Optimizing Sparse Neural Architectures for Low-Latency Anomaly Detection

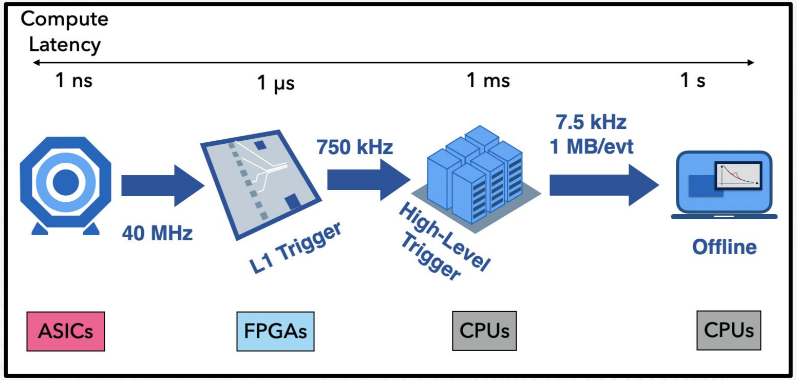
Luke McDermott[1,2], Jason Weitz[1], Javier Duarte[1], Nhan Tran[3]

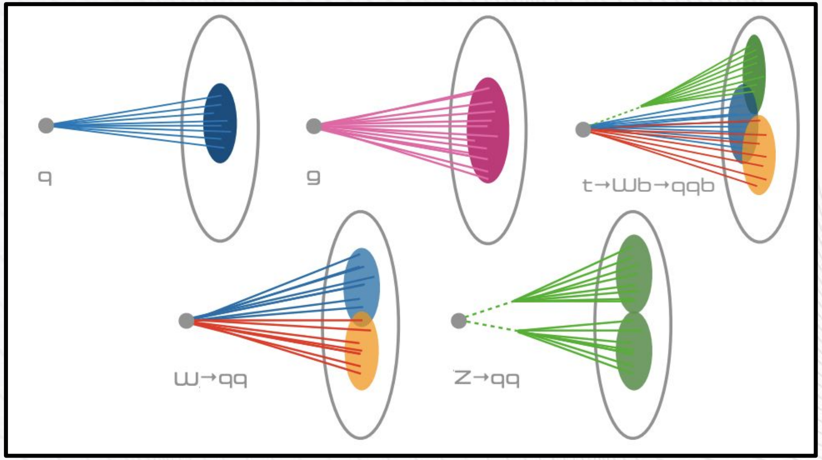
1. UC San Diego, 2. Modern Intelligence, 3. Fermilab

Anomaly Detection at the LHC at 40 MHz

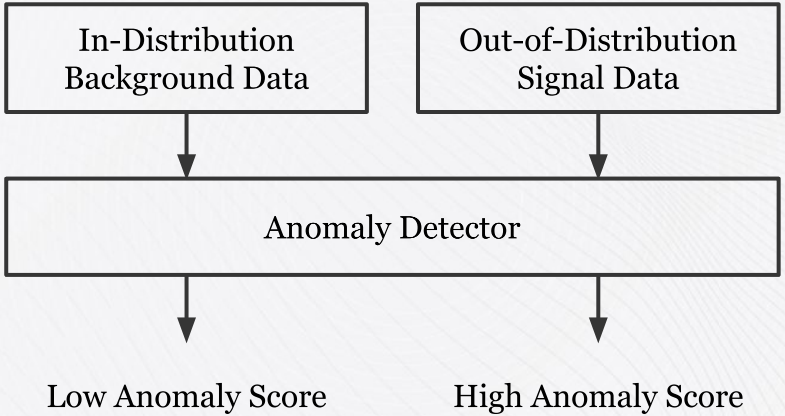
- Anomaly detection to search for new physics is an essential task
- To run AD in the first level of data selection requires algorithms with sub-microsecond latencies running on FPGAs
- Motivates studying how to optimally compress and discover sparse neural architectures
- Study using JetNet dataset [1] with q/g jets as background and t/W/Z jets as anomalies

