A Graph Neural Network-based Top Quark Reconstruction Package

Allison Xu
Haichen Wang, Xiangyang Ju

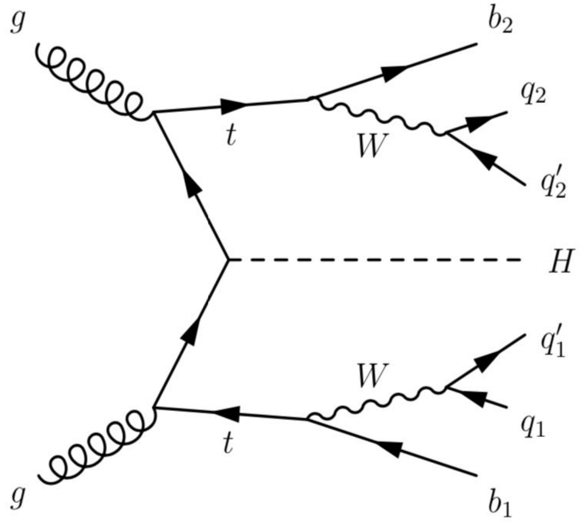
UC Berkeley
Lawrence Berkeley National Laboratory

August 10, 2020

Berkeley
UNIVERSITY OF CALIFORNIA

Overview

- Graph representation
- Graph neural network (GNN)
- GNN for ttH top reconstruction
  - Higgs diphoton decay channel
  - Top quark all-hadronic final states
- GNN performance
  - Comparison to Boosted Decision Tree (BDT)



Source: https://cds.cern.ch/record/2719502

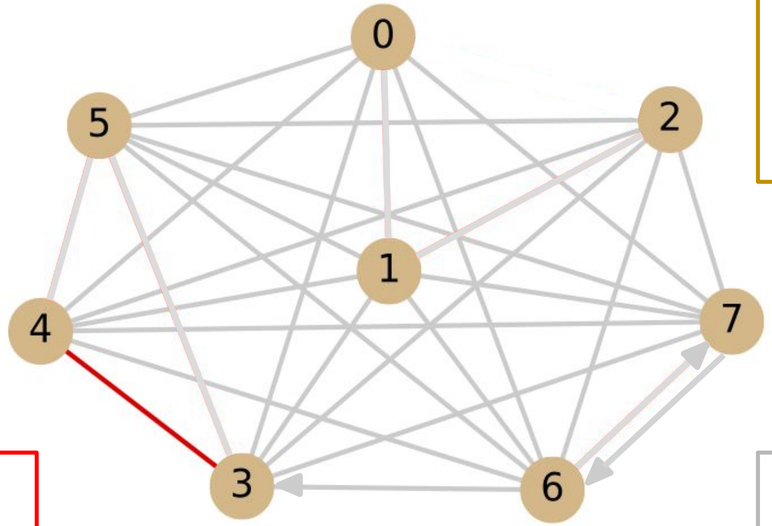
Berkeley
UNIVERSITY OF CALIFORNIA

Motivation

- Flexibility in graph construction
- Outperforms other models
- By more accurately reconstructing tops, we can construct accurate top variables, which can be useful in many models

