Modeling a top jet classifier with two-point energy correlation and geometry of soft emission

Sung Hak Lim

Theory Center, KEK

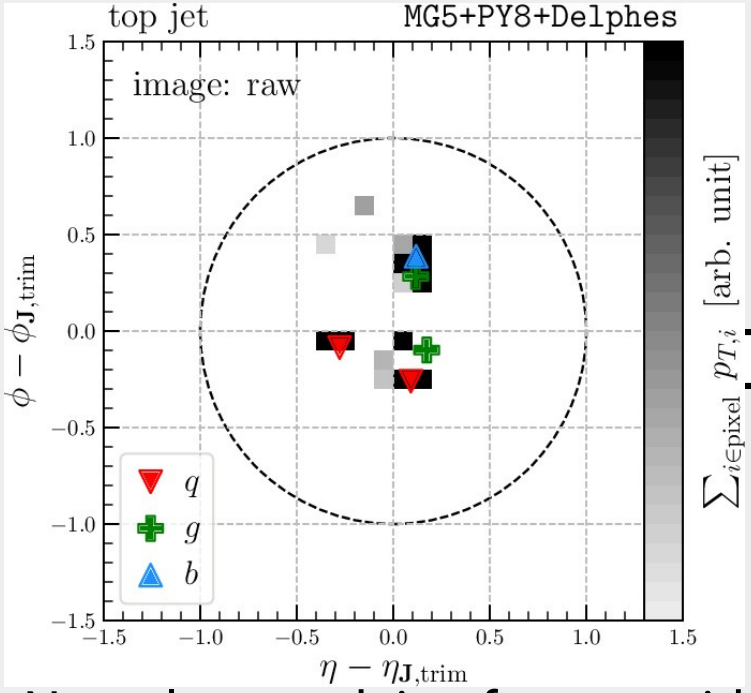
ML4Jets2020

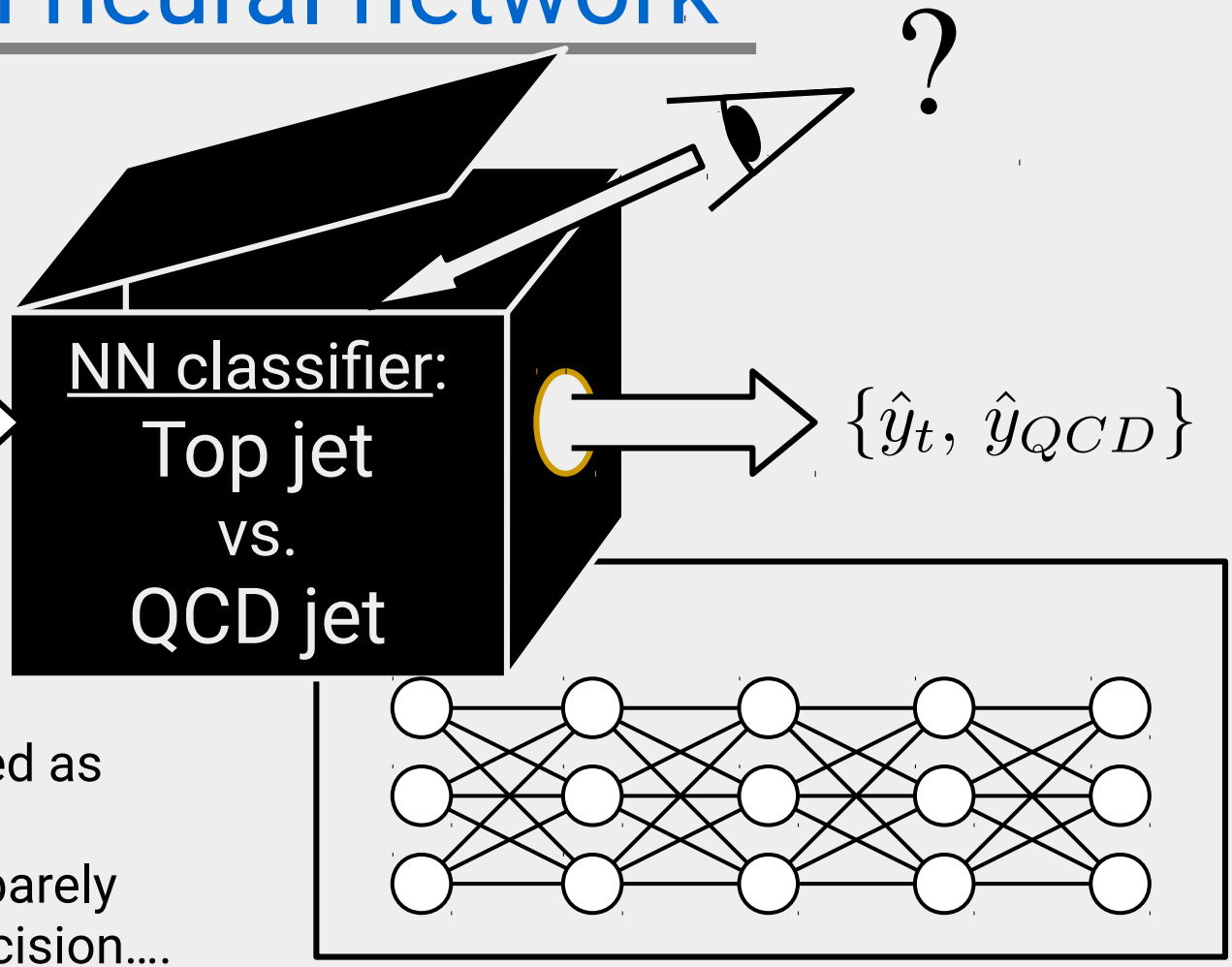
NYU, New York, USA

Jan. 2020

S. H. Lim, M. M. Nojiri, arXiv:1807.03312, JHEP10(2018)181.
A. Chakraborty, **S. H. Lim**, M. M. Nojiri, arXiv:1904.02092, JHEP07(2019)135.
A. Chakraborty, **S. H. Lim**, M. M. Nojiri, M. Takeuchi, will appear in arXiv soon

Difficulties on understanding the results from neural network





Neural network is often considered as a black box because studying its internal information barely gives you an insight about the decision....

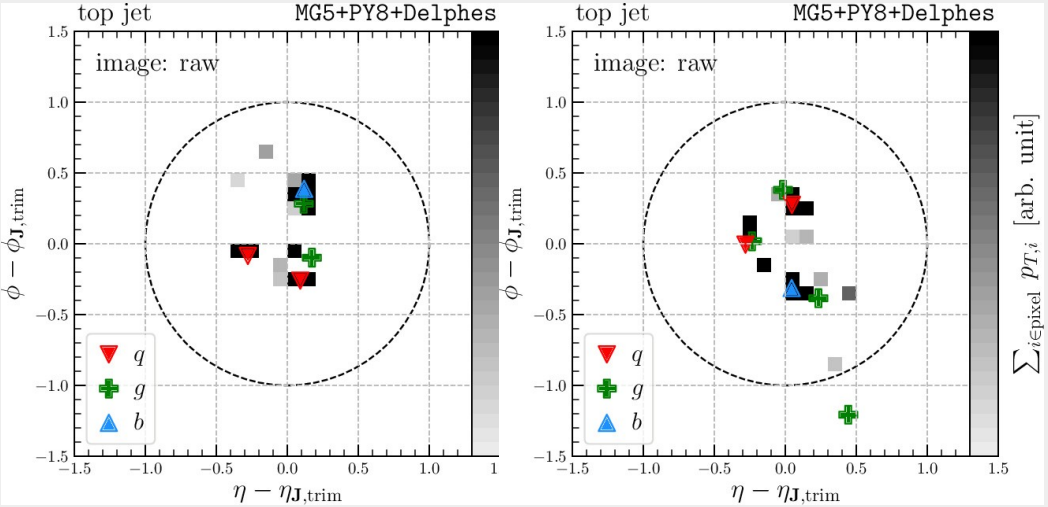
We introduce a relatively transparent framework:

