Pushing the Limit of Jet Tagging With Graph Neural Networks

Huilin Qu (CERN)

ML4Jets2021

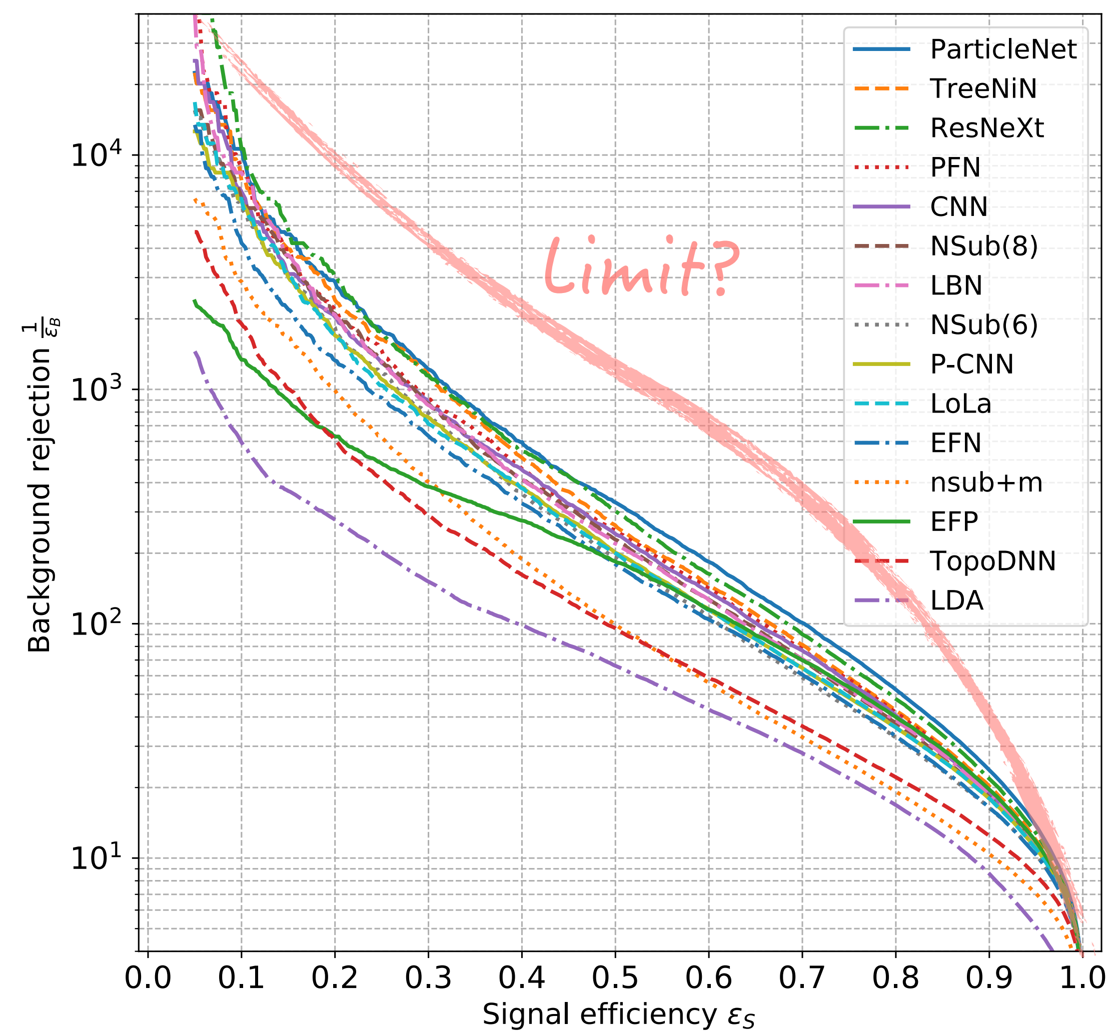
July 7, 2021

Bag of Tricks for a Better Jet Tagger

Introduction

- Jet tagging
  - powerful hammer in experimentalists' toolbox
  - fun playground for ML enthusiasts
- Graph neural networks (GNNs) have shown lots of potential for jet tagging
  - top tagging benchmark (2019)
    - GNN-based ParticleNet [1902.08570] achieves the best performance
  - since then:
    - ABCNet [2001.05311], Point Cloud Transformer [2102.05073]: better performance in quark/gluon tagging
    - LundNet [2012.08526]: surpass ParticleNet in top tagging
    - ...
  - now: can we do even better? and how?

"Top Tagging Landscape"

