ML-Based Top Taggers: Performance, Uncertainty and Impact of Tower Tracker Data

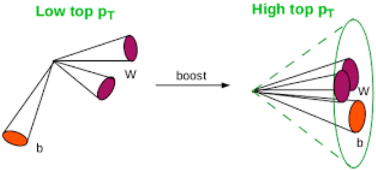
Kirtiman Ghosh, Institute Of Physics Bhubaneswar

Collaborator
Rameswar Sahu

December 19, 2023

Outline

Boosted Object Tagging



Motivation

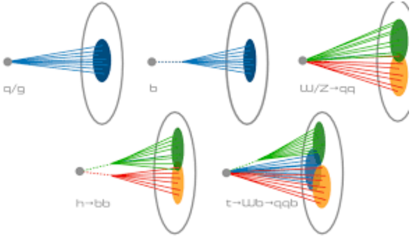
- Absence of BSM@LHC
- QCD Background
- Visible Final States

Cut-based classifiers

Based on jet shape observables like jet mass, N-subjettiness, etc. HEPTopTagger, Johns Hopkins Tagger, YSplitter etc.

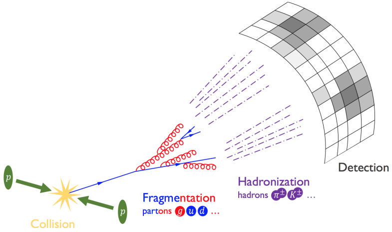
ML-based classifiers

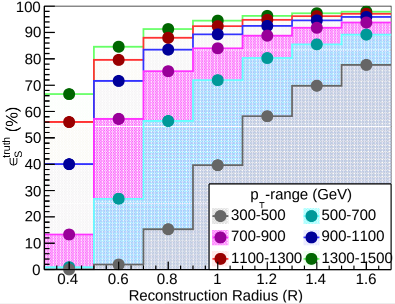
BDT-classifiers, CNN/GNN-classifiers etc.



Objectives

- Study the effect of tracking information in determining the performance of top taggers
- Study the Systematic uncertainty arising from the MC generators.
- Impact of truth level identification criteria.
- Study the variation in performance of classifiers with transverse momentum of the fat-jets.





Dataset

Top Fat-Jets

- Process Considered :

$$pp \to t(\to bq\bar{q}')\ \bar{t}(\to \bar{b}q\bar{q}')$$

- Truth-level matching: All three top decay products (at Parton level) must lie inside the cone of the fat jet.

QCD Fat-Jets

$$pp \to j\ j$$

where j = u, d, c, s, g and their anti-particles

- No Truth-level matching.

Claorimeter based dataset

Contains information of the energy deposits in Ecal and Hcal.

Tracker based dataset

- Includes tracking information
- Uses particle flow algorithm to match tracks with calorimeter energy deposits
- In the end we have three classes of data :
  - Charged particles
  - Photons
  - Neutral Hadrons

BDT Classifier

BDT_{calo}

- Utilizes High-level features like :
  - The Jet Mass
  - The N-subjettiness (τ_{43}, τ_{32}, τ_{21})
  - *b*-tag

- Trained using the TMVA 4.3 toolkit in ROOT 6.24

BDT_{trck}

- Extends the previous set by including additional track-based HLFs :
  - # of tracks inside a jet.
  - $w_{trk} = \frac{\sum_{trk \in J} p_{T,trk} \Delta R_{trk,J}}{\sum_{trk \in J} p_{T,trk}}$
  - $w_{calo} = \frac{\sum_{i \in J} p_{T,i} \Delta R_{i,J}}{\sum_{i \in J} p_{T,i}}$
  - .....
  - These extra features are defined for each sub-jet inside the fat-jet.
  - We consider only the first three leading sub-jets.
  - In the absence of three sub-jets, the missing variables are zero-padded.

CNN Classifier

CNN_{calo}

- Single layered 64 × 64 images.
- Contains transverse energy of the calorimeter cells as pixel intensities.