Graph Representation

Global
Global feature 1
Global feature 2
…



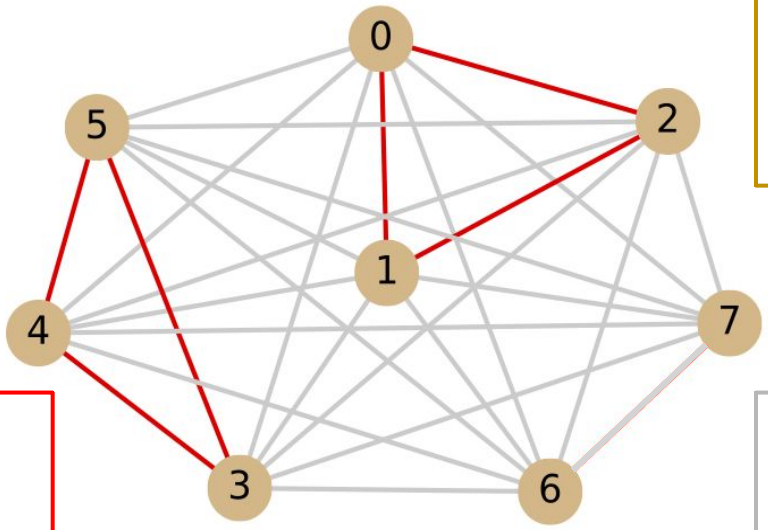
Node
Node feature 1
Node feature 2
…

Edge
Edge feature 1
Edge feature 2
…
Target 1

Edge
Edge feature 1
Edge feature 2
…
Target 0

Graph for ttH

Global
Number of jets



Node (Jet)
Four momentum
B-tagging score

Edge
Distance
Difference in angle
Mass of two-jet system

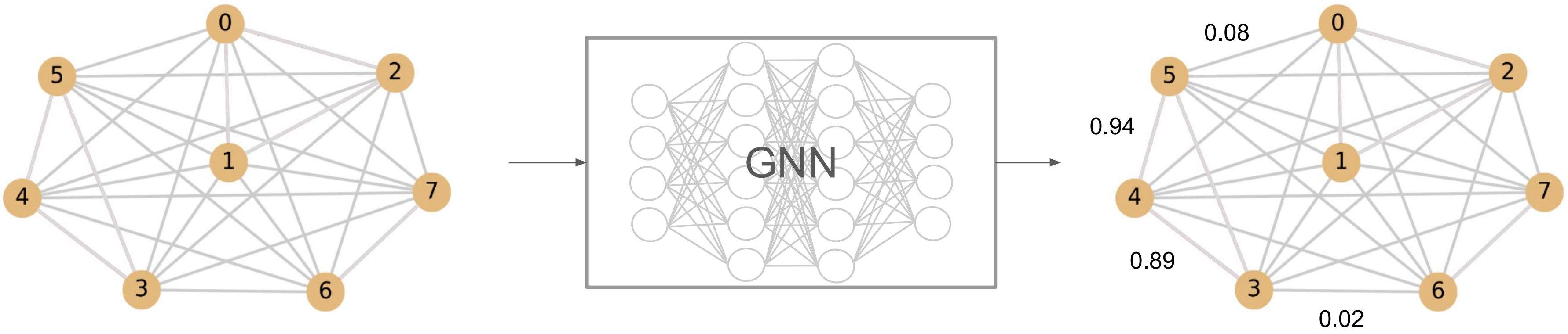
Target 1 if two jets come from the same top quark decay

Edge
Distance
Difference in angle
Mass of two-jet system

Target 0 otherwise

Graph Neural Network

- Input: Graphs without target
- Output: Graphs with scores for every edge
  - How likely the jets connected by an edge come from the same top quark decay
- Implemented using TensorFlow