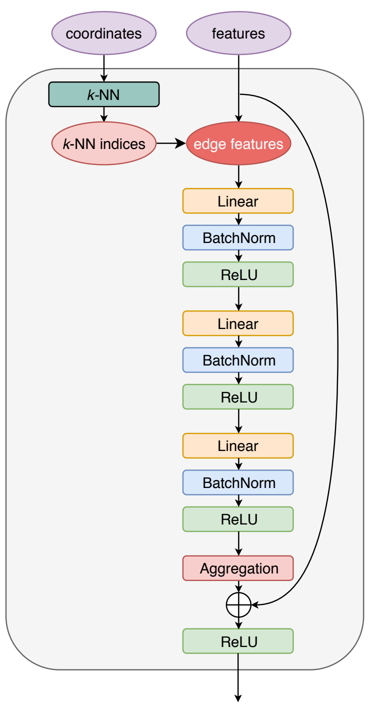


G. Kasieczka et al. [1902.09914]

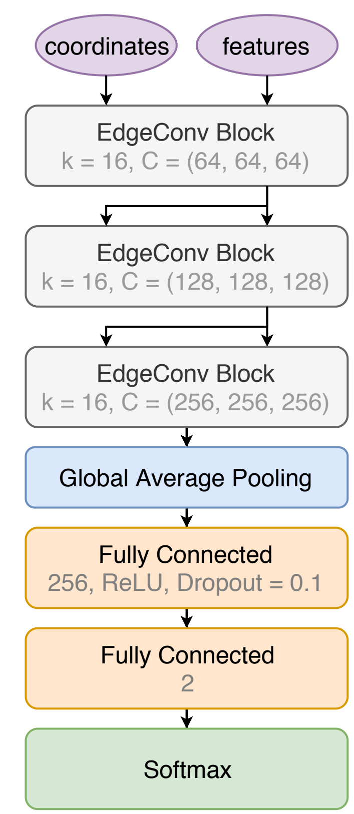
RECAP: *PARTICLENET*

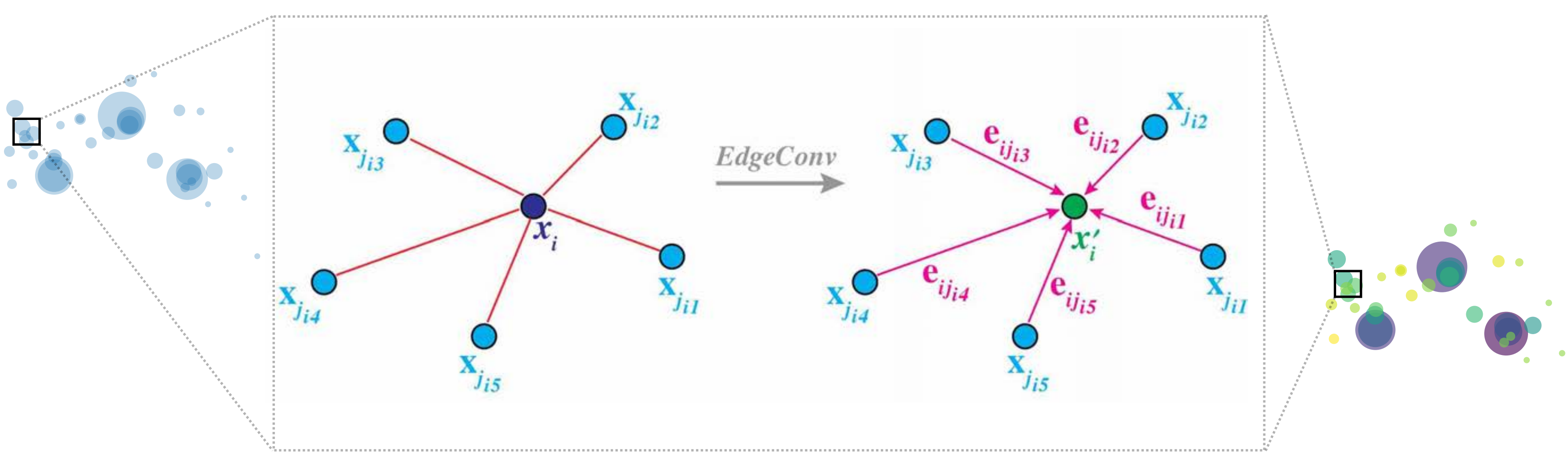
- ParticleNet
  - jet treated as a permutation-invariant point cloud
  - customized graph neural network architecture for jet tagging based on Dynamic Graph CNN [Y. Wang et al., *arXiv:1801.07829*]
- Key building block: *EdgeConv*
  - treating a point cloud as a graph: each point is a vertex
    - for each point, a local patch is defined by finding its k-nearest neighbors
  - designing a permutation-invariant "convolution" function
    - learn an "edge feature" for each center-neighbor pair: $e_{ij} = \mathrm{MLP}(x_i, x_j)$
      - same MLP for all neighbor points, and all center points, for symmetry
    - aggregate the edge features in a symmetric way: $x_i' = \mathrm{mean}_j\ e_{ij}$
- EdgeConv can be stacked to form a deep network
  - learning both local and global structures, in a hierarchical way

EdgeConv block



ParticleNet architecture

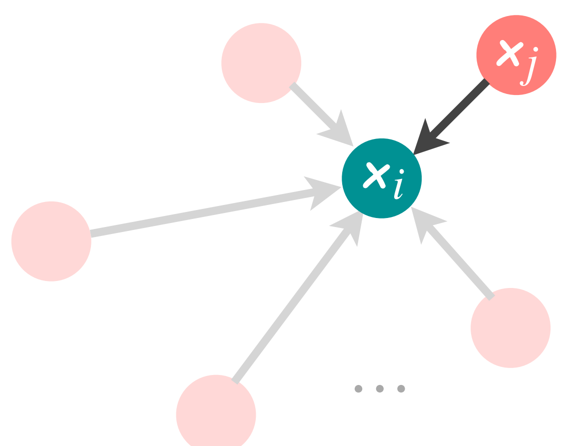




ParticleNeXt: Pairwise Features

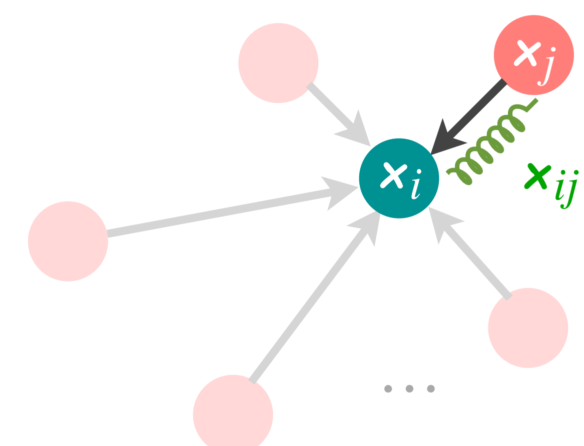
- ParticleNeXt: next-generation of ParticleNet, for better performance
- The first enhancement is the addition of (explicit) pairwise features on the edges

ParticleNet



$$e_{ij} = MLP(x_i, x_j)$$

ParticleNeXt



$$e_{ij} = MLP(x_i, x_j, x_{ij})$$

- Examples of pairwise features:

$$\Delta_{ij}^2 \equiv (y_i - y_j)^2 + (\phi_i - \phi_j)^2, \quad m^2 \equiv (p_i + p_j)^2,$$