[1] https://zenodo.org/record/6975118

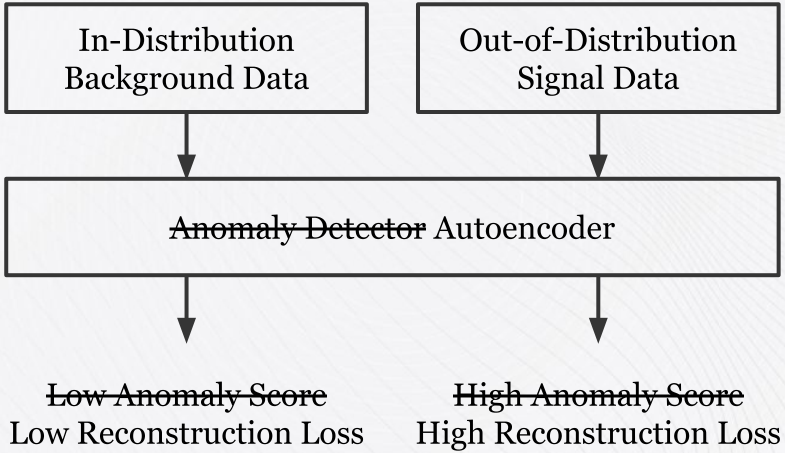




Anomaly Detectors



Anomaly Detectors

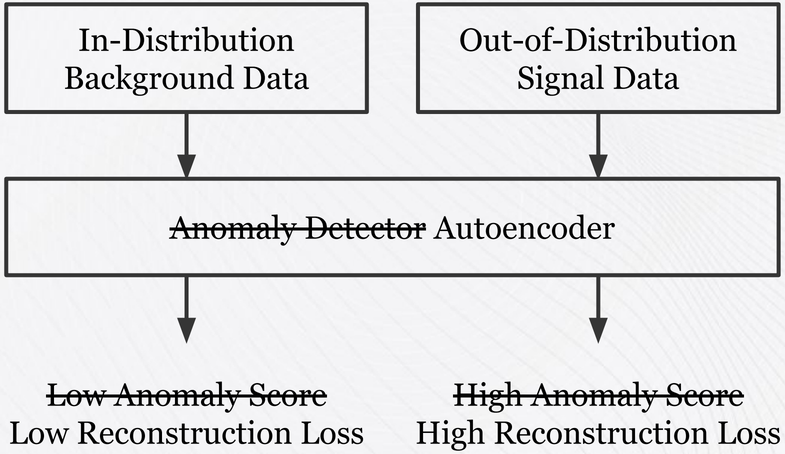


Anomaly Detectors

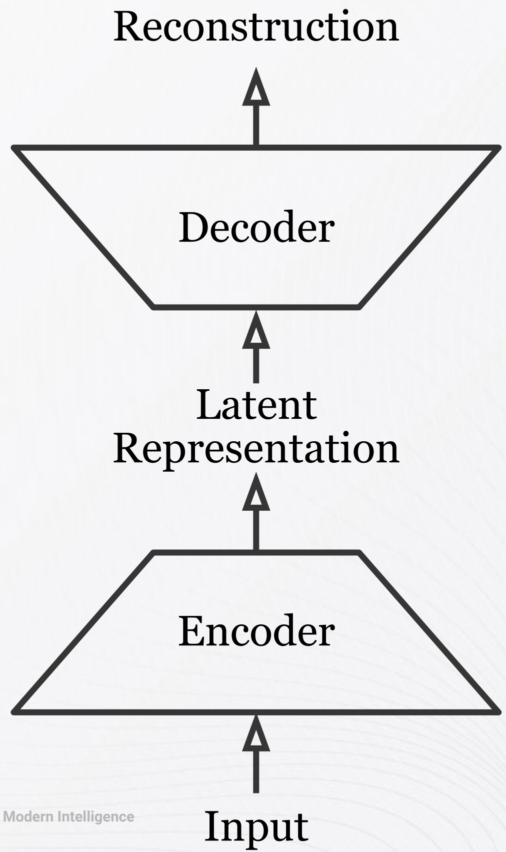
JetNet Dataset:

Background Data (In-Distribution)
Gluon & Light Quark

Signal Data (Out-of-Distribution)
Top Quark, W-Boson, & Z-Boson



Autoencoders with Reconstruction Loss in General

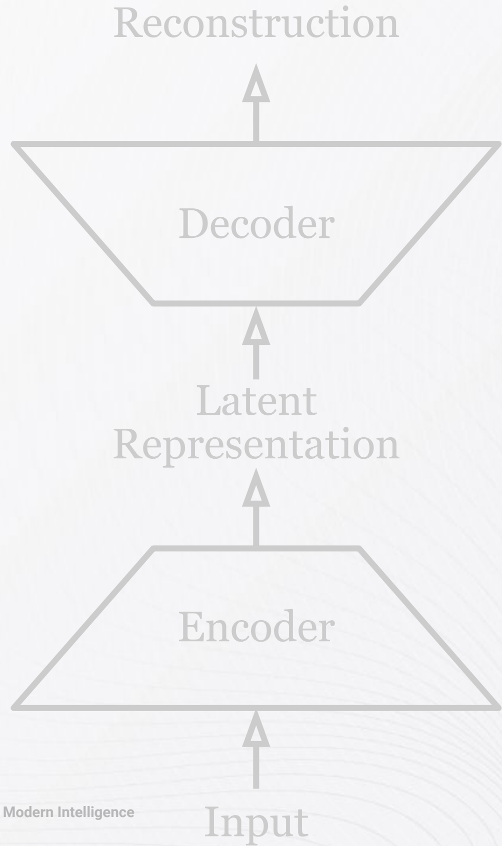


Goal #1:
Compress latent representation by modeling input distribution & removing noise

Goal #2:
Generalize to out-of-distribution data

Reconstruction Loss = MSE(Input, Reconstruction)

Reconstruction-Based Anomaly Detection

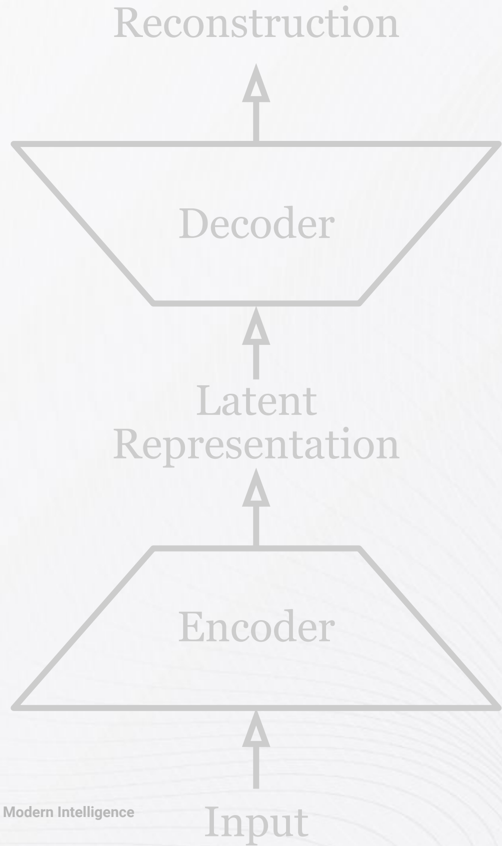


Goal #1:
Compress latent representation by modeling input distribution & removing noise

