# lui-gnn

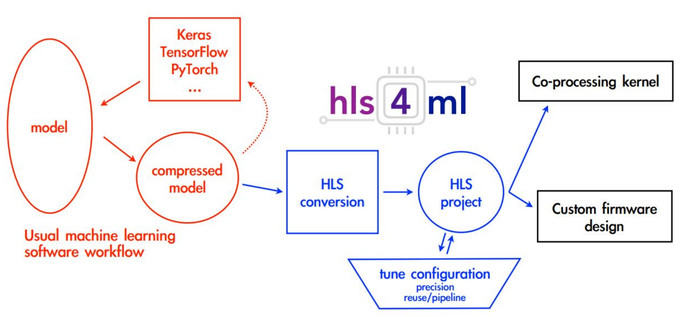
## A Graph Neural Network Surrogate Model for hls4ml

Dennis Plotnikov[1], Benjamin Hawks[2], Karla Tame-Narvaez[2], Nhan V. Tran[2]

[1] Johns Hopkins University [2] Fermi National Accelerator Laboratory

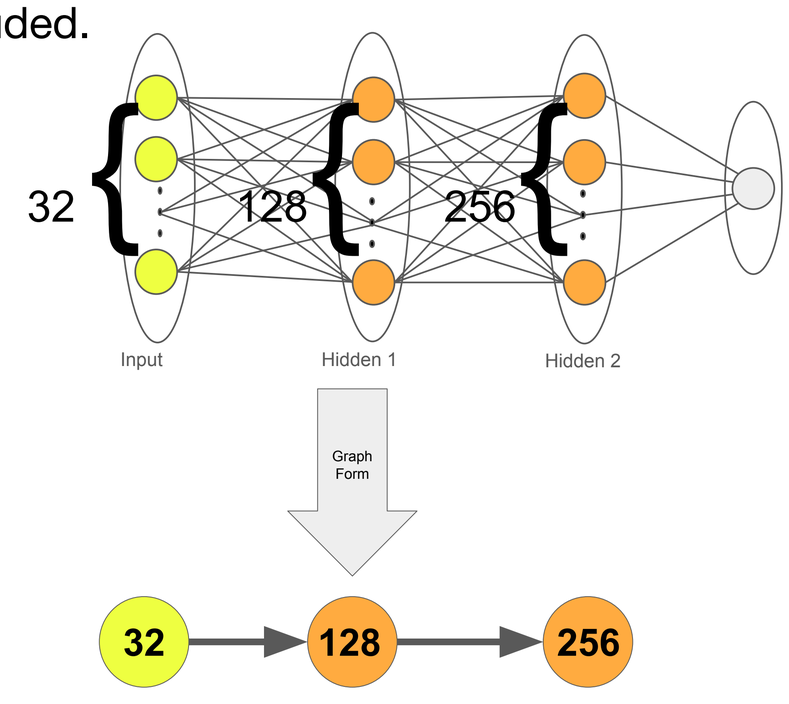
## Overview of hls4ml

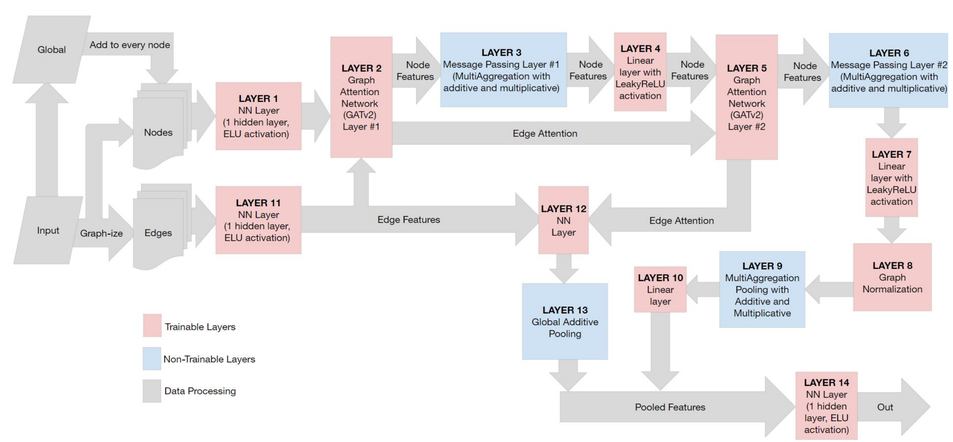
hls4ml is a pipeline used to convert machine learning models to a form that can be run on a field-programmable gate array (FPGA) or inscribed into an application-specific integrated circuit (ASIC). This has strong applications in high-energy physics, where detector triggers require latency on the scale of nanoseconds, but would benefit greatly from the power of machine learning.