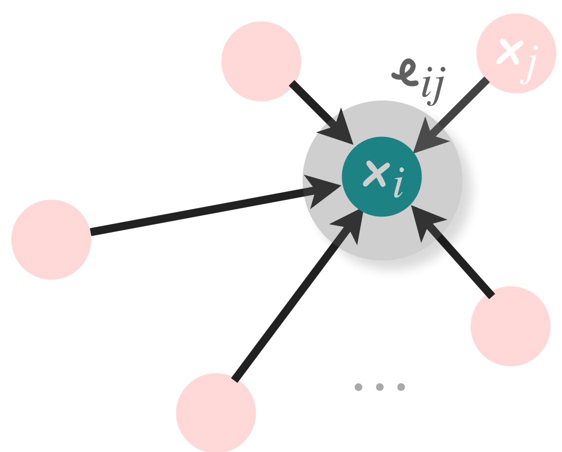
$$k_T \equiv \min(p_{T,i}, p_{T,j})\,\Delta_{ij}, \quad z \equiv \frac{\min(p_{T,i}, p_{T,j})}{p_{T,i} + p_{T,j}}$$

(use the logarithm to improve stability of the training)

ParticleNeXt: Attentive Pooling

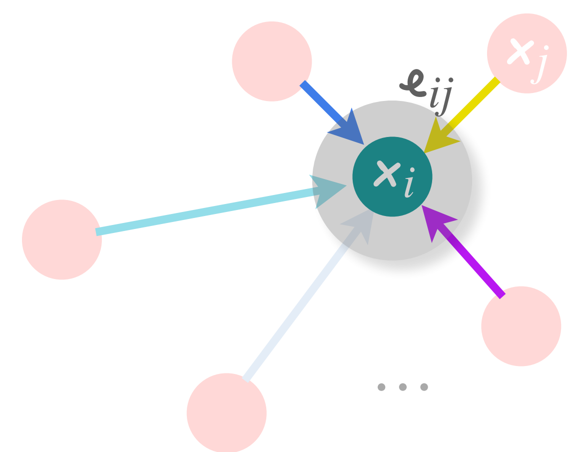
- Use attention-based pooling to increase the expressive power
  - for both the **local neighborhood pooling**, and the final **global pooling**

ParticleNet



$$z_i = \mathrm{mean}_j(e_{ij})$$

ParticleNeXt



$$\mathrm{attn}_{ij} = \mathrm{MLP}(e_{ij})$$

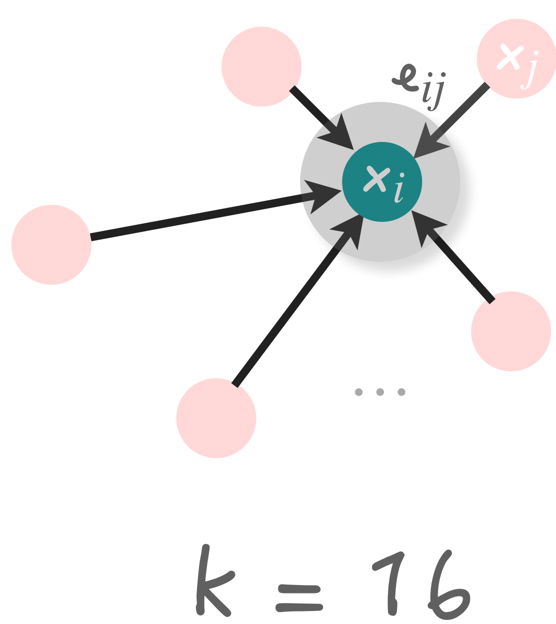
$$w_{ij} = \mathrm{softmax}_j(\mathrm{attn}_{ij})$$

$$z_i = \Sigma_j(w_{ij}\, e_{ij})$$

ParticleNeXt: Multi-Scale Aggregation

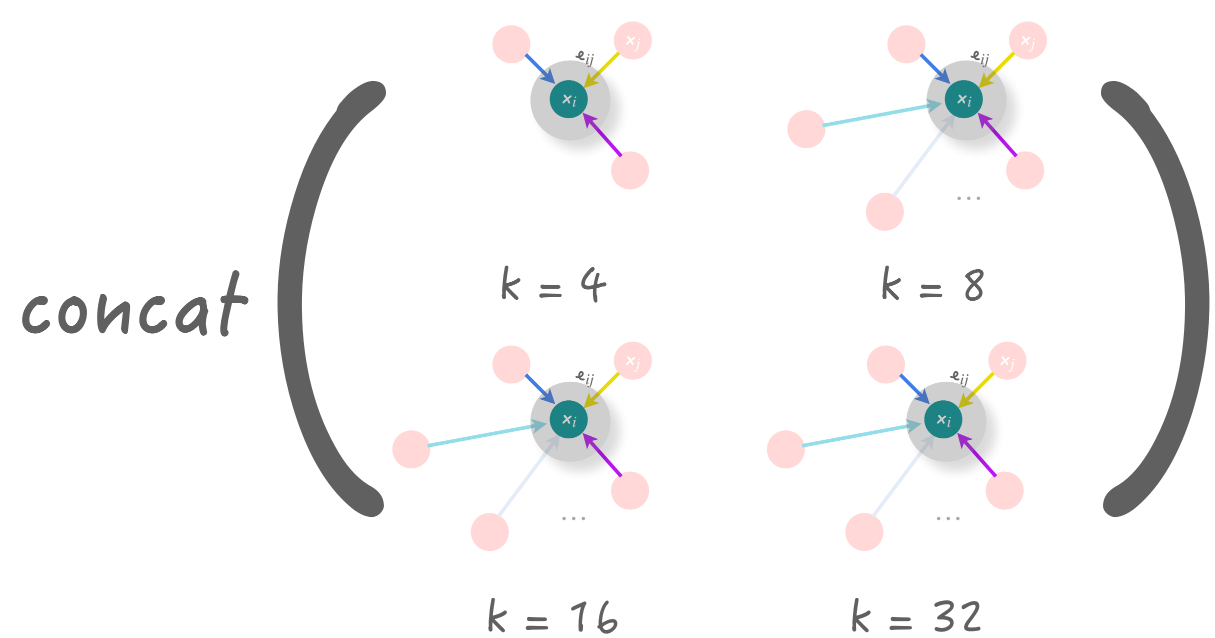
- Introduce multi-scale aggregation to better capture both short- and long-range correlations
  - perform local aggregation for the 4, 8, 16 and 32 nearest neighbors (with different attentive pooling) and combine the 4 aggregated representations with a MLP
  - on the other hand: remove dynamic kNN (based on learned features), i.e., use only kNN in $\eta-\phi$ space, to reduce computational cost
    - in this case the kNN needs to be performed only once, and then the graph connectivity is fixed