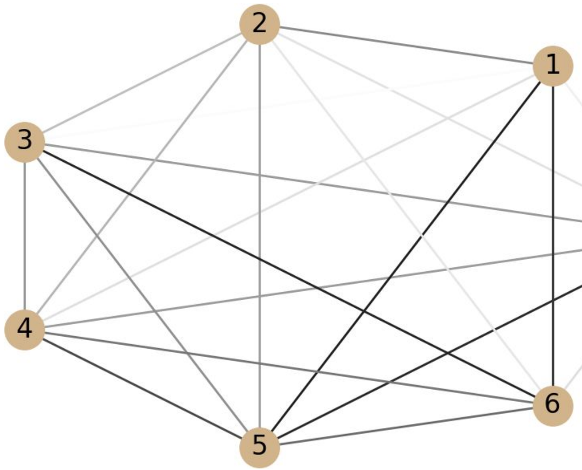


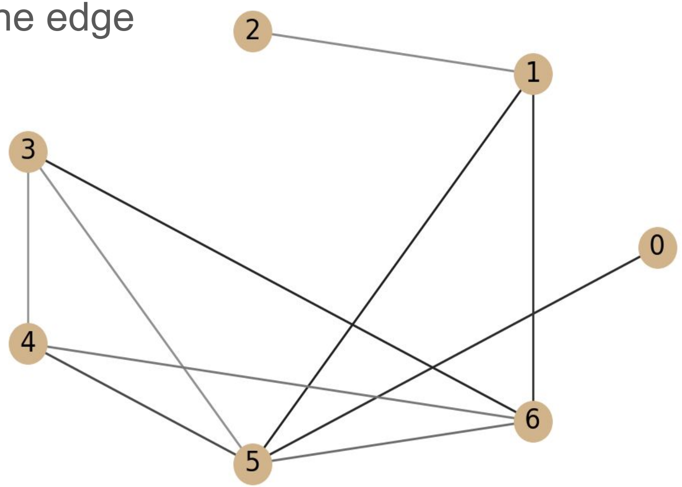
GNN Top Reconstruction

1. Construct the graphs
2. Feed graphs into the training
3. Apply the model on all events
4. Reconstruct triplets from jets using edge scores
5. Evaluate performance

GNN Top Reconstruction Algorithm

1. Remove edges with scores below the edge threshold





GNN Top Reconstruction Algorithm

2. Construct all possible triplets from the remaining edges
3. Score each triplet (e.g. sum of three edge scores)

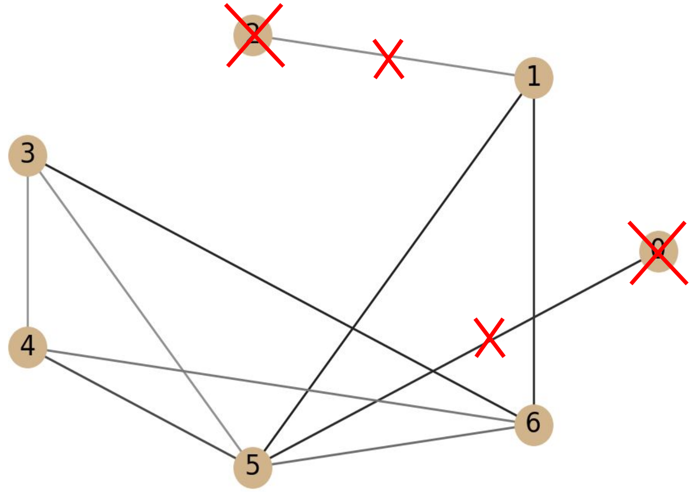
(1, 5, 6)

(3, 4, 5)

(3, 4, 6)

(3, 5, 6)

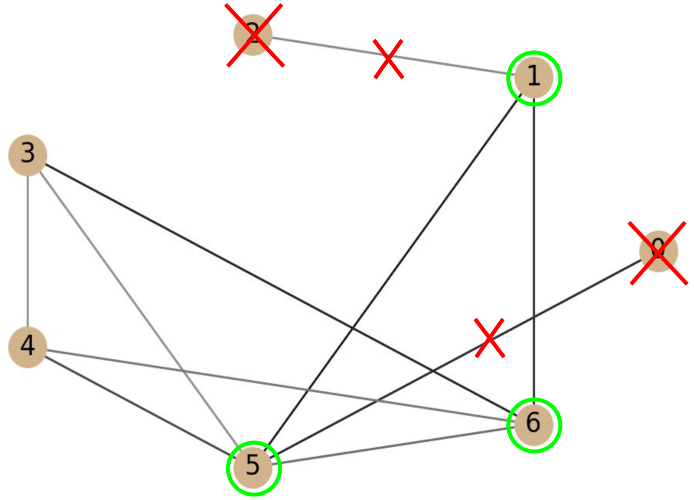
(4, 5, 6)