## Modeling Neural Network Architectures as Graphs

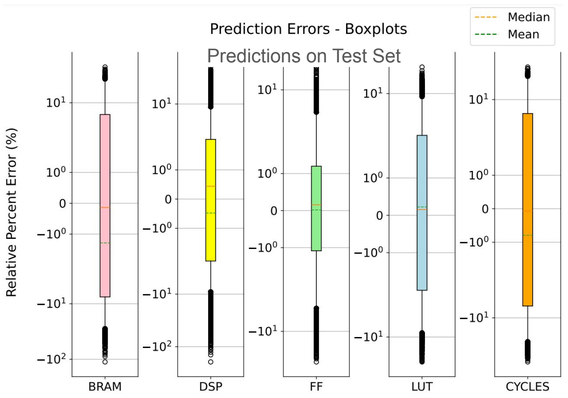
The structure of a typical neural network can be represented in the form of a directed graph. Each node represents a layer of the network, and each edge represents a feedforward connection. This allows for modeling complex architectures, including skip connections and recurrent networks. Node features (e.g. number of connections in the layer) and edge features (e.g. sparsity of the connection) can both be included.

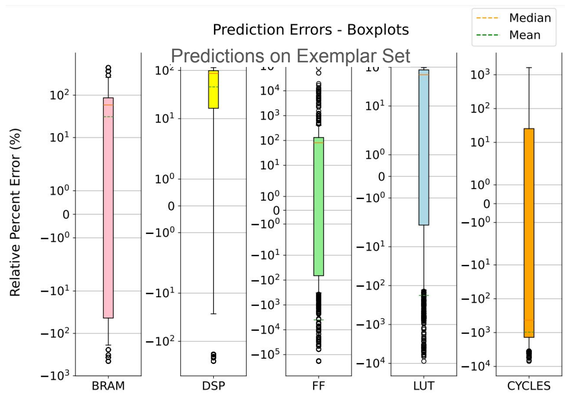



The graph-based input data for the surrogate model is best handled by a graph neural network (GNN). The network takes data in the form of graphs (represented with an adjacency list), and runs them through two GATv2 graph attention networks (arXiv:2105.14491) and two message-passing graph convolution layers, before pooling all edge and node results with global features. A final neural network layer allows all three types of data to inform a prediction.

## Results and Conclusion




The GNN was trained on the wa-hls4ml benchmark test set of randomized neural networks. On testing data similar to its training data, lui-gnn performs quite well. For a majority of the test points, the predicted values for latency and resource consumption has no more than a 10% error. On the wa-hls4ml benchmark exemplar dataset (composed of various scientific algorithms for comparison, with strongly variable model structure and synthesis parameters), the performance suffers, likely due to insufficiently heterogeneous input data. As more data becomes available for training the model, the predictions are expected to generalize more effectively.

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).This work was supported in part by National Science Foundation (NSF) awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, OAC-2112167, CNS-2100237, CNS-2120019, the University of California Office of the President, and the University of California San Diego's California Institute for Telecommunications and Information Technology/Qualcomm Institute. Thanks to CENIC for the 100Gbps networks.