ParticleNet



$$x_i' = z_i$$

ParticleNeXt

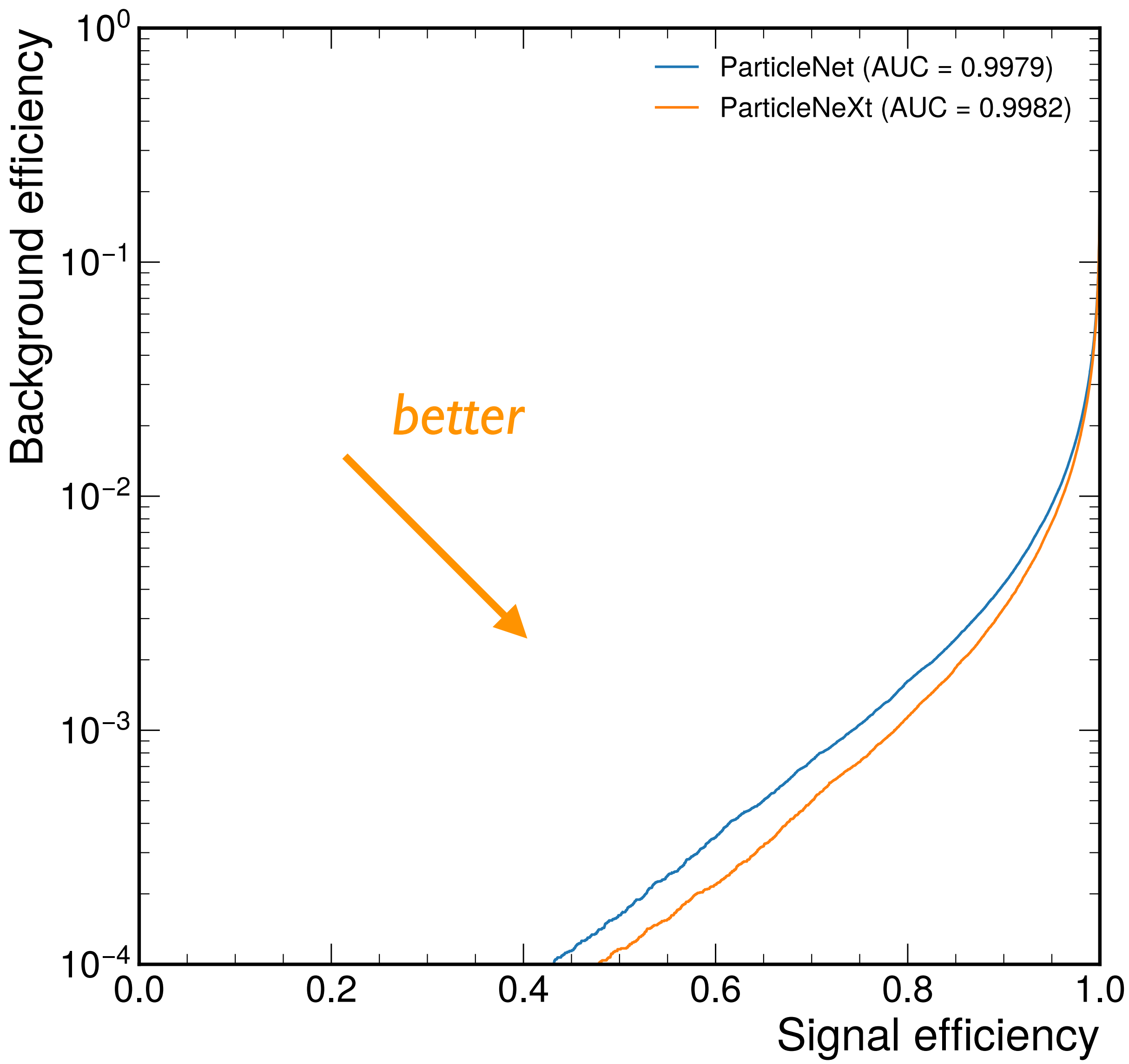


$$x_i' = \mathrm{MLP}(z_i^{\mathrm{concat}})$$

DATASET

- A new jet tagging dataset was generated for the development of ParticleNeXt
  - all events are generated with MadGraph5_aMC@NLO v3.1.1 at LO and interfaced with Pythia v8.245 for parton shower (w/ the default tune and MPI enabled)
  - fast detector simulation w/ Delphes v3.5.0, using the CMS card
    - tracking resolution parametrization based on the CMS Run1 performance [1405.6569]
  - jets clustered from the Delphes e-flow objects using the anti-kt algorithm w/ R=0.8
    - only consider jets w/ 500 < p_T 1000 GeV, and $|\eta|$ < 2
    - input features for each jet constituent particle: 4-momenta, PID, **impact parameters and errors**
  - top-tagging benchmark:
    - Top quark jets: $pp \to t\bar{t}\ (t \to bW, W \to qq')$
      - truth matching criteria: ΔR(jet, q) < 0.8 for all three quarks from hadronic top decay
    - QCD jets: $pp \to Z(\to\nu\bar{\nu}) + j\,(j = u, d, s, c, b, g)$
  - Higgs-tagging benchmark:
    - Higgs boson jets: $pp \to hh\ (h \to b\bar{b})$
      - truth matching criteria: ΔR(jet, b) < 0.8 for both quarks from the Higgs decay
    - QCD jets: $pp \to Z(\to\nu\bar{\nu}) + j\,(j = u, d, s, c, b, g)$