- Preprocessing Steps :
  - Translation.
  - Reflection
  - Rotation

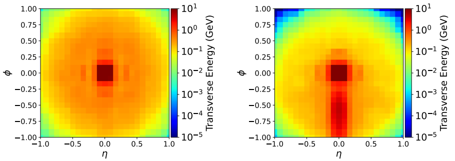
Model

- 10-layered ResNet

CNN_{trck}

- Two layered 64 × 64 images.
- First layer Contains the transverse energy of the photons and neutral hadrons as pixel intensities.
- second layer Contains transverse energy of the charged particles as pixel intensities.

Top Images after translation and rotation



GNN Classifier

- LorentzNet with 6 Lorentz Group equivariant Blocks
- Graph level classifier
- Uses the four-momentum of the jet constituents as node coordinates and the charge of the constituents as node embeddings.
- For each fat-jet, we store the four-momentum and charge of 200 constituents ordered by p_T.
- for fat-jets with less than 200 constituents, the missing entries are zero-padded.
- GNN_{calo} considers only the calorimeter energy deposits (from both charged and neutral particles) as jet constituents
- GNN_{trck} uses Charge particles, photons and neutral hadrons as jet constituents.
- For the charged particles, we replace the mass with zero by hand to avoid implementing any specific particle identification algorithm.

Composite Classifier

Models

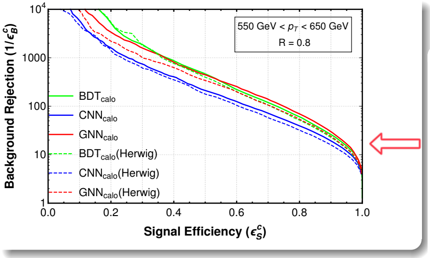
- $C_{calo}B_{calo}$, $C_{calo}B_{trck}$
- $C_{trck}B_{calo}$, $C_{trck}B_{trck}$
- $G_{calo}B_{calo}$, $G_{calo}B_{trck}$
- $G_{trck}B_{calo}$, $G_{trck}B_{trck}$

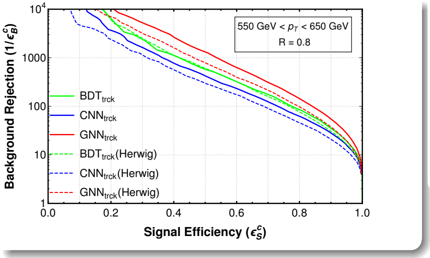
- First trains a CNN/GNN with LLFs and Uses the output score as a new HLF in addition to other important HLFs in a BDT.
- The Classifiers from top to bottom are ordered according to their complexity.

Motivation

- To introduce HLFs effective in differentiating quark jets from gluon jets that LLF-based classifiers like CNN or GNN may or may not have learned.
- To re-introduce HLFs that are lost during the preprocessing of data in CNN/GNN.
- Provides better control over uncertainty originating from the use of showering and hadronization models, i.e., different Monte Carlo event generators.

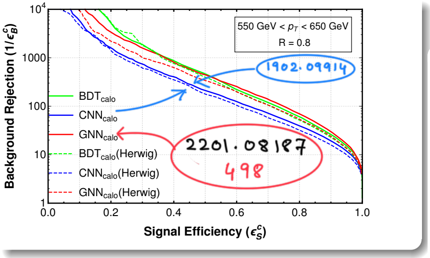
Classifiers Performance (Simple Classifiers)