Goal #2:
Generalize to out-of-distribution data

Low Reconstruction Loss = Background
High Reconstruction Loss = Anomaly

Reconstruction-Based Anomaly Detection

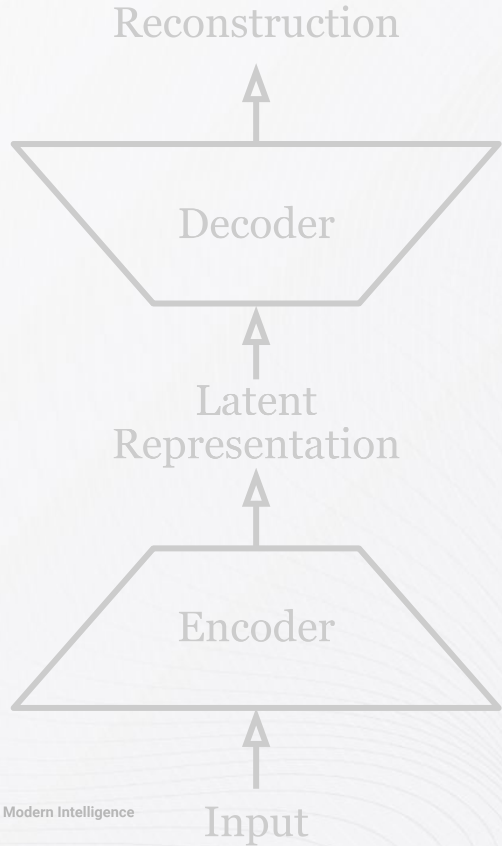


Goal #1:
Compress latent representation by modeling input distribution & removing noise

~~**Goal #2:**~~
~~Generalize to out-of-distribution data~~

Low Reconstruction Loss = Background
High Reconstruction Loss = Anomaly

Reconstruction-Based Anomaly Detection



Goal #1:
Compress latent representation by modeling input distribution & removing noise

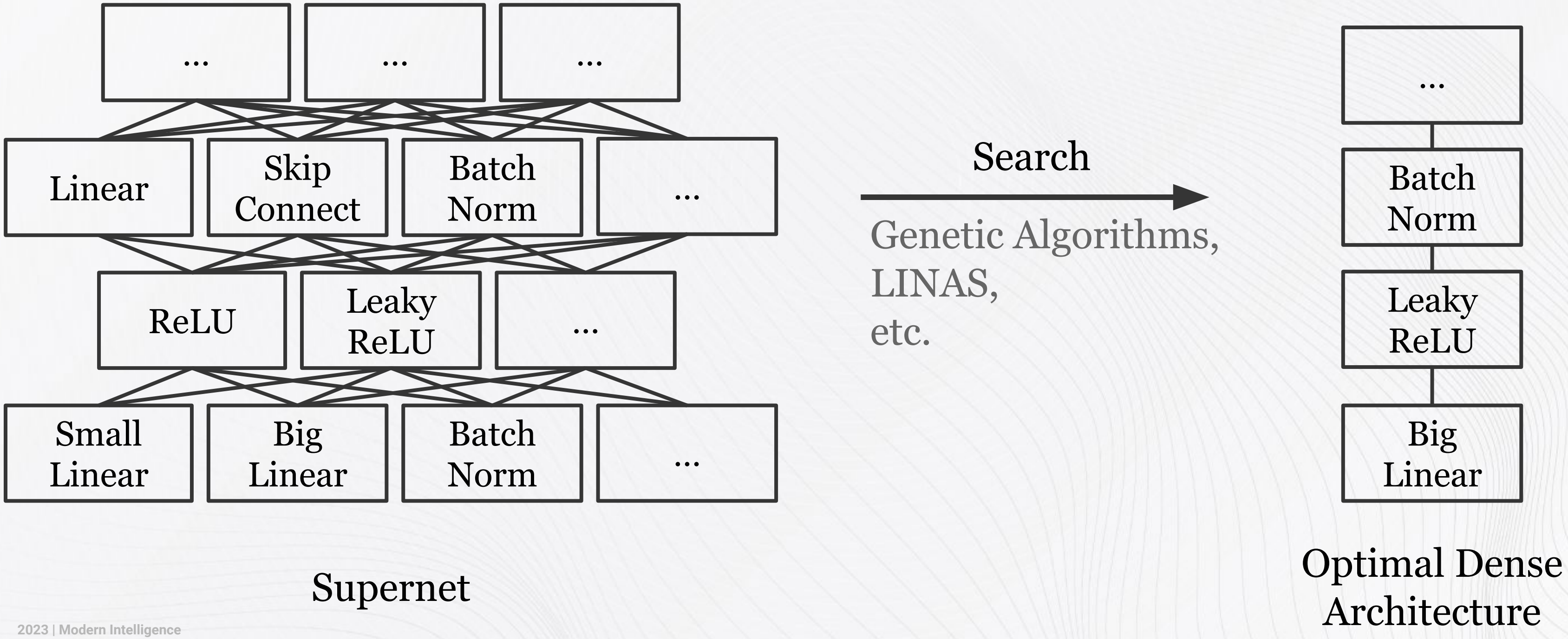
~~**Goal #2:**~~
~~Generalize to out-of-distribution data~~