Relation Network with **Two-Point Energy Correlation**
and
Minkowski Functional for **Geometry of Soft Emission**

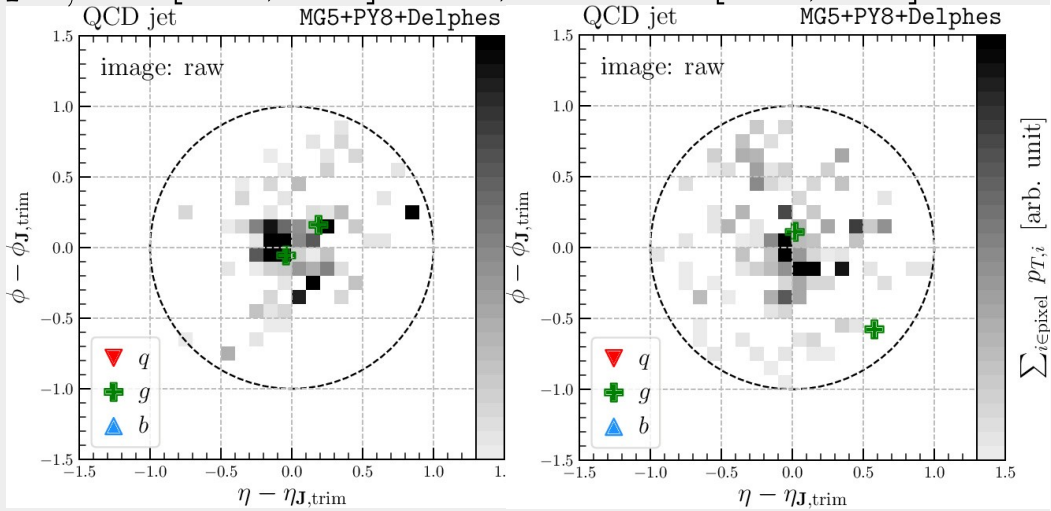
Top jets vs. QCD jets

Can we understand the reasoning of NN classifier outputs?

$p_{T,\mathbf{J}} \in [500, 600]\,\mathrm{GeV}, \quad m_{\mathbf{J}} \in [150, 200]\,\mathrm{GeV}$



$pp \to t\bar{t}$ **Top jets**



QCD jets (figures are gluon jets) $pp \to jj$

We observe two key features for the classification at a glance:

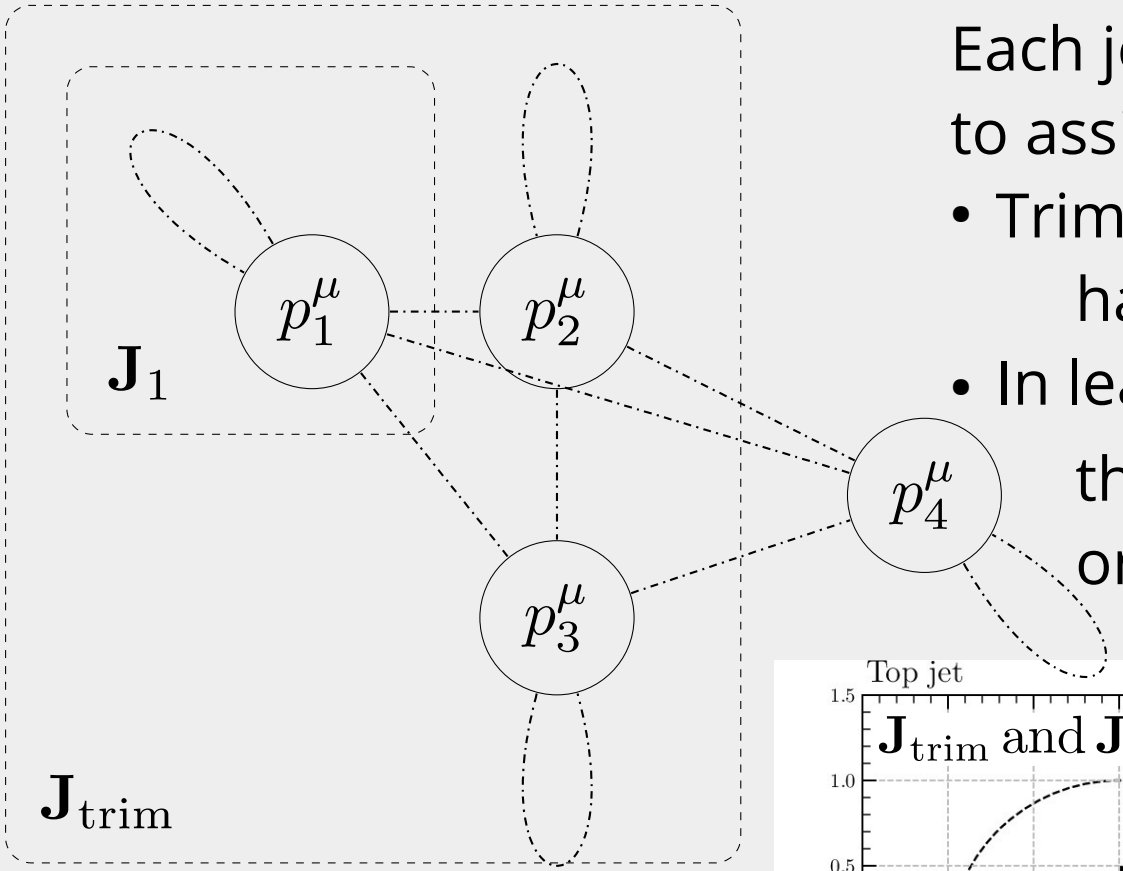
- Correlation among subclusters
- Distribution of soft emission

Analyze them by:

- Relation Network with Two-Point Energy Correlation
- Geometric Data Analysis: Minkowski Functional