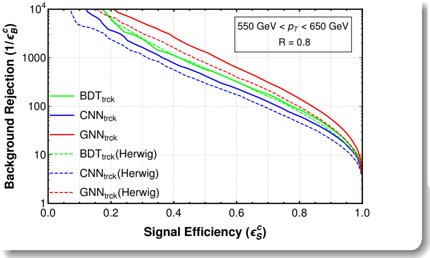




Classifier	$1/\epsilon_B^c(\epsilon_S^c = 0.7)$	$1/\epsilon_B^c(\epsilon_S^c = 0.5)$
BDT_{calo}	119(105)	467(398)
CNN_{calo}	70(57)	211(178)
GNN_{calo}	139(106)	444(341)
BDT_{trck}	175(159)	579(610)
CNN_{trck}	124(90)	423(299)
GNN_{trck}	311(214)	1322(789)
$C_{calo}B_{calo}$	176(175)	682(619)
$C_{calo}B_{trck}$	208(204)	811(737)
$C_{trck}B_{calo}$	249(218)	1023(768)
$C_{trck}B_{trck}$	257(221)	995(799)
$G_{calo}B_{calo}$	260(241)	969(842)
$G_{calo}B_{trck}$	278(256)	1141(894)
$G_{trck}B_{calo}$	489(397)	1641(1604)
$G_{trck}B_{trck}$	493(399)	1736(1666)

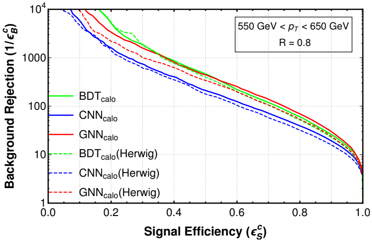
Classifiers Performance (Simple Classifiers)

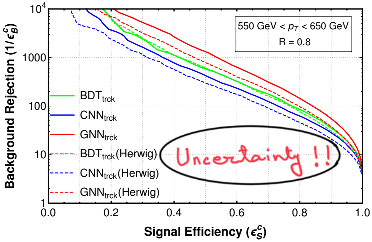




Classifier	$1/\epsilon_B^c(\epsilon_S^c = 0.7)$	$1/\epsilon_B^c(\epsilon_S^c = 0.5)$
BDT_{calo}	119(105)	467(398)
CNN_{calo}	70(57)	211(178)
GNN_{calo}	139(106)	444(341)
BDT_{trck}	175(159)	579(610)
CNN_{trck}	124(90)	423(299)
GNN_{trck}	311(214)	1322(789)
$C_{calo}B_{calo}$	176(175)	682(619)
$C_{calo}B_{trck}$	208(204)	811(737)
$C_{trck}B_{calo}$	249(218)	1023(768)
$C_{trck}B_{trck}$	257(221)	995(799)
$G_{calo}B_{calo}$	260(241)	969(842)
$G_{calo}B_{trck}$	278(256)	1141(894)
$G_{trck}B_{calo}$	489(397)	1641(1604)
$G_{trck}B_{trck}$	493(399)	1736(1666)

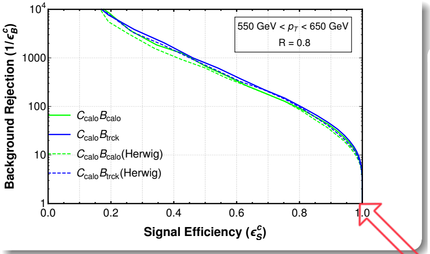
Classifiers Performance (Simple Classifiers)

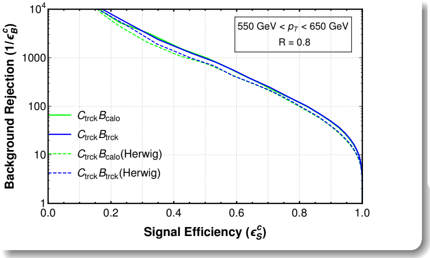




Classifier	$1/\epsilon_B^c(\epsilon_S^c = 0.7)$	$1/\epsilon_B^c(\epsilon_S^c = 0.5)$
BDT_{calo}	119(105)	467(398)
CNN_{calo}	70(57)	211(178)
GNN_{calo}	139(106)	444(341)
BDT_{trck}	175(159)	579(610)
CNN_{trck}	124(90)	423(299)
GNN_{trck}	311(214)	1322(789)
$C_{calo}B_{calo}$	176(175)	682(619)
$C_{calo}B_{trck}$	208(204)	811(737)
$C_{trck}B_{calo}$	249(218)	1023(768)
$C_{trck}B_{trck}$	257(221)	995(799)
$G_{calo}B_{calo}$	260(241)	969(842)
$G_{calo}B_{trck}$	278(256)	1141(894)
$G_{trck}B_{calo}$	489(397)	1641(1604)
$G_{trck}B_{trck}$	493(399)	1736(1666)