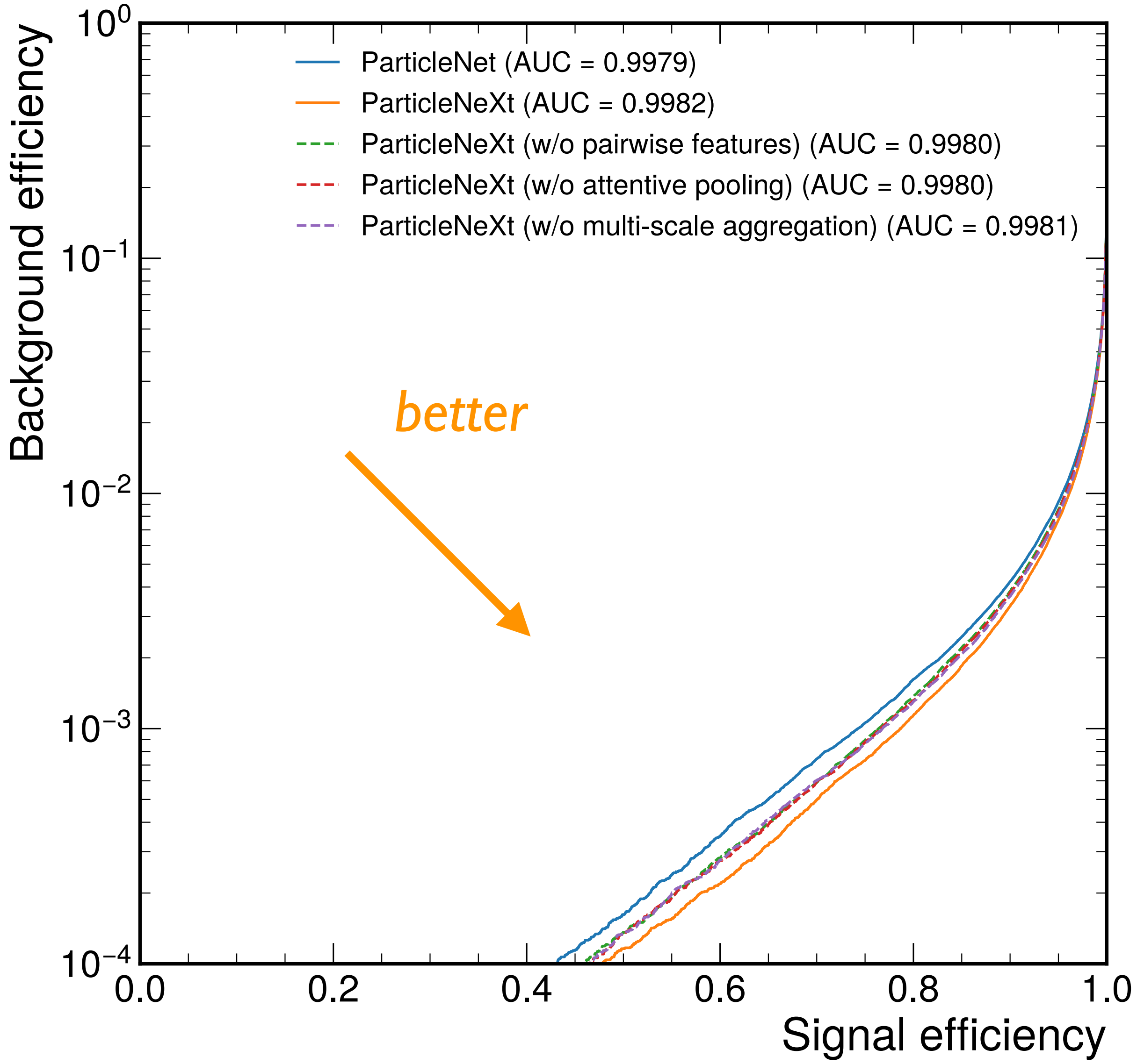
PERFORMANCE: TOP TAGGING



- Training/validation/test splitting:
  - 1.6M / 0.4M / 2M
- Training repeated for 3 times starting from randomly initialized weights
  - the median-accuracy training is reported, and the standard deviation of the 3 trainings is quoted as the uncertainty
- Significant improvement in background rejection w/ ParticleNeXt
  - ~50% higher BKG rejection (@ϵ_S = 70%)
  - computational cost still under control

	Accuracy	AUC	$1/\varepsilon_b$ at $\varepsilon_s = 70\%$	$1/\varepsilon_b$ at $\varepsilon_s = 50\%$	Parameters	Inference time (CPU)	Inference time (GPU)	Training time (GPU)
ParticleNet	0.980	0.9979	1342 ± 4	6173 ± 425	366k	23 ms	0.30 ms	1.0 ms
ParticleNeXt	**0.981**	**0.9982**	$\mathbf{2008 \pm 75}$	$\mathbf{8621 \pm 309}$	560k	30 ms	0.54 ms	1.7 ms