Low Reconstruction Loss = Background
High Reconstruction Loss = Anomaly

Revised Goal #2:
Only reconstruct in-distribution data

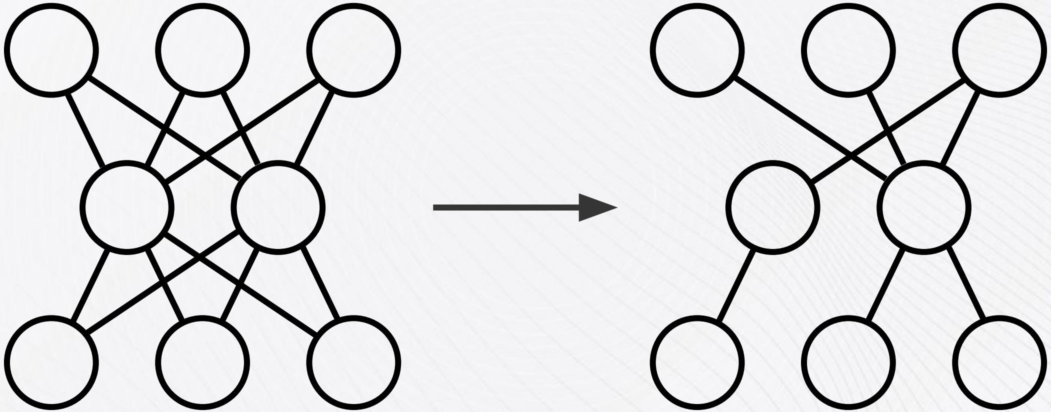
Global Architecture Search with Supernetworks

Search across dense architectures & train once (One-Shot NAS)