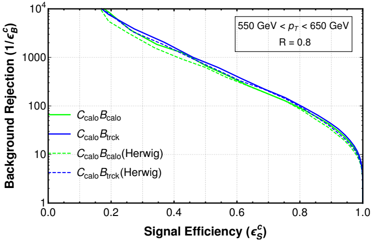
Classifiers Performance (Composite Classifiers CNN-based)

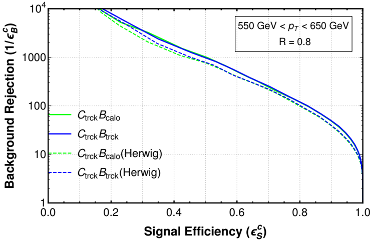




Classifier	$1/\epsilon_B^c(\epsilon_S^c = 0.7)$	$1/\epsilon_B^c(\epsilon_S^c = 0.5)$
BDT_{calo}	119(105)	467(398)
CNN_{calo}	70(57)	211(178)
GNN_{calo}	139(106)	444(341)
BDT_{trck}	175(159)	579(610)
CNN_{trck}	124(90)	423(299)
GNN_{trck}	311(214)	1322(789)
$C_{calo}B_{calo}$	176(175)	682(619)
$C_{calo}B_{trck}$	208(204)	811(737)
$C_{trck}B_{calo}$	249(218)	1023(768)
$C_{trck}B_{trck}$	257(221)	995(799)
$G_{calo}B_{calo}$	260(241)	969(842)
$G_{calo}B_{trck}$	278(256)	1141(894)
$G_{trck}B_{calo}$	489(397)	1641(1604)
$G_{trck}B_{trck}$	493(399)	1736(1666)

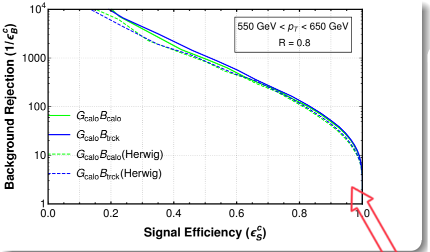
Classifiers Performance (Composite Classifiers CNN-based)

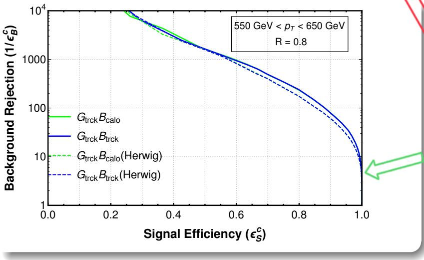




Classifier	$1/\epsilon_B^c(\epsilon_S^c = 0.7)$	$1/\epsilon_B^c(\epsilon_S^c = 0.5)$
BDT_{calo}	119(105)	467(398)
CNN_{calo}	70(57)	211(178)
GNN_{calo}	139(106)	444(341)
BDT_{trck}	175(159)	579(610)
CNN_{trck}	124(90)	423(299)
GNN_{trck}	311(214)	1322(789)
$C_{calo}B_{calo}$	176(175)	682(619)
$C_{calo}B_{trck}$	208(204)	811(737)
$C_{trck}B_{calo}$	249(218)	1023(768)
$C_{trck}B_{trck}$	257(221)	995(799)
$G_{calo}B_{calo}$	260(241)	969(842)
$G_{calo}B_{trck}$	278(256)	1141(894)
$G_{trck}B_{calo}$	489(397)	1641(1604)
$G_{trck}B_{trck}$	493(399)	1736(1666)

Classifiers Performance (Composite Classifiers GNN-based)