See also
ParticleNet
arXiv:1902.08570

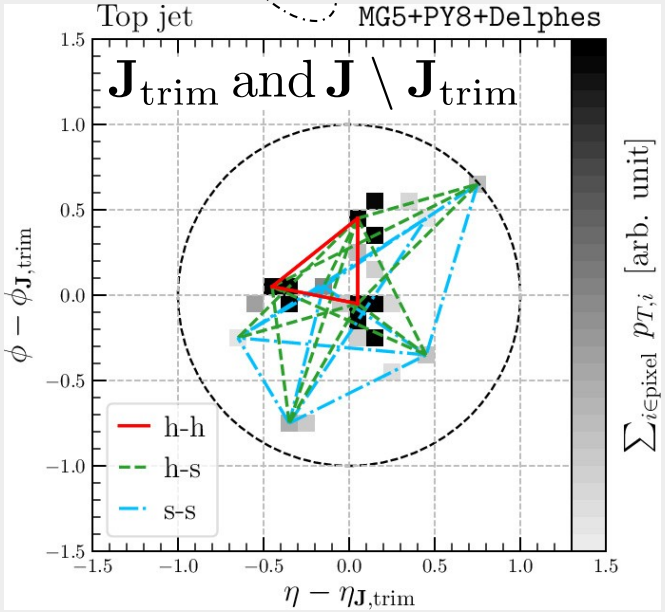
Jet as a Graph

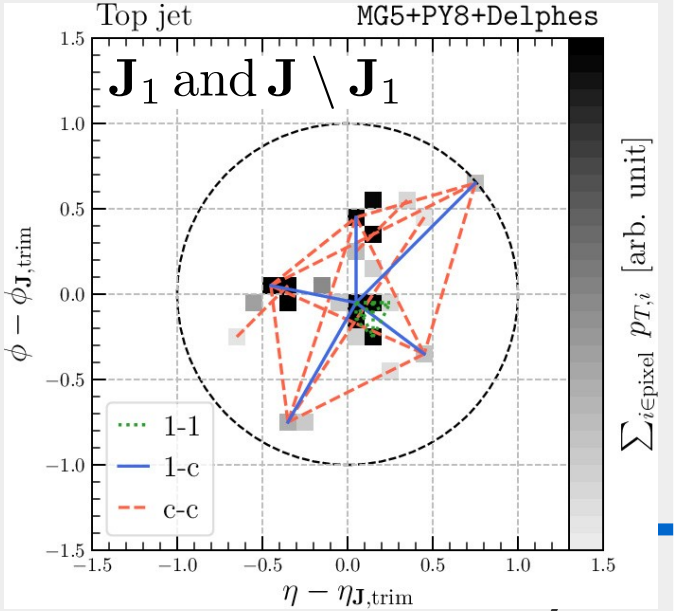


Each jet constituents has two labels to assist the substructure identification:

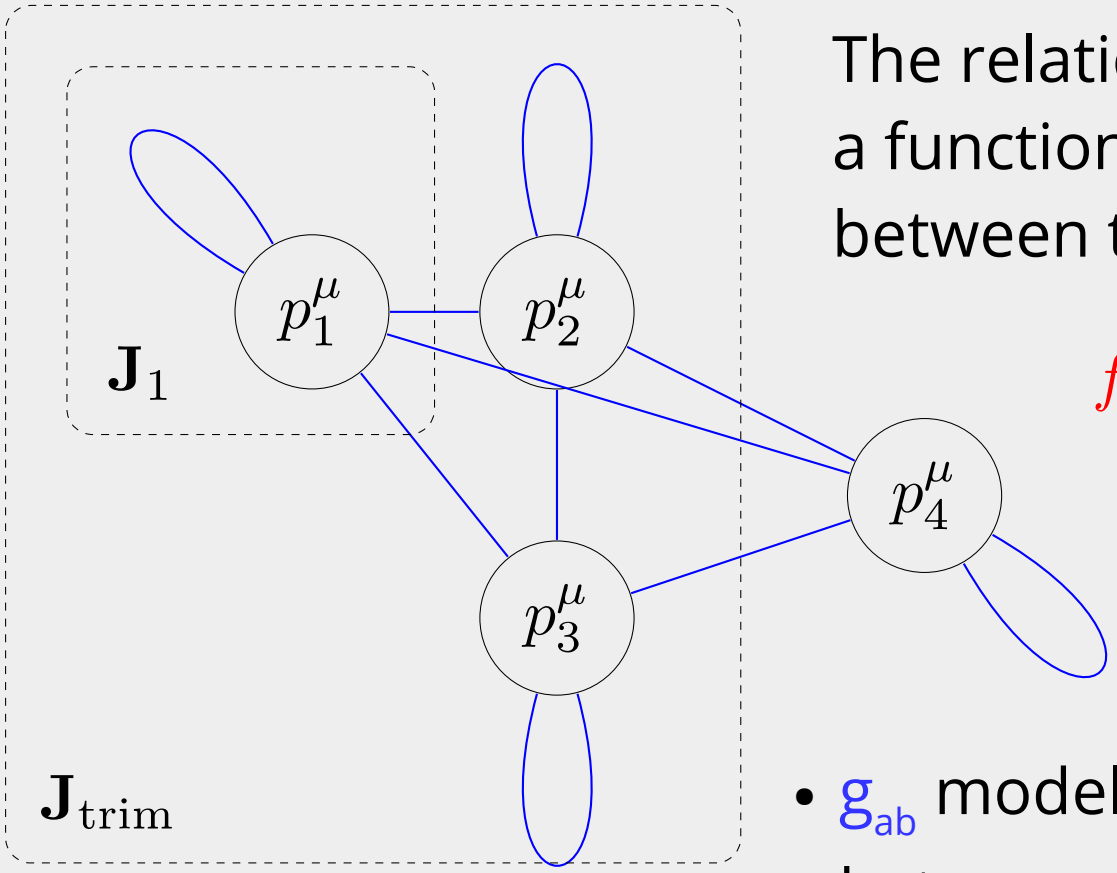
- Trimmed or not:
  hard and soft substructure
- In leading p_T subjet or not:
  three-prong jet is factored into one-prong jet and two-prong jet.

We use a vertex-labeled, fully-connected graph embedding of a jet.





Jet as a Graph, and Relation Network



The relation network (RN) is
a functional model that learns correlation between two points.

$$f\left(\sum_{i\in \mathrm{J}_a, j\in \mathrm{J}_b} g_{ab}(p_i^\mu, p_j^\mu)\right)$$

arXiv:1702.05068, 1706.01427
See also Yang-Ting's 2PCNN!
arXiv:1911.02020

- g_{ab} models two-point correlation between particles.
- Summation aggregates the correlations.
- f is the output model of the aggregated correlation.

Relation Network and Two-Point Energy Correlation Spectra

If we add IRC safety constraints on the RN, the model transforms into an analysis of the two-point energy correlation.

$$f\left(\sum_{i\in \mathrm{J}_a, j\in \mathrm{J}_b} g_{ab}(p_i^\mu, p_j^\mu)\right)$$

Relation Network

arXiv:1702.05068, 1706.01427

IRC safety

$$f\left(\sum_{i\in \mathrm{J}_a, j\in \mathrm{J}_b} p_{T,i} p_{T,j} w_{ab}(\vec{R}_i, \vec{R}_j)\right)$$

Two-Point Energy Correlator

Two-Point Energy Flow Polynomials

arXiv:hep-ph/9601308, 1712.07124

Use relative distance only

$$f\left(\int dR\, S_{2,ab}(R) w_{ab}(R)\right)$$

Two-Point Energy Correlation Spectra

arXiv:1807.03312, 1904.02092

More constraints...

$$\sum_{i\in \mathrm{J}_a, j\in \mathrm{J}_b} p_{T,i} p_{T,j} w_{ab}(R_{ij}) = \int dR\, S_{2,ab}(R) w_{ab}(R)$$