Berkeley UNIVERSITY OF CALIFORNIA

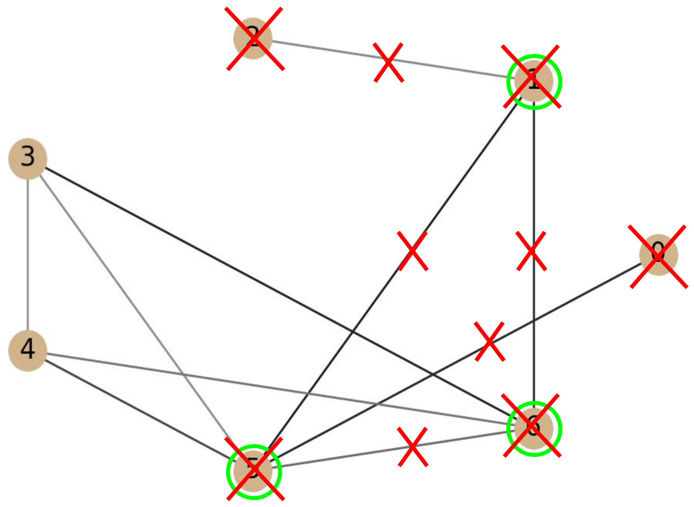
GNN Top Reconstruction Algorithm

4. Select the highest scoring triplet, if possible



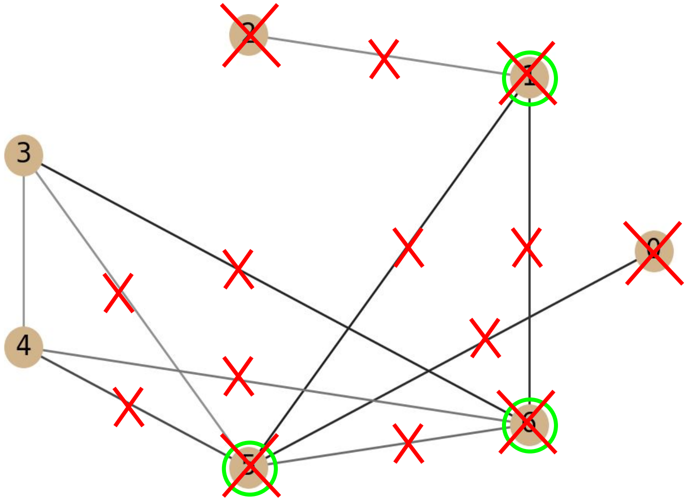
GNN Top Reconstruction Algorithm

5. Eliminate triplets containing any of the jets in the highest scoring triplet



GNN Top Reconstruction Algorithm

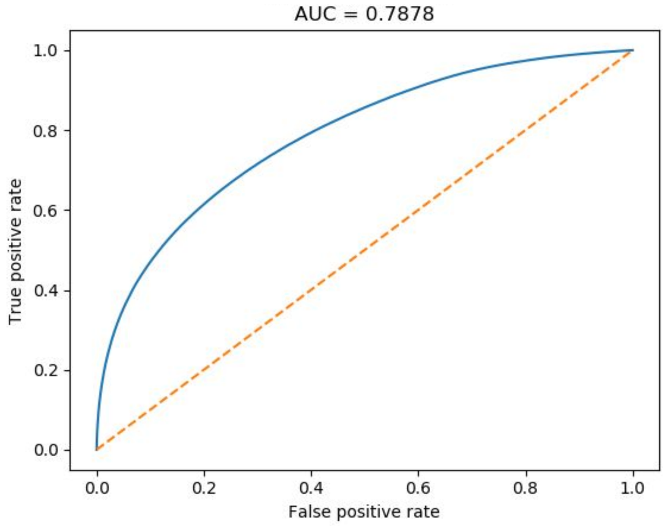
5. Eliminate triplets containing any of the jets in the highest scoring triplet
6. Select the next highest scoring triplet, if possible



GNN Performance

- Efficiency = Number of correctly identified triplets / Number of top quarks
- ROC curve and AUC (area under ROC)

True positive rate = Number of true triplets that pass a cut on triplet score / Number of true triplets
False positive rate = Number of false triplets that pass a cut on triplet score / Number of false triplets