Classifier	$1/\epsilon_B^c(\epsilon_S^c = 0.7)$	$1/\epsilon_B^c(\epsilon_S^c = 0.5)$
BDT_{calo}	119(105)	467(398)
CNN_{calo}	70(57)	211(178)
GNN_{calo}	139(106)	444(341)
BDT_{trck}	175(159)	579(610)
CNN_{trck}	124(90)	423(299)
GNN_{trck}	311(214)	1322(789)
$C_{calo}B_{calo}$	176(175)	682(619)
$C_{calo}B_{trck}$	208(204)	811(737)
$C_{trck}B_{calo}$	249(218)	1023(768)
$C_{trck}B_{trck}$	257(221)	995(799)
$G_{calo}B_{calo}$	260(241)	969(842)
$G_{calo}B_{trck}$	278(256)	1141(894)
$G_{trck}B_{calo}$	489(397)	1641(1604)
$G_{trck}B_{trck}$	493(399)	1736(1666)

Systematic Uncertainty

$C_{trck}B_{calo}$: Ranking Of Variables

Variable	Ranking
M	0.3625
score	0.309
$\tau_{2,1}$	0.099
τ_{32}	0.09
b-tag	0.0714
τ_{43}	0.0676

$G_{trck}B_{calo}$: Ranking Of Variables

Variable	Ranking
score	0.3517
M	0.3142
$\tau_{2,1}$	0.0968
τ_{32}	0.093
b-tag	0.075
τ_{43}	0.069

Classifier	$1/\epsilon_B^c(\epsilon_S^c = 0.7)$	$1/\epsilon_B^c(\epsilon_S^c = 0.5)$
BDT_{calo}	119(105)	467(398)
CNN_{calo}	70(57)	211(178)
GNN_{calo}	139(106)	444(341)
BDT_{trck}	175(159)	579(610)
CNN_{trck}	124(90)	423(299)
GNN_{trck}	311(214)	1322(789)
$C_{calo}B_{calo}$	176(175)	682(619)
$C_{calo}B_{trck}$	208(204)	811(737)
$C_{trck}B_{calo}$	249(218)	1023(768)
$C_{trck}B_{trck}$	257(221)	995(799)
$G_{calo}B_{calo}$	260(241)	969(842)
$G_{calo}B_{trck}$	278(256)	1141(894)
$G_{trck}B_{calo}$	489(397)	1641(1604)
$G_{trck}B_{trck}$	493(399)	1736(1666)

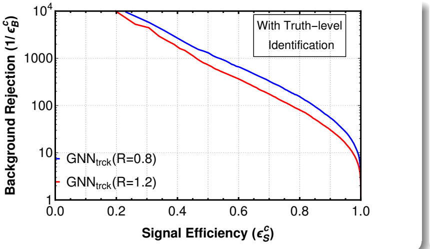
Enhanced performances & reduced uncertainties

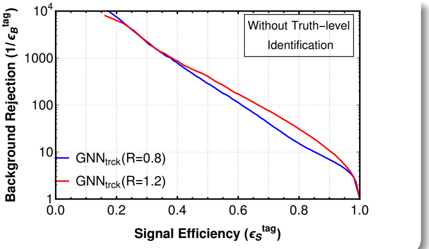
- We expect GNN_{trck} with the full tracking information to be efficient enough to provide the best performance.
- The observed reduction in performance is because of the masking of the mass information of the charged track.

MC generator	GNN_{trck}	$G_{trck}B_{trck}$
Pythia8	1769	1736
Herwig7	1025	1666

Classifier	$1/\epsilon_B^c(\epsilon_S^c = 0.7)$	$1/\epsilon_B^c(\epsilon_S^c = 0.5)$
BDT_{calo}	119(105)	467(398)
CNN_{calo}	70(57)	211(178)
GNN_{calo}	139(106)	444(341)
BDT_{trck}	175(159)	579(610)
CNN_{trck}	124(90)	423(299)
GNN_{trck}	311(214)	1322(789)
$C_{calo}B_{calo}$	176(175)	682(619)
$C_{calo}B_{trck}$	208(204)	811(737)
$C_{trck}B_{calo}$	249(218)	1023(768)
$C_{trck}B_{trck}$	257(221)	995(799)
$G_{calo}B_{calo}$	260(241)	969(842)
$G_{calo}B_{trck}$	278(256)	1141(894)
$G_{trck}B_{calo}$	489(397)	1641(1604)
$G_{trck}B_{trck}$	493(399)	1736(1666)