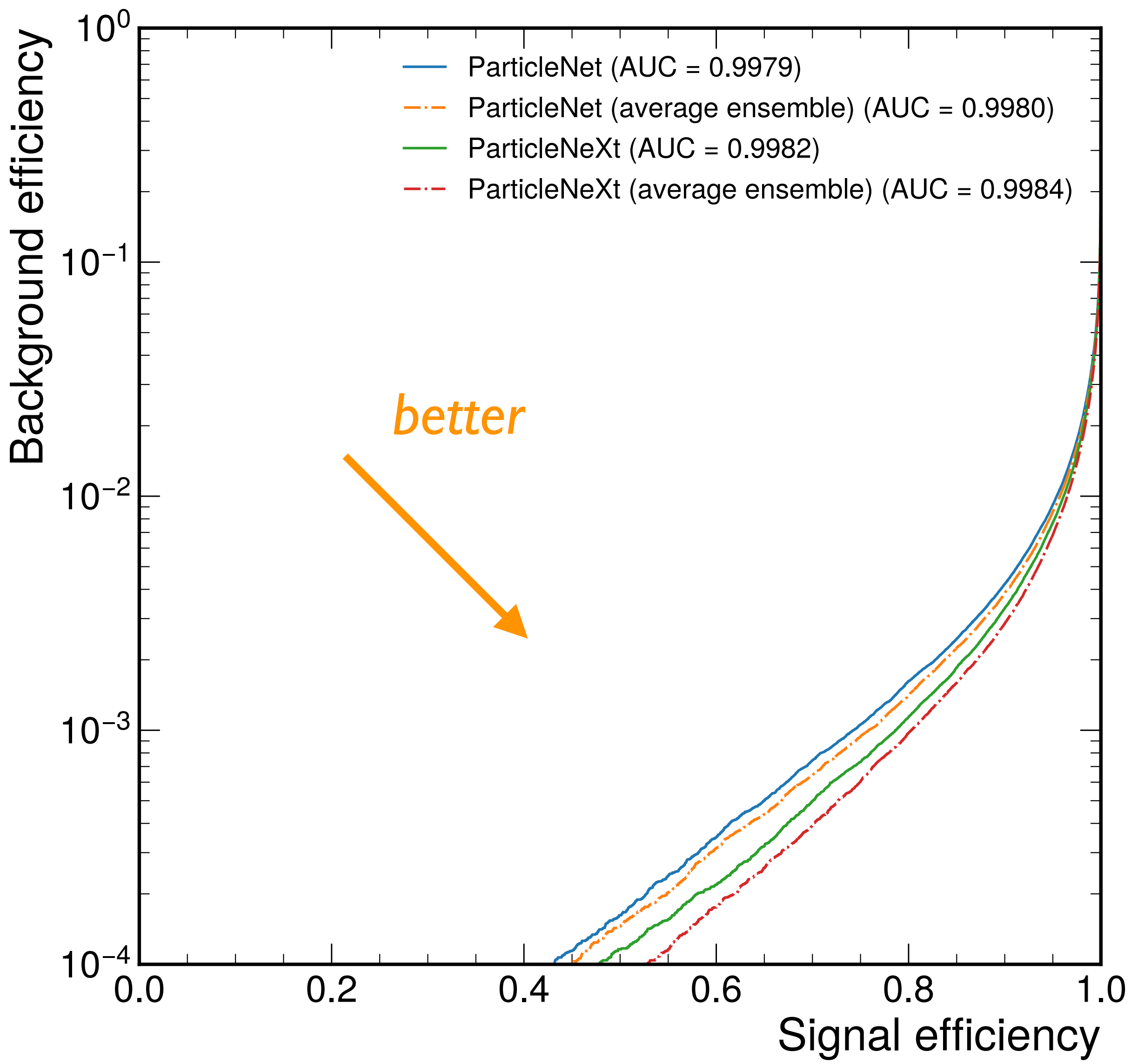
Ablation Study



- Investigated the effects of the new features of ParticleNeXt by removing each of them and repeat the training
  - all the new features contribute
  - ~20% loss in BKG rejection if any of the three is removed

	Accuracy	AUC	$1/\varepsilon_b$ at $\varepsilon_s = 70\%$	$1/\varepsilon_b$ at $\varepsilon_s = 50\%$
ParticleNet	0.980	0.9979	1342 ± 4	6173 ± 425
ParticleNeXt	**0.981**	**0.9982**	$\mathbf{2008 \pm 75}$	$\mathbf{8621 \pm 309}$
ParticleNeXt (w/o pairwise features)	0.980	0.9980	1695 ± 70	7353 ± 193
ParticleNeXt (w/o attentive pooling)	0.980	0.9981	1689 ± 72	7463 ± 696
ParticleNeXt (w/o multi-scale aggregation)	0.981	0.9980	1664 ± 57	7407 ± 193