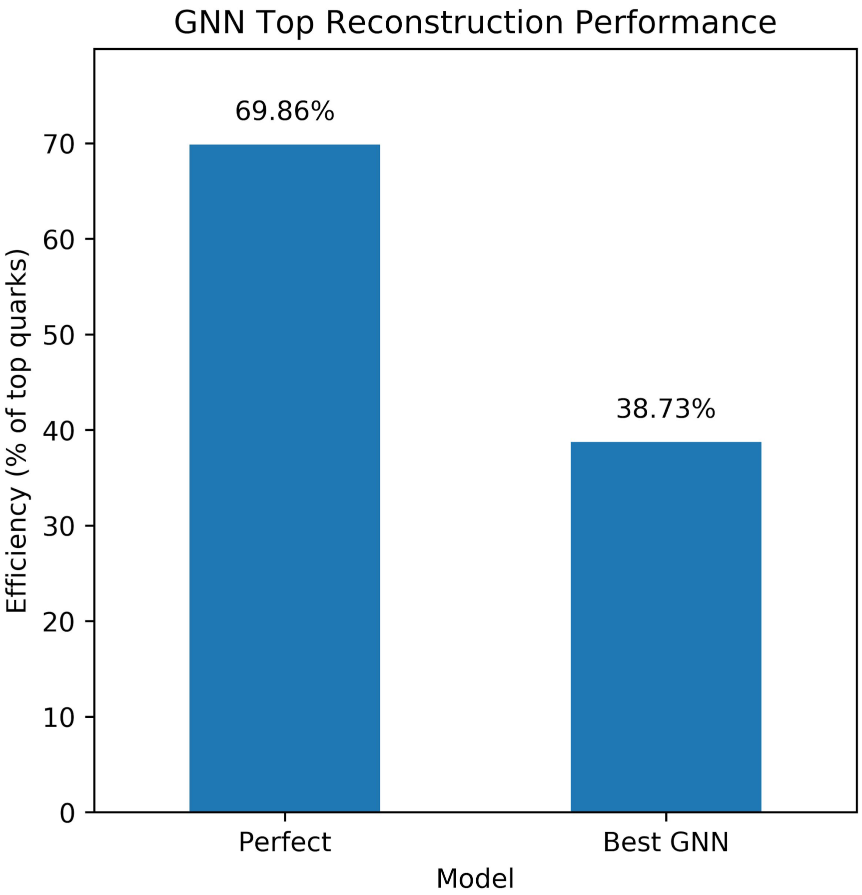


Berkeley UNIVERSITY OF CALIFORNIA

GNN Performance

Preselection:

- 2 truth-matched triplets
- 0 leptons



BDT Top Reconstruction

- Boosted decision tree is the current model used for top reconstruction
- Fair comparison of GNN with BDT using efficiency and AUC
- Applied in both hadronic and semi-leptonic channels
- Resource: https://cds.cern.ch/record/2719502

BDT Comparison

Preselection: 2 truth-matched triplets, 0 leptons

Model	Efficiency	AUC
Perfect	69.9%	100%
A	33.0%	75.2%
B	40.5%	80.5%
Best GNN	38.7%	78.8%
BDT	36.7%	72.0%

Models

- Perfect: Achieves the maximum possible efficiency and AUC
- A: Initial GNN model
- B: Trained on events with two truth-matched triplets

