# Heavy-Flavour Jet Tagging at LHCb Using Graph Neural Networks

**Nate Grieser**, Gabriella Pesticci

University of Cincinnati, CERN

**2024 ML4Jets Workshop**

05-11-2024

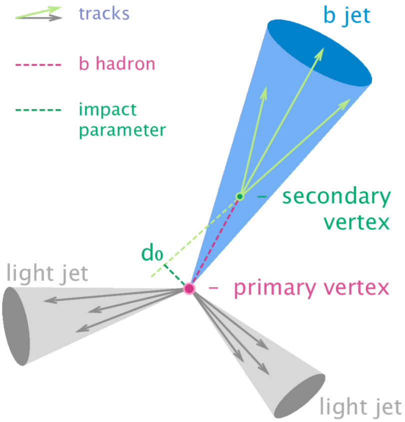
# Motivation For HF Tagging

**Jet Hadronization**

- *pp* collisions at the LHC produce quarks and gluons which hadronize due to QCD confinement
- Longer lifetime of HF-jets creates secondary vertex

**Goal:**

- Classify jets originating from different HF quarks ($b$, $c$)



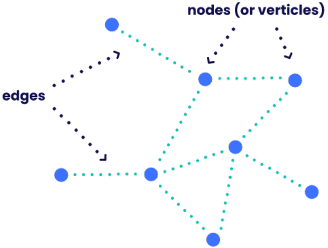
# Leveraging Graph Neural Networks

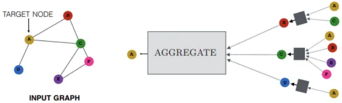
## GNN Advantages

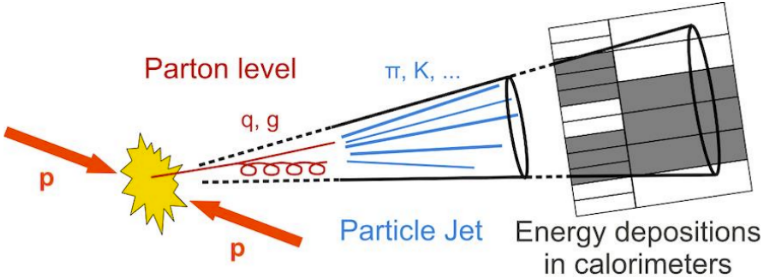
- Variable number of nodes and edges
- Capture complex relationships to represent system

## Message Passing:

- Node information aggregated from neighboring nodes
    - Learn features of neighbors

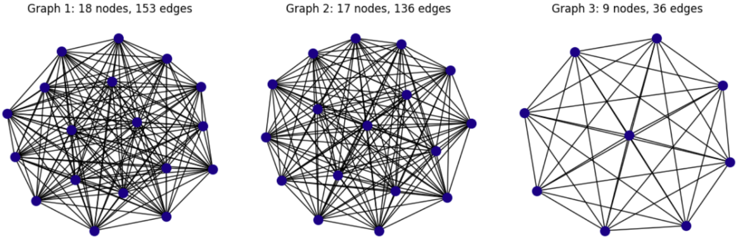





- 1 graph = 1 jet
- 1 node = 1 daughter
- Fully Connected edges

- Number of nodes vary between graphs
- **Features:** jet-level and daughter-level kinematics



# Node Features

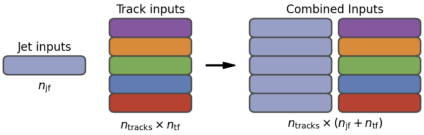
GitHub

## Jet Features

- Top-level jet kinematics
- SV-tagger output variables

## Daughter Features

- Unique daughter kinematics




ATL-PHYS-PUB-2022-027

| Jet Features | Daughter Features |
|---|---|
| Eta | E |
| pT | pT |
| | ID |
| | pX |
| **SV Tagging** | pY |
| fdrMin | pZ |
| ptSvrJet | Eta |
| nTrk | Phi |
| nTrkJet | Q |
| drSvrJet | IP |
| absQSum | IPCHI2 |
| m | IPraw |
| mCor | NNe |
| fdChi2 | NNk |
| ipChi2Sum | NNp |
| tau | NNpi |
| z | NNmu |
| pT | Chi2 |
| | QoverP |
| | trackX |
| | trackY |
| | trackZ |
| | trackVX |
| | trackVY |
| | trackVZ |
| | CaloNeutralEcal |
| | CaloNeutralHcal2Ecal |
| | CaloNeutralE49 |
| | CaloNeutralPrs |

# Sample Preparation and Definition

## Dataset

- 1.2M fully reconstructed di-jet simulated events per flavour
  - Only leading jet used for graph
- 80:20 Training and validation split