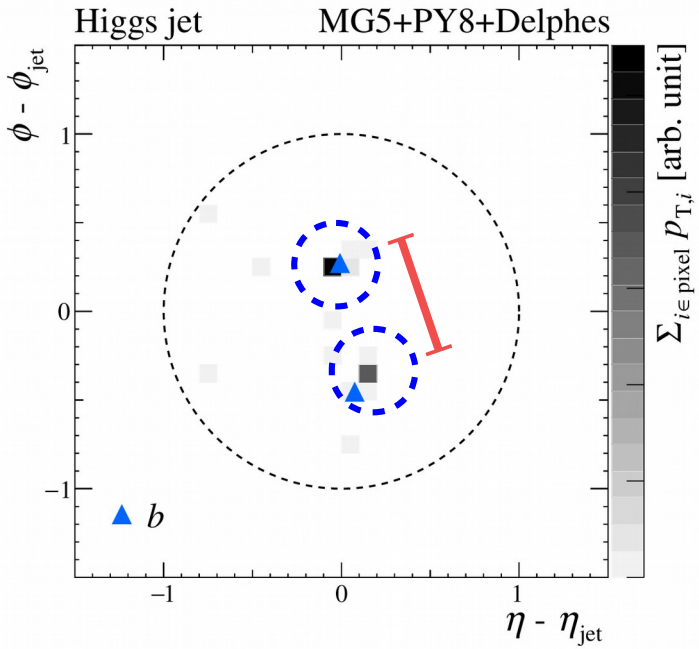
Two-point energy correlation spectrum

See also
Energy-Energy Correlation
Basham, et. al.
Phys. Rev. Lett. 41, 1585

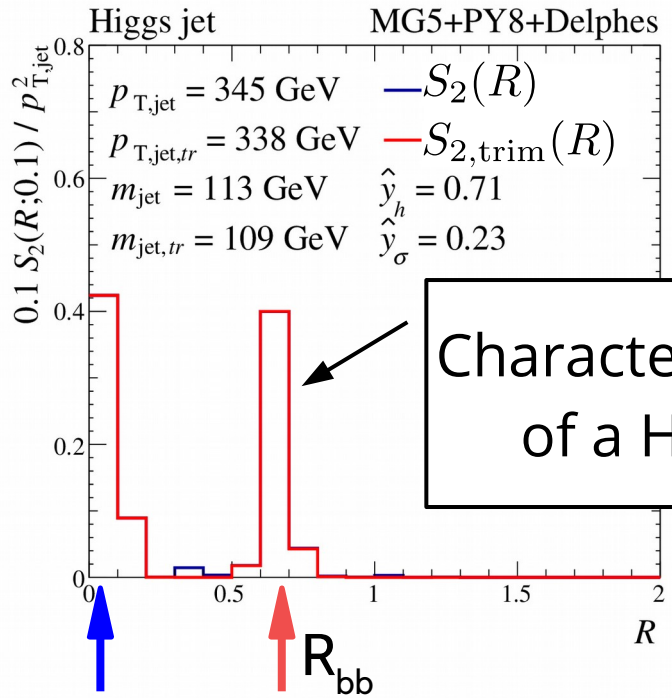
Two-point energy correlation spectrum is an aggregated energy correlation between two constituents at a distance R.

$$S_{2,ab}(R) = \int d\vec{R}_1 \, d\vec{R}_2 \; P_{T,a}(\vec{R}_1) P_{T,b}(\vec{R}_2) \delta(R - R_{12})$$

$$P_{T,a}(\vec{R}) = \sum_{i \in \mathbf{J}_a} p_{T,i} \, \delta(\vec{R} - \vec{R}_i)$$