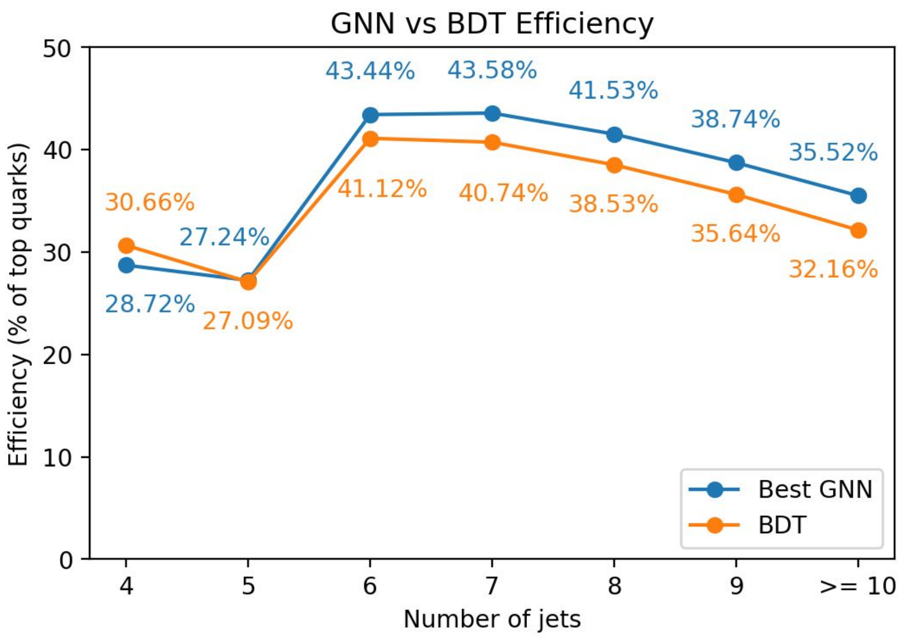
  To remove ambiguity in edges with target 0 for events with zero or one truth-matched triplet
- Best GNN: Tuned composition of training events

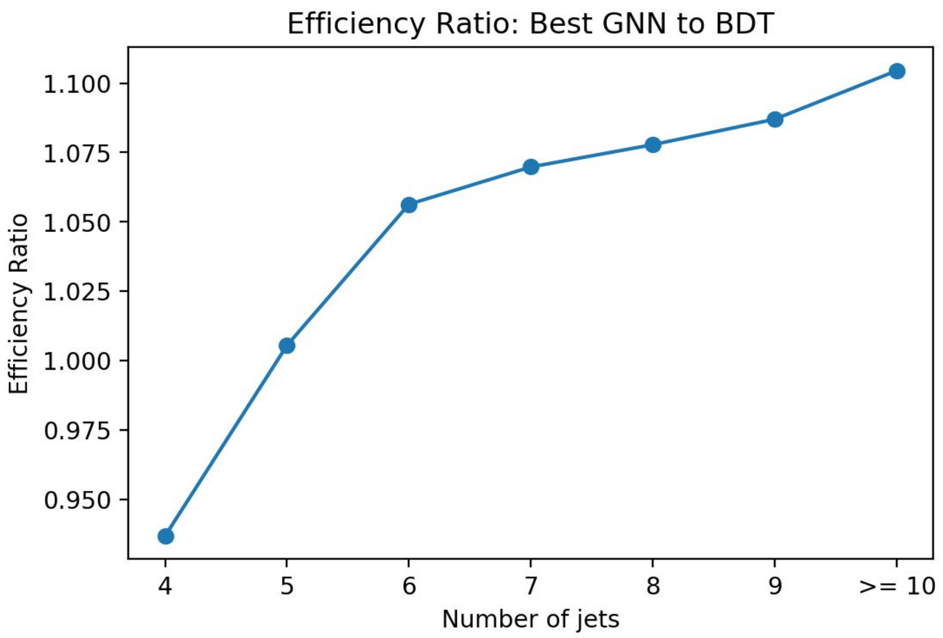
  To remove dependence between edge scores and number of jets in an event