## Truth Matching

- Reco jet matched to truth jet
- Energy fraction of daughters used to select b/c/q

| General Selection Requirements |
|---|
| 20 GeV $< p_T <$ 50 GeV |
| $2.2 < \eta < 4.4$ |

## Classifier Types

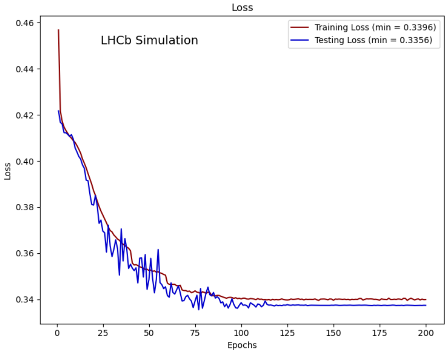
- *b* vs *q*
- *c* vs *b*

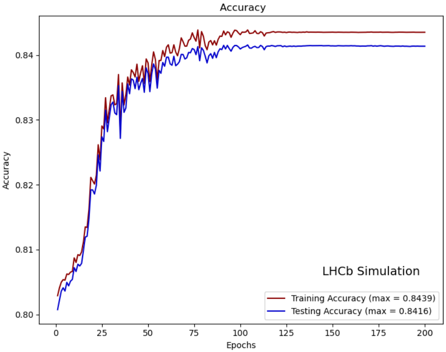
**Truth Matching**

| $b$ **Selection** | $c$ **Selection** | $q$ **Selection** |
|---|---|---|
| MC Match $= 1$ | MC Match $= 1$ | MC Match $= 1$ |
| MC Jet EfB $> 0.6$ | MC Jet EfD $> 0.6$ | MC Jet EfD $< 0.6$ |
| | | MC Jet EfB $< 0.6$ |