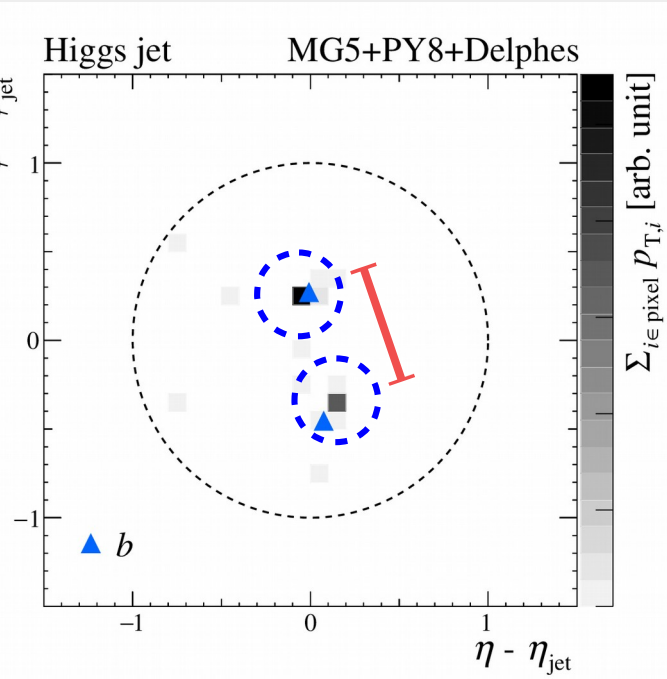
Calorimeter jet, anti-kt, R=1.0

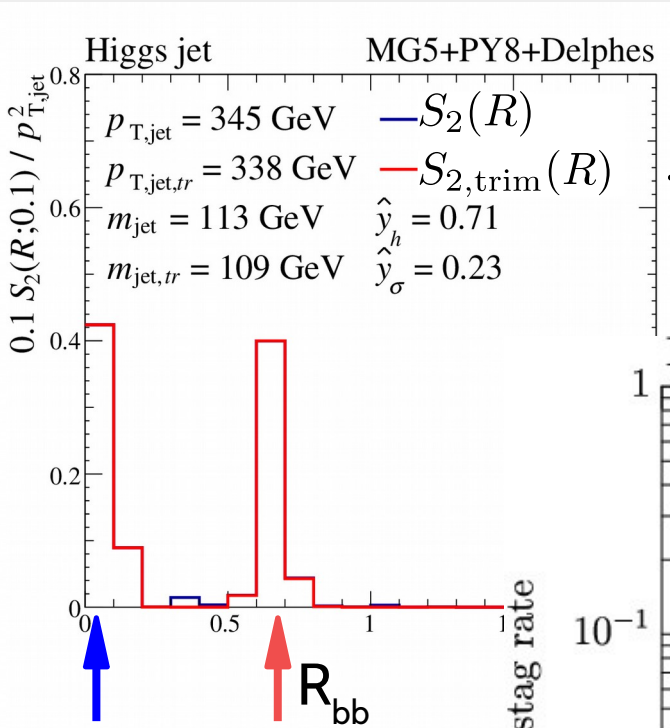


Characterstic peak of a Higgs jet

Higgs jets vs. QCD jets

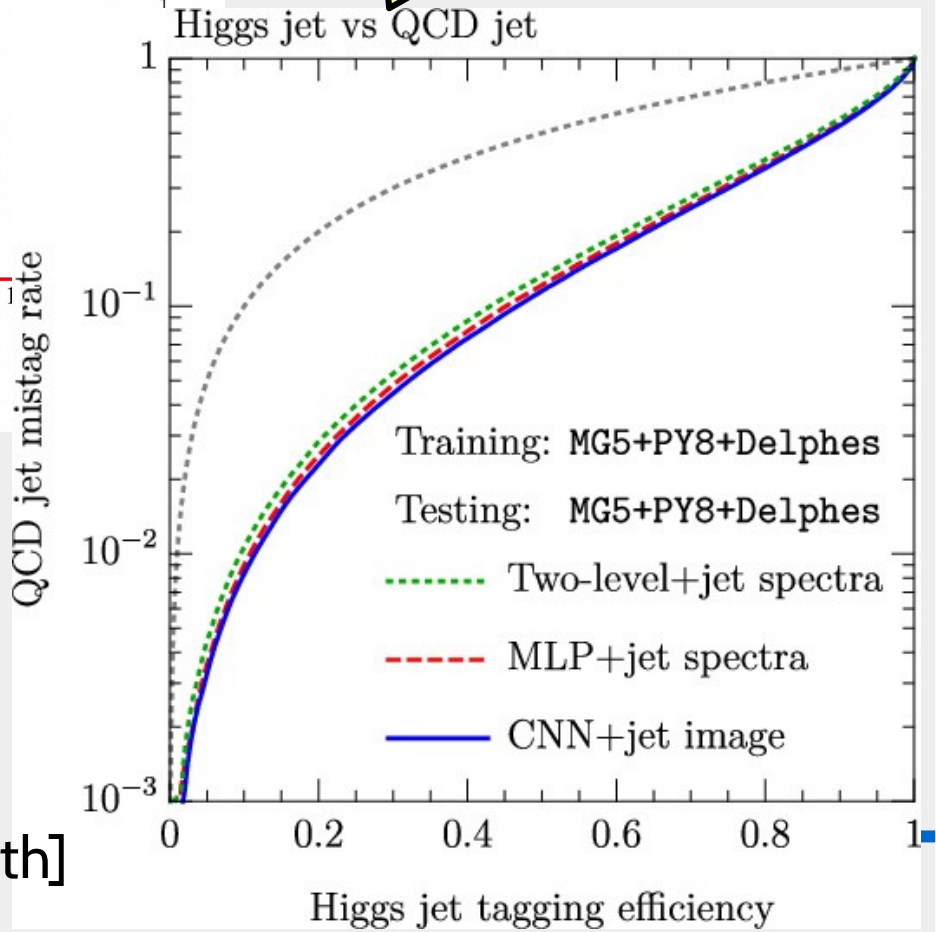
arXiv:1904.02092





Calorimeter jet, anti-kt, R=1.0

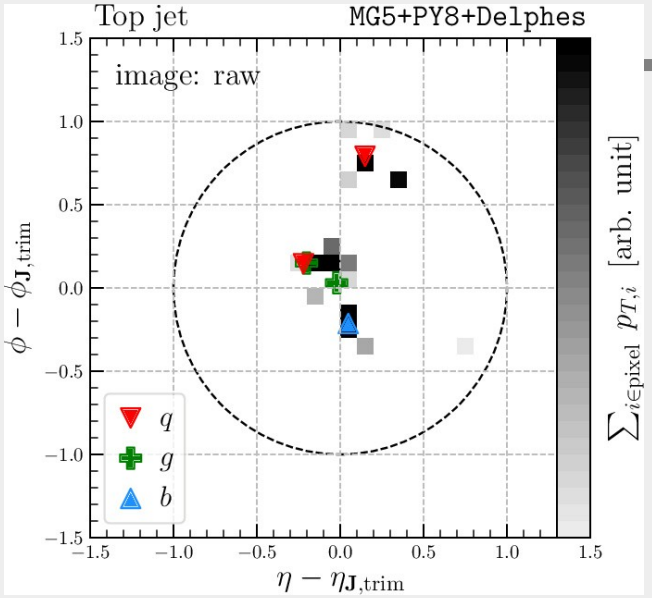
Good performance

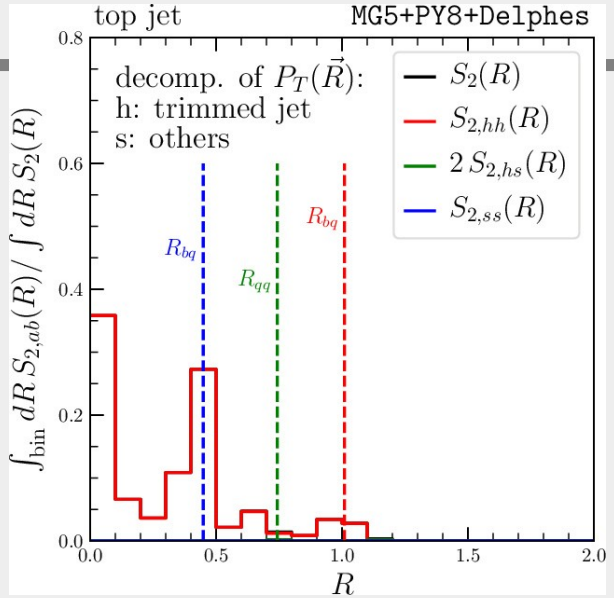


The dimension of inputs is also reduced compared to the jet image so that the network is computationally cheap.

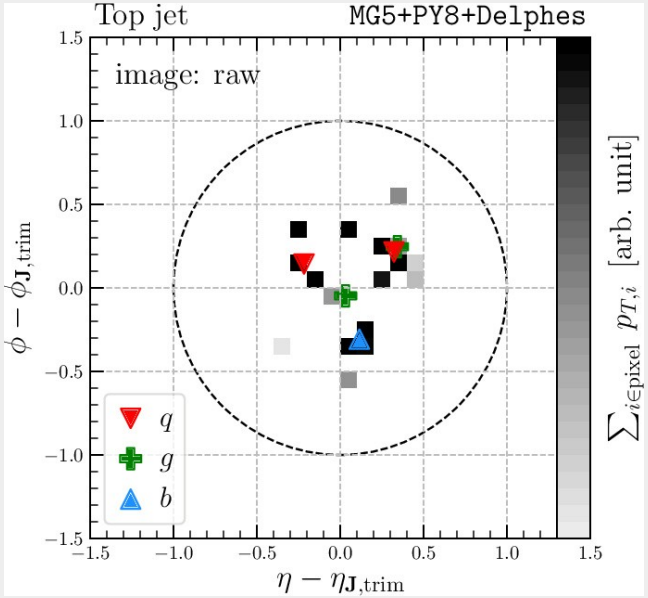
[Length/bin width]^2 → [Length/bin width]

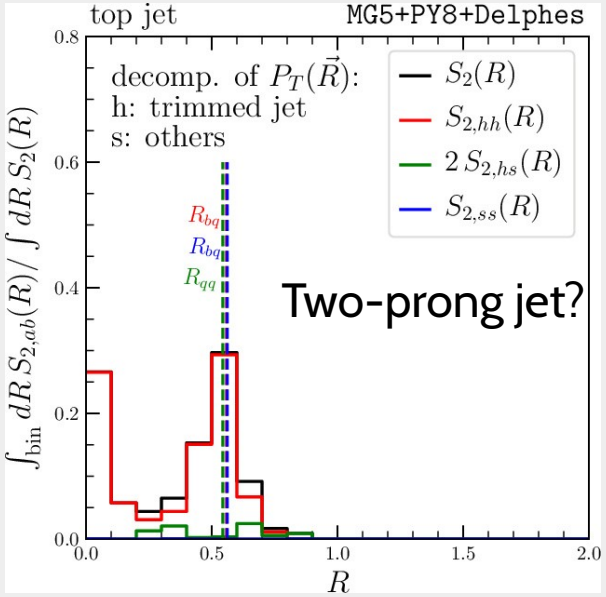
Top Jets





Two-point energy correlation spectrum of trimmed three-prong jet has three characteristic peaks.





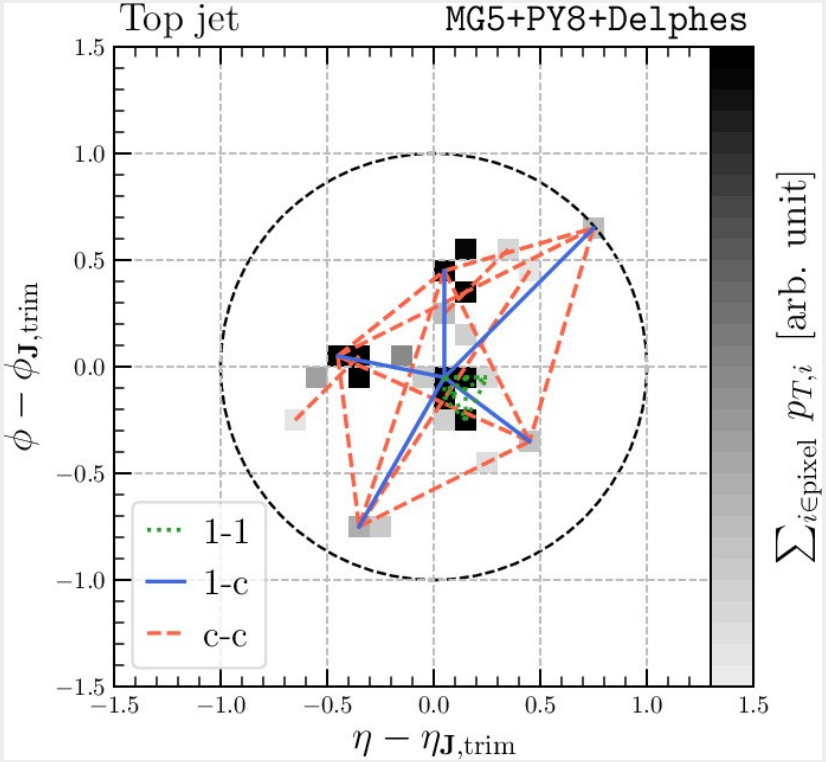
Need more information to resolve overlapping peaks...