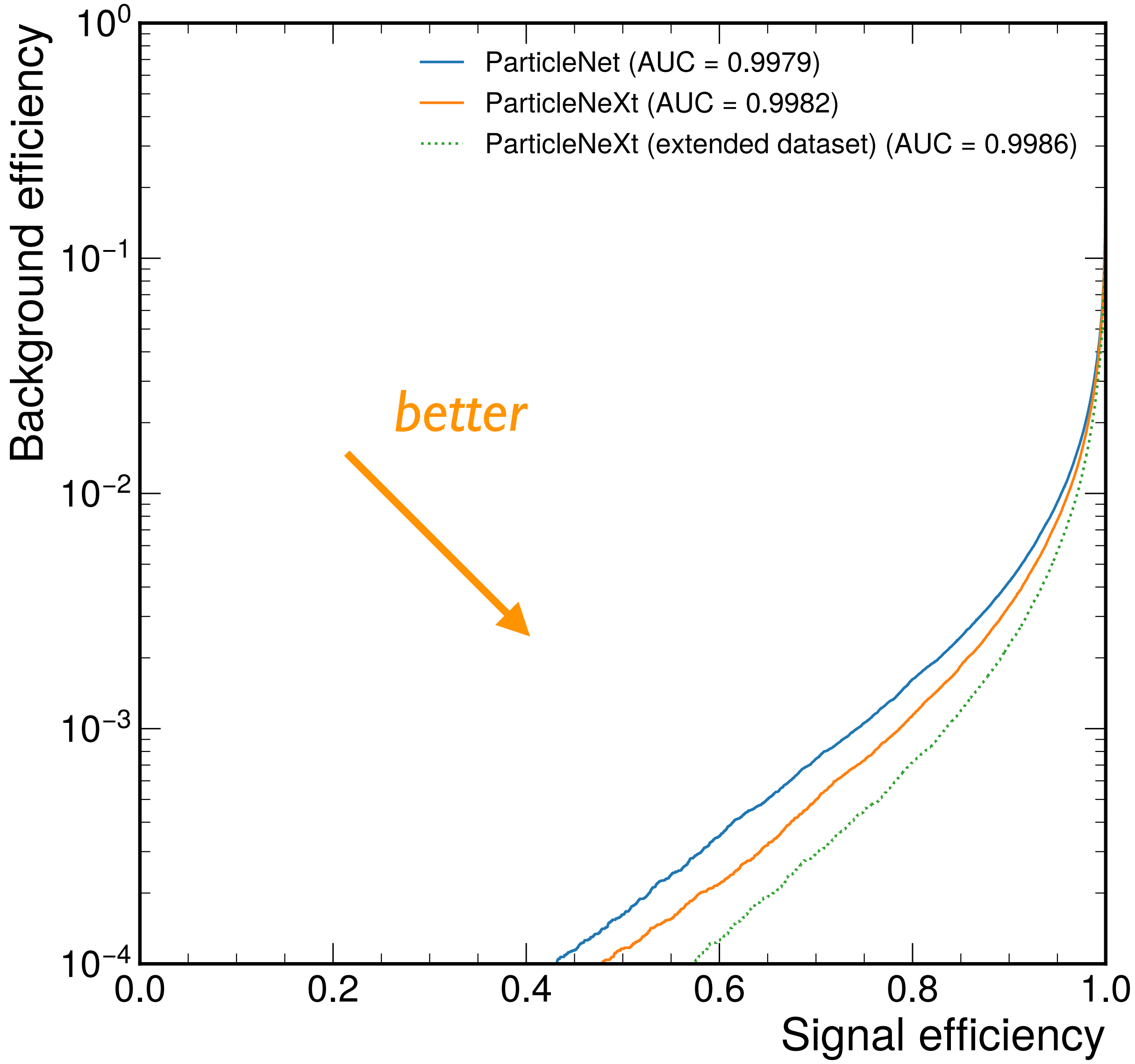
Model Ensembling



- Model ensembling still helps, even for the new ParticleNeXt
  - ensembling method: average the DNN outputs from the 3 independent trainings
  - ~30% improvement for ParticleNeXt with the 3-model ensemble
    - ~15% for ParticleNet

	Accuracy	AUC	$1/\varepsilon_b$ at $\varepsilon_s = 70\%$	$1/\varepsilon_b$ at $\varepsilon_s = 50\%$
ParticleNet	0.980	0.9979	1342 ± 4	6173 ± 425
ParticleNeXt	0.981	0.9982	2008 ± 75	8621 ± 309
ParticleNet (average ensemble)	0.980	0.9980	1558	6897
ParticleNeXt (average ensemble)	**0.982**	**0.9984**	**2558**	**11494**

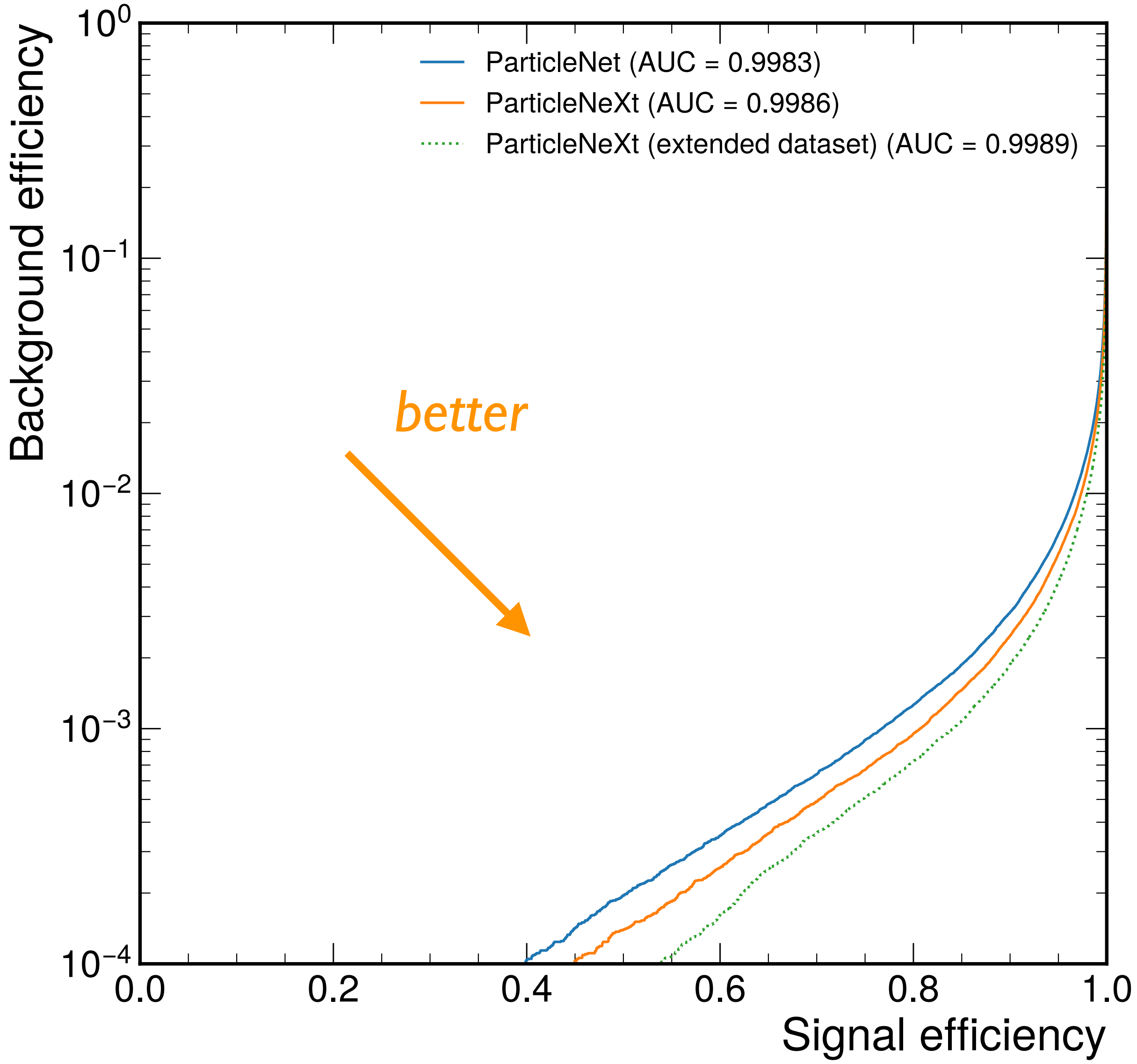
EXTENDED TRAINING DATASET



- Training on a larger dataset
  - training/validation/test splitting:
    - 10M / 1M / 2M
  - i.e., 5x more jets for training compared to the baseline dataset
- Substantial gain in performance
  - ~70% higher BKG rejection (@ϵ_S = 70%)
- *Question:* Can we encode more physics into the network to make the training more data-efficient?