Local Architecture Search through Pruning

Remove unnecessary parameters for faster inference

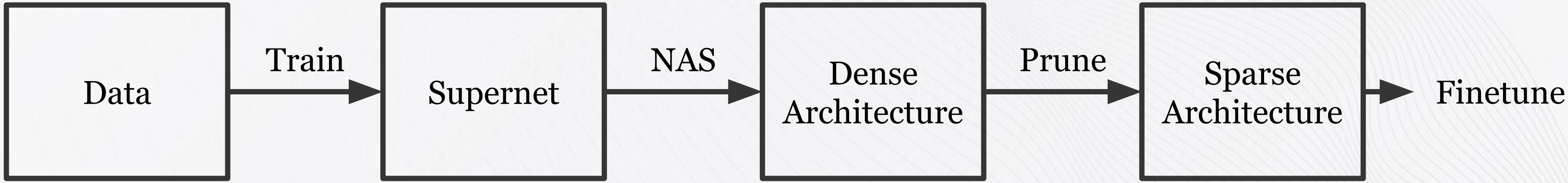


Dense Architecture

Sparse Architecture

We employ Iterative Magnitude Pruning w/ Weight Rewinding

Our Framework



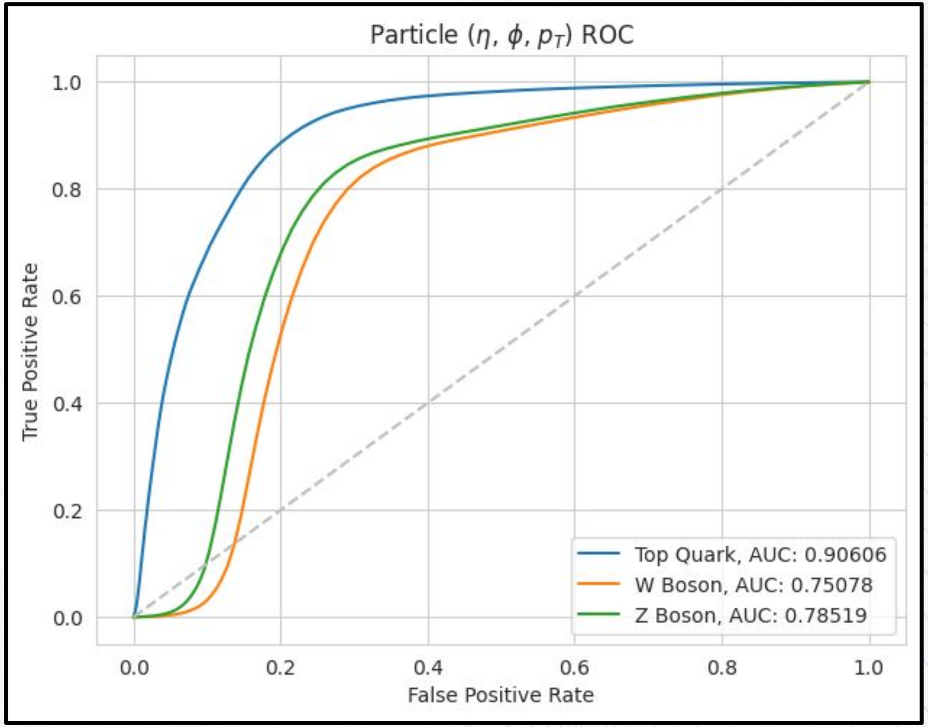
Optimal Architecture

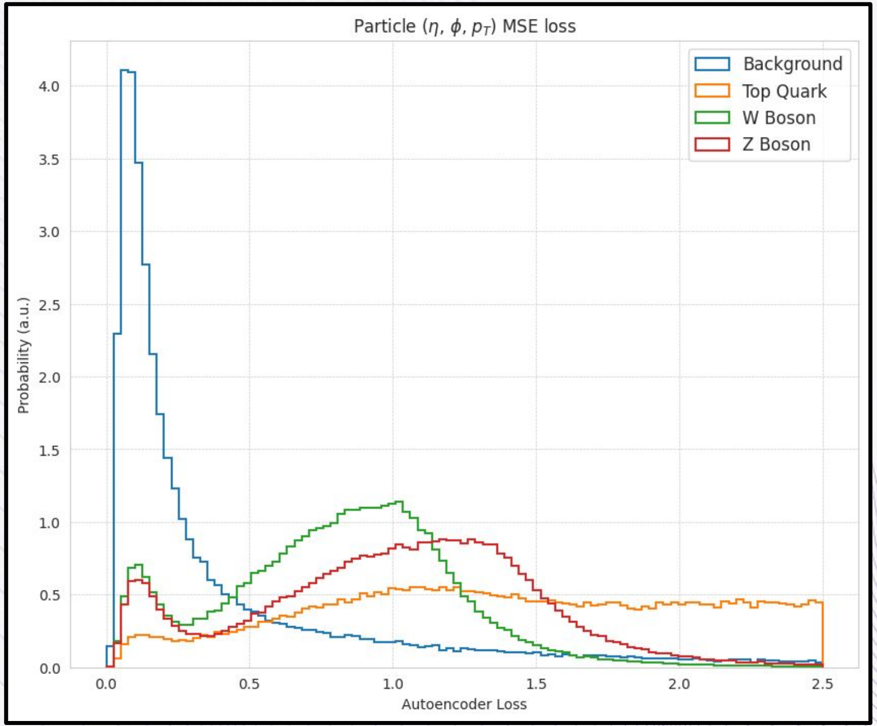
Layer	Sparsity | # params
Batch Norm	
Linear 30x16 + L-ReLU	76% | 114
Linear 16x8 + L-ReLU	56% | 56
Linear 8x4	22% | 25
Latent Dim = 4	
Linear 4x8	31% | 22
Batch Norm + L-ReLU	
Linear 8x16 + L-ReLU	45% | 70
Linear 16x30	73% | 132