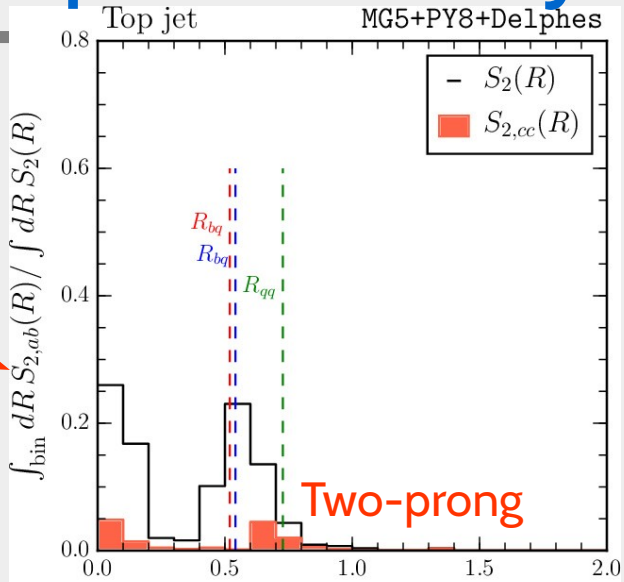
Calorimeter jet, anti-kt, R=1.0

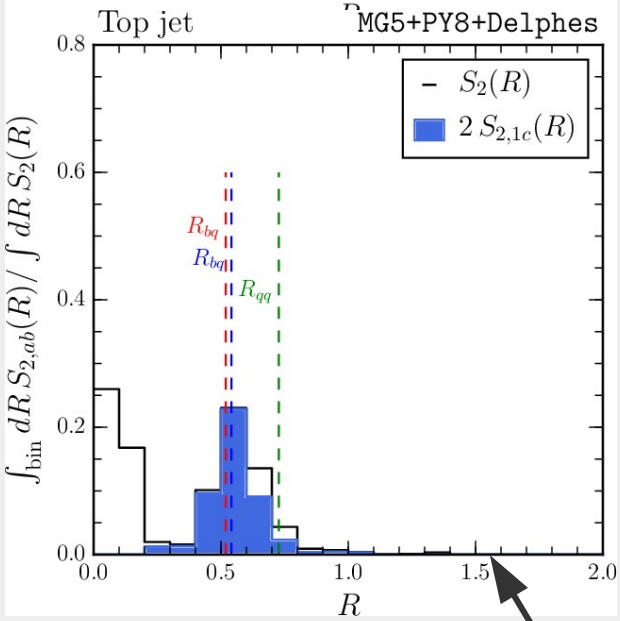
Two-Point Correlation Spectrum: Leading p_T subjet and its complementary

With leading p_T subjet labeling, the three-prong substructure identification is decomposed to

- Two-prong substructure in the complementary set,
- Correlation between the leading p_T subjet and the rest.







See also **Telescoping Jet Substructure**
arXiv: 1711.11041

A top tagger architecture with two-point energy correlation spectra

Hard and soft substructure analyzer

$S_{2,\mathrm{trim}}(R)$

$S_{2,\mathrm{soft}}(R)$

Leading subjet and its complmentary substructure analyzer

$S_{2,11}(R)$ $S_{2,cc}(R)$

$S_{2,1c}(R)$

kinematics

$p_{T,\mathbf{J}}$, $m_{\mathbf{J}}$

$p_{T,\mathbf{J}_{\mathrm{trim}}}$, $m_{\mathbf{J}_{\mathrm{trim}}}$

$p_{T,\mathbf{J}\setminus\mathbf{J}_1}$, $m_{\mathbf{J}\setminus\mathbf{J}_1}$

* some inputs might not be necessary.

MLP

MLP

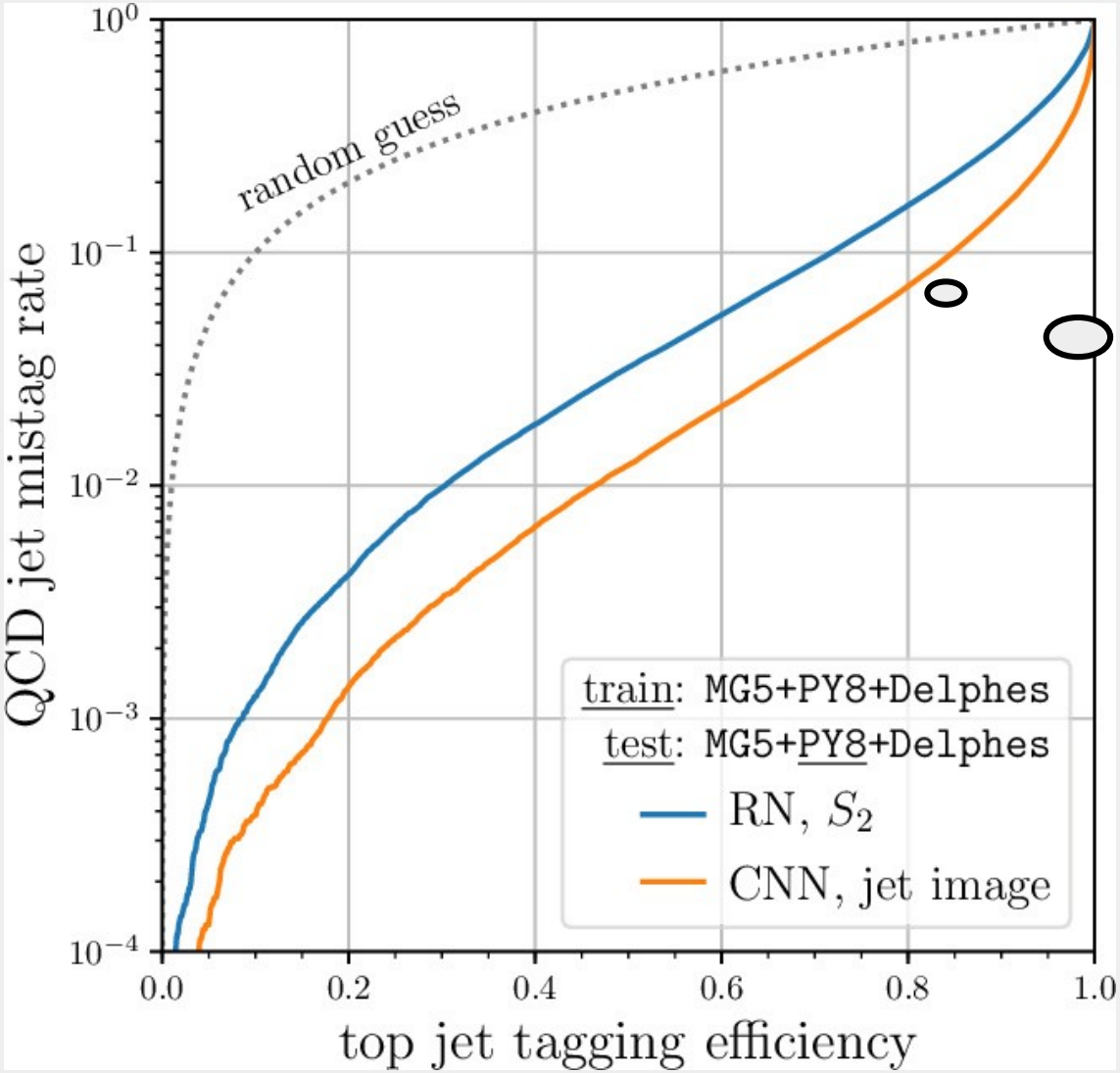
MLP

$\hat{y}(x) = q(\mathrm{top}|x)$

Train the network output with the cross-entropy.