Truth-Level Tagging





Variable	$1/\epsilon_B^c$ ($\epsilon_s^c = 50\%$)	$1/\epsilon_B^{tag}$ ($\epsilon_s^{tag} = 50\%$)
$R = 0.8$	1298	274
$R = 1.2$	711	424

p_T - Dependance

With TLT : $1/\epsilon_B^c$ ($\epsilon_S^c = 50\%$)

p_T [GeV]	BDT_{calo}	BDT_{trck}	CNN_{trck}	GNN_{trck}	$C_{trck}B_{calo}$	$G_{trck}B_{calo}$
300-500	388	456	159	587	762	1413
500-700	136	276	184	765	455	1178
700-900	168	345	278	845	538	1409
900-1100	79	247	256	971	466	1175
1100-1300	56	167	214	882	318	872
1300-1500	39	127	217	877	273	850

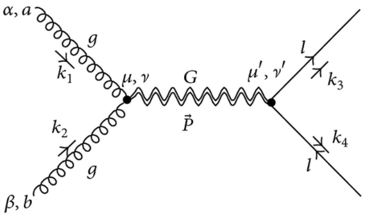
Without TLT : $1/\epsilon_B^{tag}$ ($\epsilon_S^{tag} = 50\%$)

p_T [GeV]	BDT_{calo}	BDT_{trck}	CNN_{trck}	GNN_{trck}	$C_{trck}B_{calo}$	$G_{trck}B_{calo}$
300-500	95	119	54	121	157	250
500-700	83	152	110	303	243	581
700-900	84	166	147	421	258	582
900-1100	57	148	168	534	279	789
1100-1300	45	124	157	540	234	651
1300-1500	34	101	167	609	217	662

Future Directions

Application for BSM searches

Supersymmetry, extra-dimensional models, leptoquark models, different gauge and field extensions of the SM, etc.



Classifier development

Systematic Uncertainties, Performance in the high-p_T region, variable radius jets, etc.

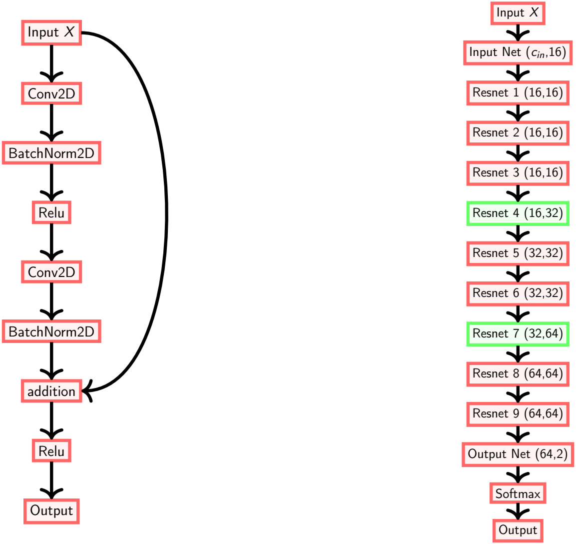
Questions?

Questions?

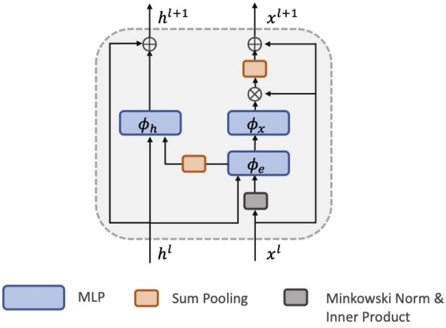
Backup

CNN Model

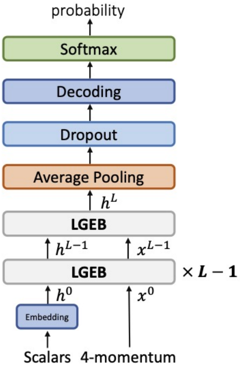
ResNet Block (left) and full ResNet model (right) :



GNN Model



Lorentz Group Equivariant Block (LGEB)



LorentzNet