Berkeley
UNIVERSITY OF CALIFORNIA

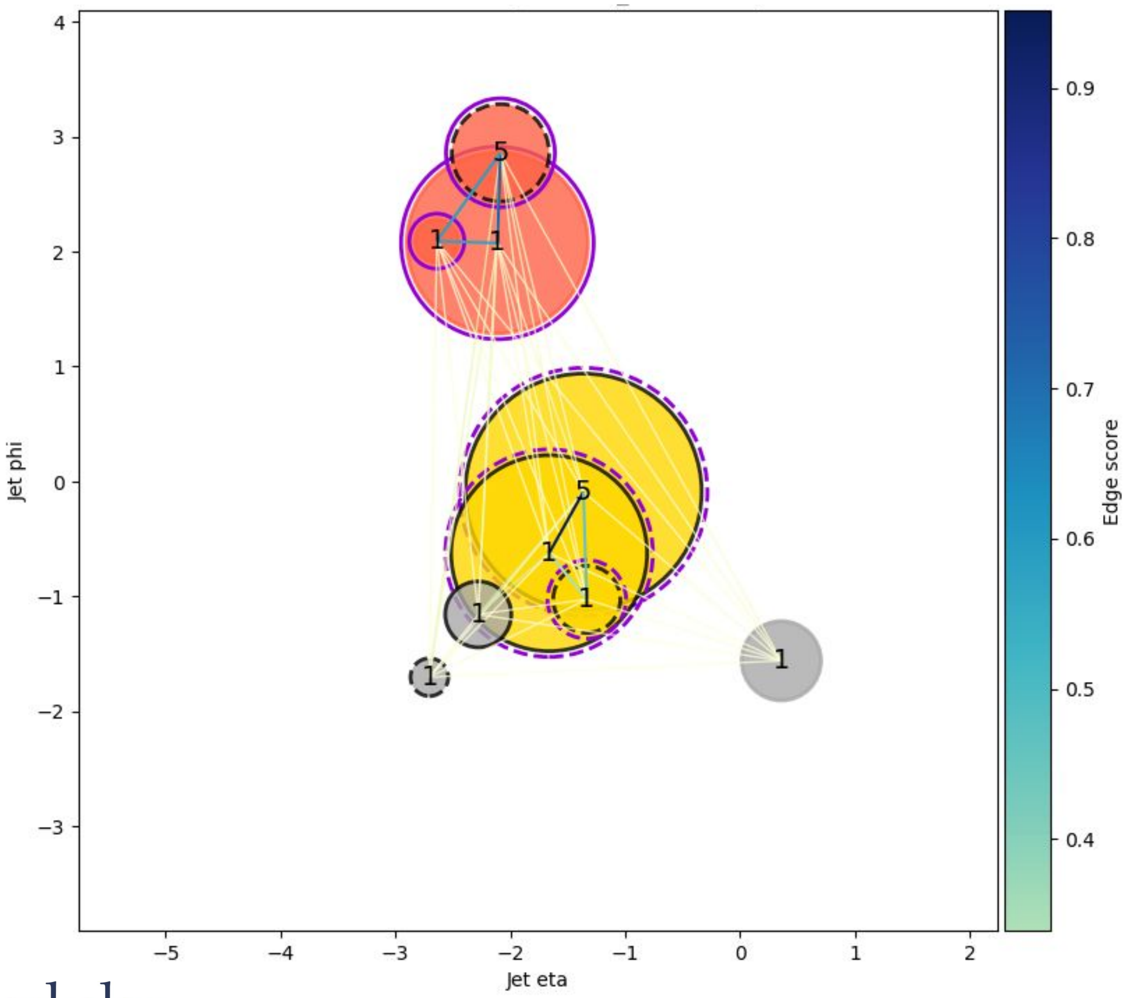
BDT Comparison

Preselection: 2 truth-matched triplets, 0 leptons





Berkeley
UNIVERSITY OF CALIFORNIA



GNN correctly identifies both triplets
BDT correctly identifies none

- Truth-matched triplets - red and yellow nodes
- Other jets - gray nodes
- GNN triplets - purple outline
- BDT triplets - black outline

- Area of node proportional to transverse momentum of jet
- B-tagging score - node label

- Edges colored by edge score

Berkeley
UNIVERSITY OF CALIFORNIA

Conclusion

- GNN outperforms BDT, specifically in events with 6 or more jets

- Experiment with different graph configurations
- Evaluate model for semi-leptonic final states
- Use top variables as training variables for other models
- Reconstruction of other complex processes

Acknowledgements

Prof. Haichen Wang (PI)

Dr. Xiangyang Ju

Dr. Shuo Han

Ryan Roberts

US ATLAS SUPER Program

Lawrence Berkeley National Laboratory

University of California, Berkeley