$$\mathcal{L} = \sum_{i=0}^{N_{\mathrm{event}}} w^{(i)}(x|y^{(i)})\, y^{(i)}\, \log \hat{y}^{(i)}(x)$$

ROC and Performance Comparison



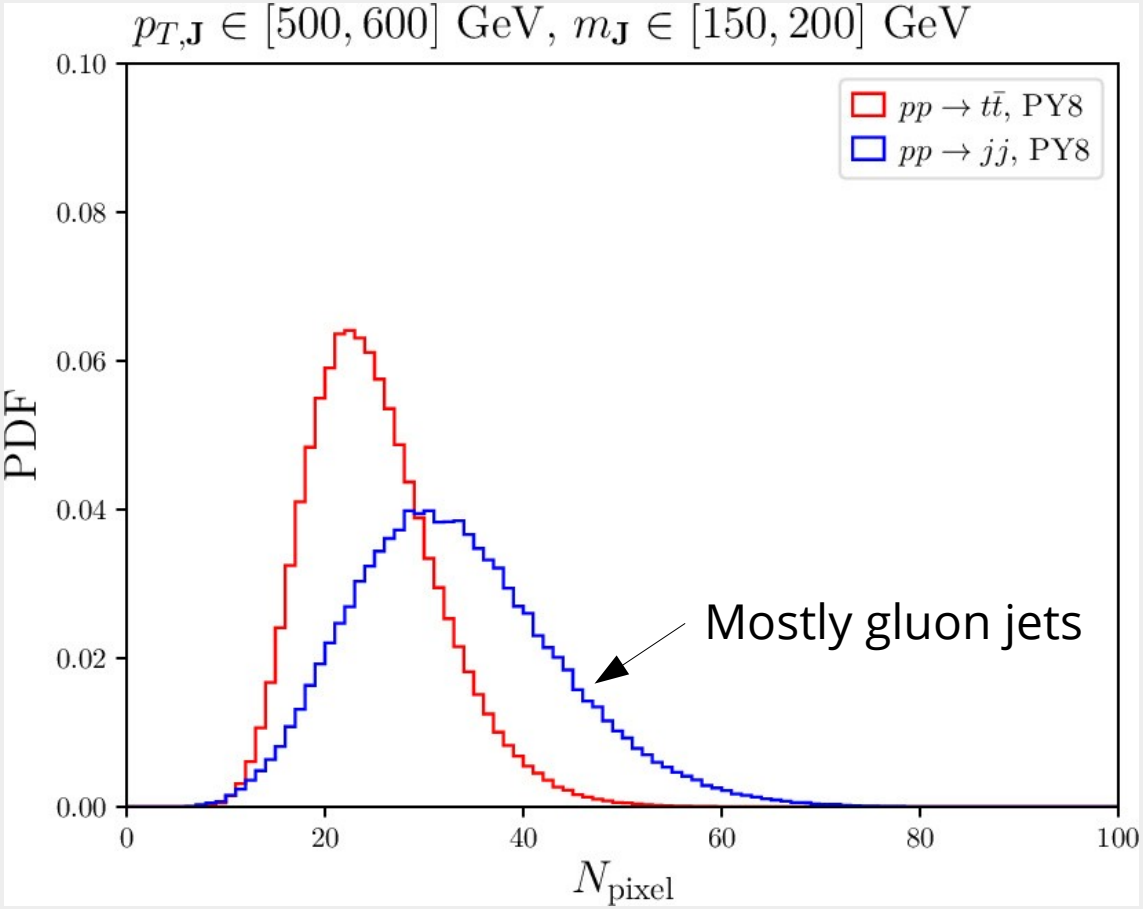
Is CNN better in performance?
- Yes, because we are only using IRC safe two-point energy correlators.

Gap is there.

S2's are not enough for top tagger compared to the Higgs jet tagger. Nevertheless, the gap means that there are information that did not consider. Adding it will fill the gap.

geometry of soft emission

N_{pixel} distribution: top jet and gluon jet



This N_{pixel} is useful for studying color coherence, such as quark jet vs. gluon jet discrimination.

This information is not covered by the energy correlation,
but still important for the jet classification.

Including N_{pixel} as an input of our setup enhance the classification performance, but can we do something more?

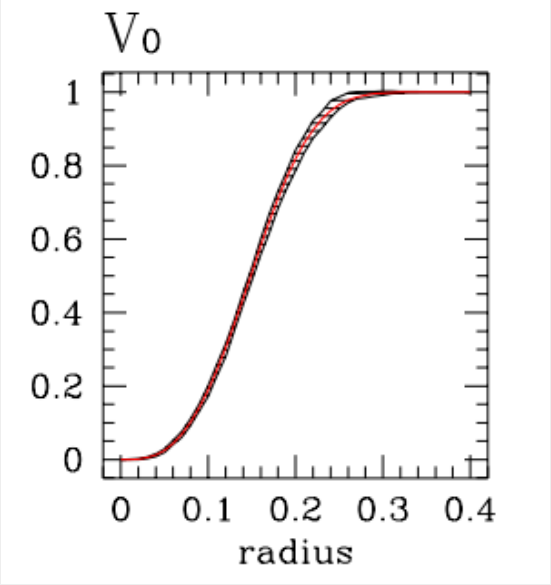
Geometric Data Analysis: Minkowski Functional

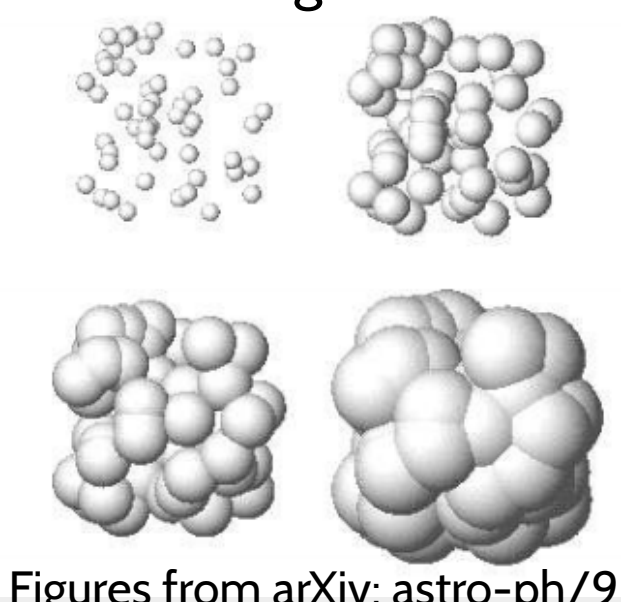
See also applications of
Minkowski Functional
arXiv: astro-ph/9508154
Persistence Homology
arXiv: 1812.06960

We considered a set of geometric measure, called Minkowski functional, which includes volume, surface area, and topological measure, in the case of points on 3D space.

