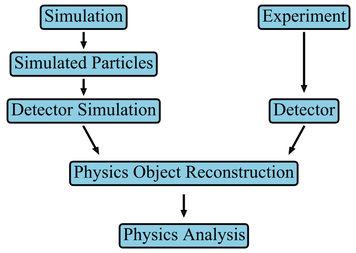
Graph Generative Adversarial Networks for Sparse Data Generation in High Energy Physics

Raghav Kansal[1], Javier Duarte[1], Breno Orzari[2], Thiago Tomei[2], Maurizio Pierini[3], Mary Touranakou[3, 4], Jean-Roch Vlimant[5], Dimitrios Gunopoulos[4]

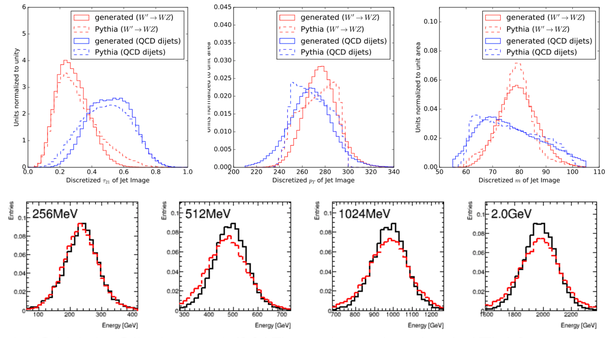
[1]UC San Diego, [2]Universidade Estadual Paulista, [3]CERN, [4]National and Kapodistrian University of Athens, [5]California Institute of Technology

Particle Collision Simulations

- Simulations of collision events are important for data analysis in high energy physics

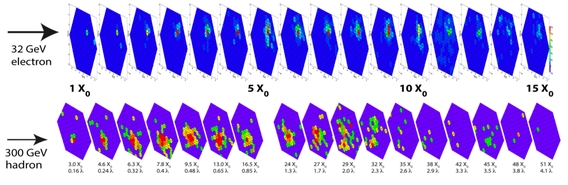


- Classical physics simulation programs such as GEANT4 are accurate but can be slow - O(min) per event
- Generative machine learning models are potential alternatives where we approximate simulations at much higher speeds - O(ms) per event
- Current work in this direction has produced impressive results:



LAGAN de Oliveira *et. al.* (2017) results on jets (top), ATLAS FastCaloGAN results ATLAS collaboration (2020) results (bottom)

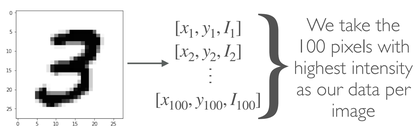
- But involves linear and convolutional architectures which aren't conducive to **the sparsity of the data** and the potentially **irregular underlying geometry**:



Datasets

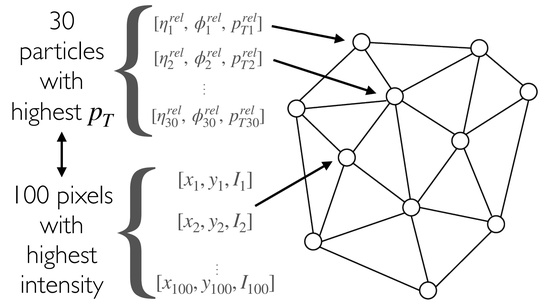
- Linear/CNNs represent jets and detector data as images
- An alternative approach is to represent this data graphically, which naturally suits the sparsity and adapts to any geometry:
- We test on two MNIST-derived graph datasets

Sparsified MNIST
(in analogy to jets or detector data)



MNIST Superpixels
(Monti et al. (2016))

- And a dataset of high transverse momentum (p_T) gluon jets where we take the 30 highest p_T particles and their $[\eta^{rel}, \phi^{rel}, p_T^{rel}]$ features
- All three are represented as graphs:

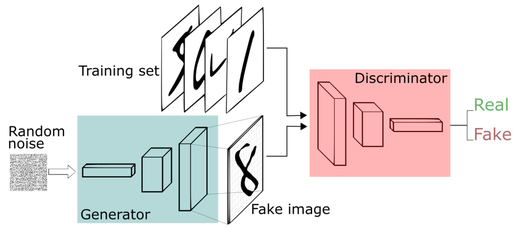


Evaluation

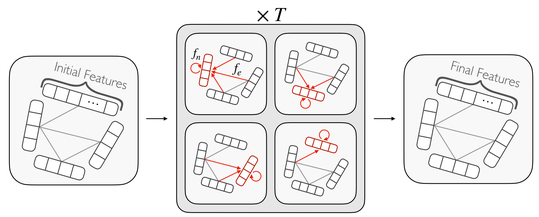
- We numerically evaluate the MNIST datasets using a graph based Fréchet Inception Distance metric - comparing the statistics of the activations of the pre-trained MoNet graph classifier
- Jets are evaluated by comparing the 1-Wasserstein (W_1) distance between the particle and jet level feature distributions - using bootstrapping for a baseline within pairs of 100 real samples and between real and fake

Graph GAN

- A generative adversarial network trains a discriminator and generator network iteratively and adversarially

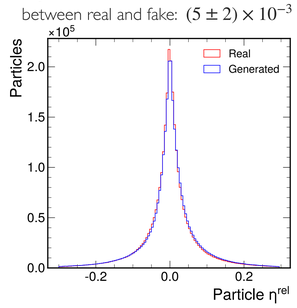


- We use a **message passing neural network** architecture for both networks:

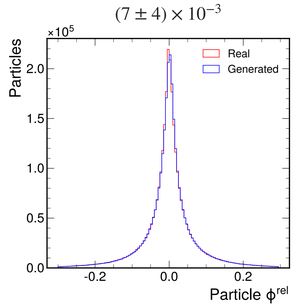


- We were successfully able to reproduce all three datasets with high fidelity, although there is some evidence of mode collapse with the superpixels

W_1 within real samples: $(6 \pm 2) \times 10^{-3}$
between real and fake: $(5 \pm 2) \times 10^{-3}$



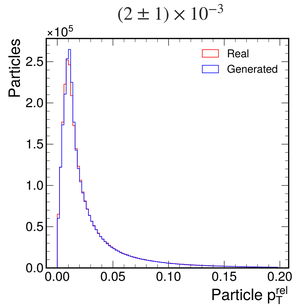
$(6 \pm 2) \times 10^{-3}$
$(7 \pm 4) \times 10^{-3}$



Results

- Sparse MNIST (top) and Superpixels (bottom) shown above, we achieved an average GFD of 0.52 and 0.30 respectively
- Jet feature distributions and the respective W_1 scores shown below

$(1.4 \pm 0.5) \times 10^{-3}$
$(2 \pm 1) \times 10^{-3}$



$(6 \pm 2) \times 10^{-3}$
$(6 \pm 2) \times 10^{-3}$