v}^k)$, $G_{u,v}^{'k} = \mathrm{Sigmoid}(M_{u,v}^k)$ is the gate controlling feature passing from neighbor to the node



PUPPI metric

$$\alpha_i = \log \sum_{j \in \text{event}} \xi_{ij} \times \Theta(R_{\min} \le \Delta R_{ij} \le R_0),$$

$$\text{where } \xi_{ij} = \frac{p_{Tj}}{\Delta R_{ij}}.$$



Training Datasets

using datasets from PUPPIML

- Pythia 8.223 + Delphes 3.3.2 for simulation
- Z(vv)+jets signal processes
- pythia-generated QCD events as pileup; Poisson distribution sampled with the average pileup of 80
- charged particle flag for the LV and PU is set to be perfect