	Accuracy	AUC	$1/\varepsilon_b$ at $\varepsilon_s = 70\%$	$1/\varepsilon_b$ at $\varepsilon_s = 50\%$
ParticleNet	0.980	0.9979	1342 ± 4	6173 ± 425
ParticleNeXt	0.981	0.9982	2008 ± 75	8621 ± 309
ParticleNeXt (extended dataset)	**0.983**	**0.9986**	**3378**	**15873**

PERFORMANCE: HIGGS TAGGING



- Baseline dataset:
  - training/validation/test splitting:
    - 1.6M / 0.4M / 2M
- Extended dataset:
  - training/validation/test splitting:
    - 10M / 1M / 2M
- Consistent improvement for ParticleNeXt in Higgs tagging as well
  - ~30% higher BKG rejection than ParticleNet
  - another 30% when trained on the extended dataset

	Accuracy	AUC	$1/\varepsilon_b$ at $\varepsilon_s = 70\%$	$1/\varepsilon_b$ at $\varepsilon_s = 50\%$
ParticleNet	0.983	0.9983	1562 ± 24	5128 ± 237
ParticleNeXt	0.985	0.9986	2045 ± 29	7143 ± 349
ParticleNeXt (extended dataset)	**0.986**	**0.9989**	**2770**	**13699**

PERFORMANCE ON PUBLIC BENCHMARKS

Top tagging landscape

	AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.3$) single	mean	median	#Param
CNN [16]	0.981	0.930	914±14	995±15	975±18	610k
ResNeXt [30]	0.984	0.936	1122±47	1270±28	1286±31	1.46M
TopoDNN [18]	0.972	0.916	295±5	382± 5	378 ± 8	59k
Multi-body N-subjettiness 6 [24]	0.979	0.922	792±18	798±12	808±13	57k
Multi-body N-subjettiness 8 [24]	0.981	0.929	867±15	918±20	926±18	58k
TreeNiN [43]	0.982	0.933	1025±11	1202±23	1188±24	34k
P-CNN	0.980	0.930	732±24	845±13	834±14	348k
ParticleNet [47]	0.985	0.938	1298±46	1412±45	1393±41	498k
LBN [19]	0.981	0.931	836±17	859±67	966±20	705k
LoLa [22]	0.980	0.929	722±17	768±11	765±11	127k
Energy Flow Polynomials [21]	0.980	0.932	384			1k
Energy Flow Network [23]	0.979	0.927	633±31	729±13	726±11	82k
Particle Flow Network [23]	0.982	0.932	891±18	1063±21	1052±29	82k
GoaT	0.985	0.939	1368±140		1549±208	35k
ParticleNet-Lite	**0.984**	**0.937**	**1262±49**			**26k**
ParticleNet	**0.986**	**0.940**	**1615±93**			**366k**
ParticleNeXt	**0.987**	**0.942**	**1923±48**			**560k**

G. Kasieczka et al. [1902.09914]

Quark/gluon tagging

	Acc	AUC	$1/\epsilon_B$ ($\epsilon_S = 0.5$)	$1/\epsilon_B$ ($\epsilon_S = 0.3$)
ResNeXt-50 [16]	0.821	0.9060	30.9	80.8
P-CNN [16]	0.827	0.9002	34.7	91.0
PFN [32]	-	0.9005	34.7±0.4	-
ParticleNet-Lite [16]	0.835	0.9079	37.1	94.5
ParticleNet [16]	0.840	0.9116	39.8±0.2	98.6±1.3
ABCNet [17]	0.840	0.9126	42.6±0.4	**118.4±1.5**
SPCT	0.824	0.899	34.4±0.4	100.3±1.5
PCT	**0.841**	**0.9140**	**43.3±0.7**	117.5±1.4
ParticleNeXt	0.841	0.9129	41±0.1	105±1.0

V. Mikuni, F. Canelli [2102.05073]

Summary & Outlook

- ParticleNeXt
  - new GNN architecture for jet tagging
  - enhanced expressiveness w/ several new features in the network design
  - significant performance improvement as demonstrated in the top and Higgs tagging benchmarks
  - paper, code and dataset to come soon — stay tuned!
- Still, performance can be further improved via:
  - model ensembling
  - extending training dataset
- Models that better incorporate physically-motivated inductive biases are likely to bring better data-efficiency, and to improve the performance as well

Backups

Input Features

Variable	Definition
$\Delta\eta$	difference in pseudorapidity between the particle and the jet axis
$\Delta\phi$	difference in azimuthal angle between the particle and the jet axis
$\log p_T$	logarithm of the particle's p_T
$\log E$	logarithm of the particle's energy
$\log \frac{p_T}{p_T(\text{jet})}$	logarithm of the particle's p_T relative to the jet p_T
$\log \frac{E}{E(\text{jet})}$	logarithm of the particle's energy relative to the jet energy
ΔR	angular separation between the particle and the jet axis ($\sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$)
q	electric charge of the particle
`isElectron`	if the particle is an electron
`isMuon`	if the particle is a muon
`isChargedHadron`	if the particle is a charged hadron
`isNeutralHadron`	if the particle is a neutral hadron
`isPhoton`	if the particle is a photon
$\tanh d_0$	hyperbolic tangent of the transverse impact parameter of the track (in units of mm)
$\tanh d_z$	hyperbolic tangent of the longitudinal impact parameter of the track (in units of mm)
σ_{d_0}	error of the transverse impact parameter
σ_{d_z}	error of the longitudinal impact parameter