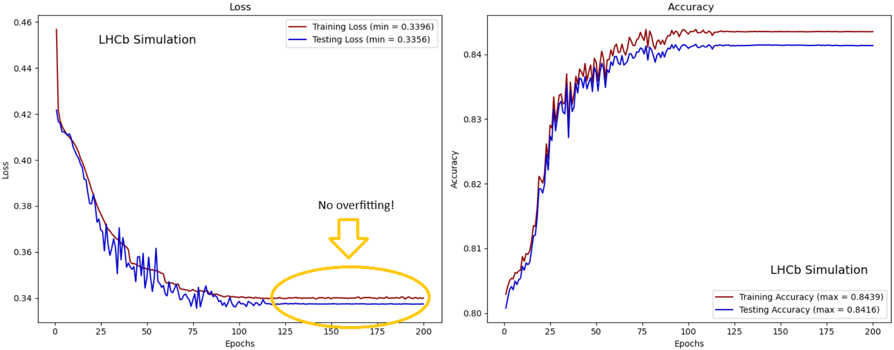
# $b$ vs. $q$ Classifier Training

90 minutes
NVIDIA GeForce RTX2080Ti

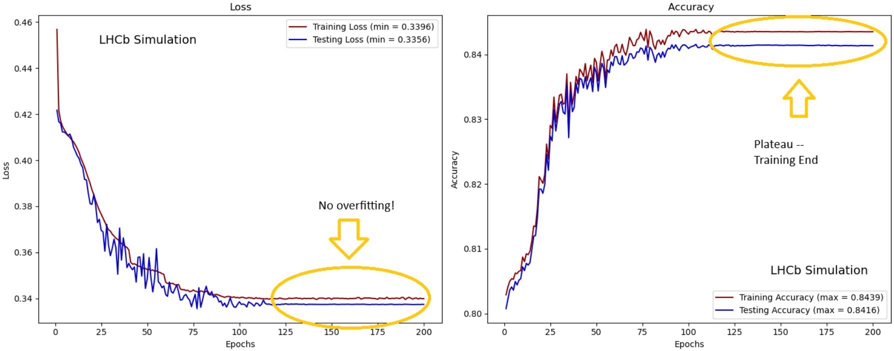


# $b$ vs. $q$ Classifier Training



# $b$ vs. $q$ Classifier Training

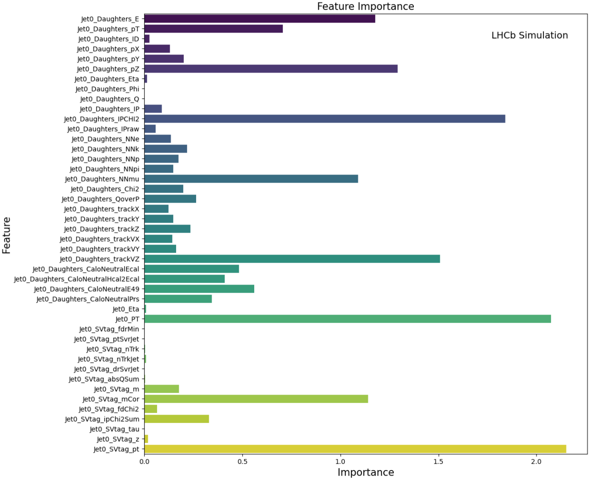


# *b* vs. *q* Feature Ranking

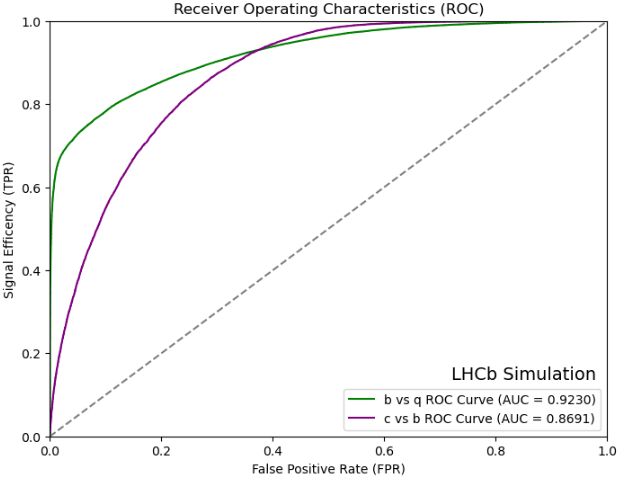
## Feature Ablation

- Remove one feature → Compare predictions

| Feature | Importance |
|---|---|
| Jet0_SVtag_pt | 2.151547 |
| Jet0_PT | 2.074978 |
| Jet0_Daughters_IPCHI2 | 1.840426 |
| Jet0_Daughters_trackVZ | 1.508102 |
| Jet0_Daughters_pZ | 1.290995 |
| Jet0_Daughters_E | 1.178953 |
| Jet0_SVtag_mCor | 1.141651 |
| Jet0_Daughters_NNmu | 1.090061 |
| Jet0_Daughters_pT | 0.705798 |
| Jet0_Daughters_CaloNeutralE49 | 0.560570 |
| Jet0_Daughters_CaloNeutralEcal | 0.482262 |
| Jet0_Daughters_CaloNeutralHcal2Ecal | 0.410697 |
| Jet0_Daughters_CaloNeutralPrs | 0.345381 |
| Jet0_SVtag_ipChi2Sum | 0.329392 |
| Jet0_Daughters_QoverP | 0.264155 |

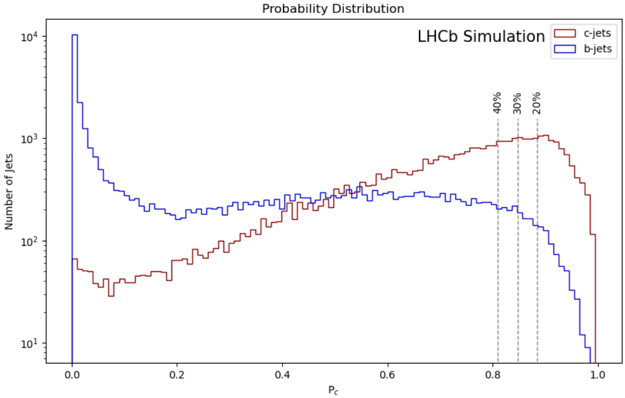


# Performance Results (WIP)



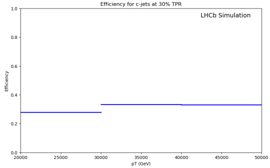

| $b$ Efficiency | $q$ FPR |
|---|---|
| 0.85 | 0.1938 |
| 0.80 | 0.1178 |
| 0.75 | 0.0685 |
| 0.70 | 0.0315 |
| 0.65 | 0.0130 |
| 0.60 | 0.0070 |
| **$c$ Efficiency** | **$b$ FPR** |
| 0.40 | 0.0559 |
| 0.30 | 0.0340 |
| 0.20 | 0.0175 |

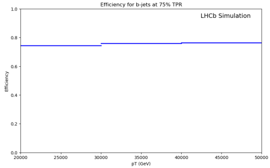
# Classifier Application




→ Charm probability distribution, including tentative charm efficiency WPs

↓ Validating $p_T$ dependence of performance



# Conclusions and Prospects

**Summary**

- First GNN Jet Tagger developed for use at LHCb
- Application significantly broadens compared to SV-required taggers
- Unique access to PID information strengthens training at LHCb

**Things to Come**

- Development to expand to more classifiers ($c$ vs. $q$, $bc$ vs. $q$)
- Develop calibration techniques $\rightarrow$ apply to analyses!