Overall Sparsity = 67%

Number of Parameters = 419

Preliminary Results



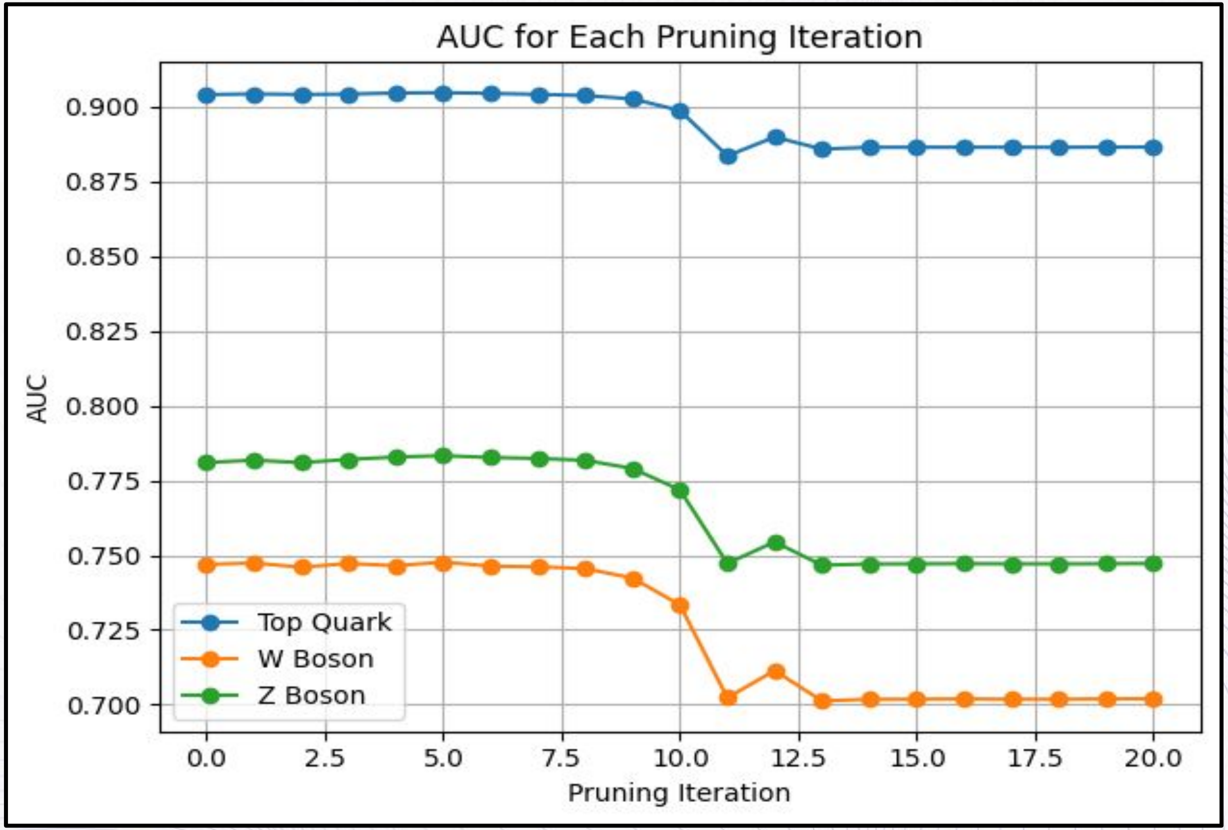


Comparison to Complex Models on JetNet Dataset

Model	Top Quark AUC	W Boson AUC	Z Boson AUC
Sparse AE (ours)	**0.9061**	<u>0.7508</u>	<u>0.7852</u>
LG AE-Min-Max	0.8539	0.6938	0.7400
LG AE-Mix	0.8669	0.7489	**0.7909**
GNN AE-JL	0.8530	0.5937	0.6545
GNN AE-PL	0.8917	**0.7558**	0.7805
CNN AE	<u>0.8962</u>	0.6886	0.7700

Key: **Best Model**, <u>Second Best Model</u>

Hao, Z., Kansala, R., Duarte, J., & Chernyavskaya, N. 2023. Lorentz group equivariant autoencoders. arXiv preprint arXiv:2212.07347.


AUC for Each Pruning Iteration
0.900
0.875
0.850
0.825
0.800
0.775
0.750
0.725
0.700
AUC
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Pruning Iteration
Top Quark
W Boson
Z Boson

Takeaways & Future Work

- Benchmark across more difficult anomaly detection datasets
    - Simple statistical baselines perform well on past datasets
- Implement Quantization Aware Training in the Inner Loop of Neural Architecture Search
- Optimize for Mixed-Precision Quantization
- Implement Hardware-Aware NAS Frameworks for FGPA optimization
- Promote the use of AutoML in the FastML community