# BACKUP

# GNN Layers Using PyTorch Geometric

**SAGEConv**

- Aggregates information from neighbors – mean
- $\boldsymbol{x}_i' = \boldsymbol{W}_1 \boldsymbol{x}_1 + \boldsymbol{W}_2 \cdot \mathrm{mean}_{j \in \mathcal{N}(i)} \boldsymbol{x}_j$

**LayerNorm**

- Normalize inputs across all features independently
- $y = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} * \gamma + \beta$

**ReLU**

- Introduces non-linearity
- $R(z) = \max(0, z)$

**Dropout**

- Zero elements with probability, *p*
- Scale by factor of $\frac{1}{1-p}$

**Global Add Pooling**

- After convolutional layers, add outputs
- $\boldsymbol{r}_i = \sum_{n=1}^{N_i} \boldsymbol{x}_n$

**Linear**

- Reduce dimensionality of outputs
- $y = xA^T + b$

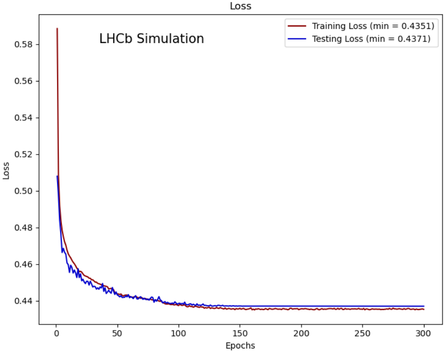
**Binary Cross Entropy Loss (with sigmoid layer)**

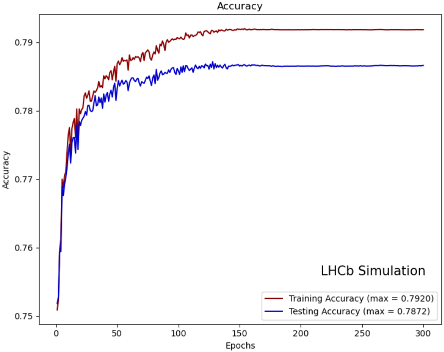
- Computes difference between prediction and truth labels
- $\ell(x, y) = L = \{l_1, \dots, l_N\}^\top, l_n = -w_n[y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n))$

**AdamW Optimizer**

- Minimizes loss function – stochastic gradient descent
- Separates weight decay from gradients

# $c$ vs. $b$ Classifier Training



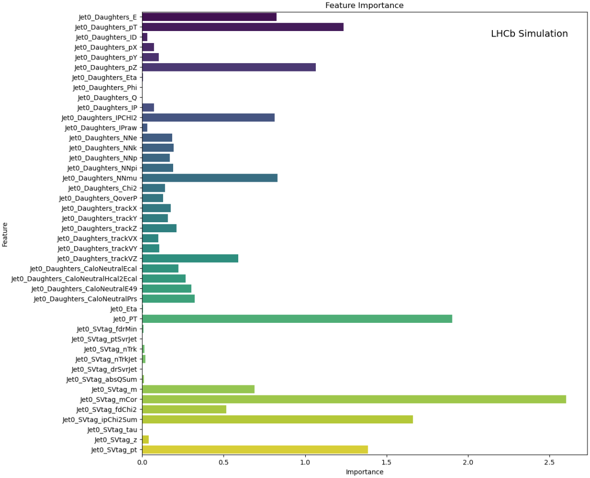



# $c$ vs. $b$ Feature Ranking

## Feature Ablation

- Remove one feature → Compare predictions

| Feature | Importance |
|---|---|
| Jet0_SVtag_mCor | 2.601232 |
| Jet0_PT | 1.903217 |
| Jet0_SVtag_ipChi2Sum | 1.661599 |
| Jet0_SVtag_pt | 1.386090 |
| Jet0_Daughters_pT | 1.234802 |
| Jet0_Daughters_pZ | 1.065050 |
| Jet0_Daughters_NNmu | 0.831524 |
| Jet0_Daughters_E | 0.825311 |
| Jet0_Daughters_IPCHI2 | 0.812331 |
| Jet0_SVtag_m | 0.689729 |
| Jet0_Daughters_trackVZ | 0.589573 |
| Jet0_SVtag_fdChi2 | 0.516657 |
| Jet0_Daughters_CaloNeutralPrs | 0.323624 |
| Jet0_Daughters_CaloNeutralE49 | 0.301616 |
| Jet0_Daughters_CaloNeutralHcal2Ecal | 0.267001 |



# Classifier Application – $b$ vs. $q$ ($P_b$)