The Minkowski functional encodes the geometry of the points well when we consider the union of balls whose centers are the points.

We measure the Minkowski functional against the radius of ball.





Figures from arXiv: astro-ph/9508154

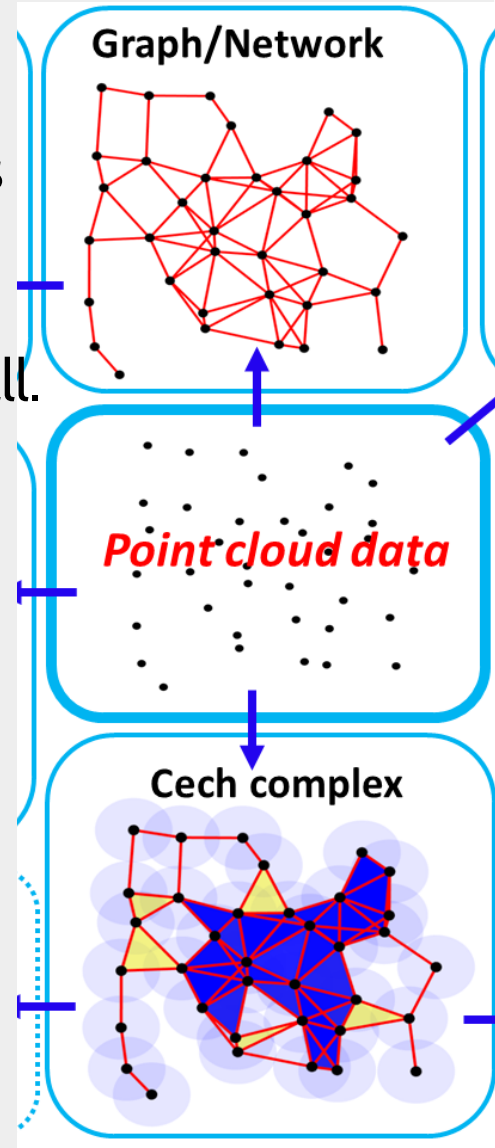


Diagram from arXiv: 1811.00252

In the case of our jet analysis, we only use 2D volume, i.e., **area**, for the simplicity.

Sequence of Pixel Counting

Our detectors have a finite resolution,
so we use approximated areas using jet images.

We calculate the sequence of area as follows:

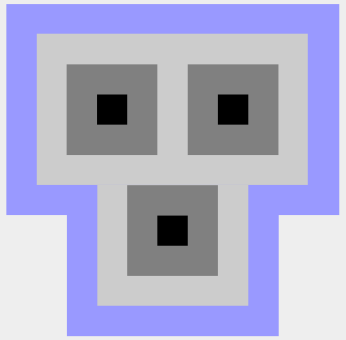
N_0: number of active pixels in the jet image

dN_n: Number of pixels surrounding pixels used in N_{n-1}

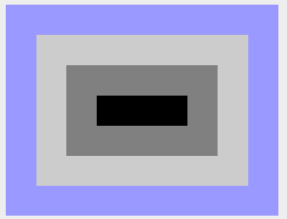
N_n: sum of the numbers of pixels, N_0, ... , dN_n

If all the pixels are isolated, the ratio $N_1 / N_0 = 9$.

If there are connected pixels,
this number is getting smaller.



$N_0 = 3$
$N_1 = 27\ (9N_0)$
$N_1 / N_0 = 9$



$N_0 = 3$
$N_1 = 12\ (=3N_0+6)$
$N_1 / N_0 = 5$

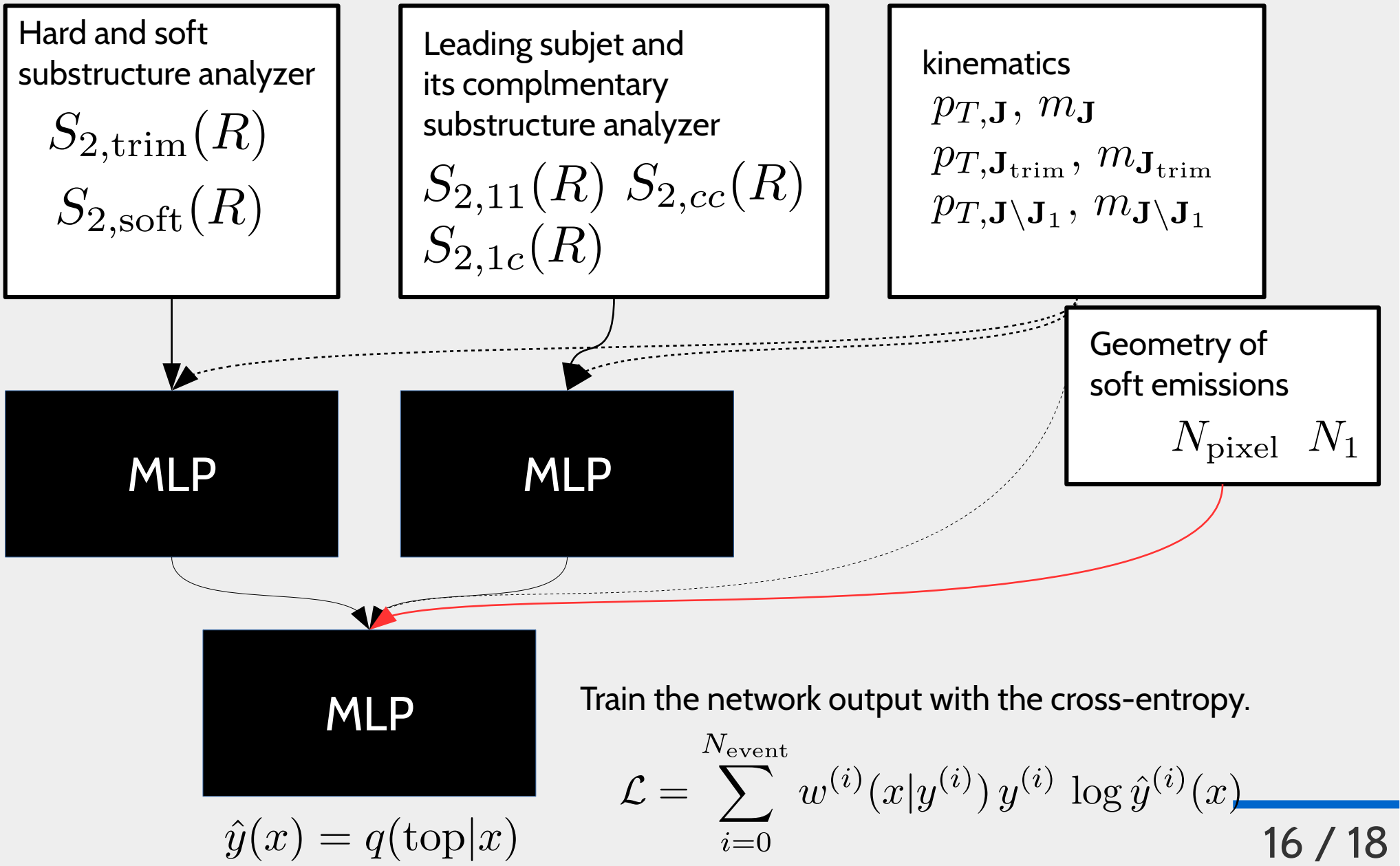
■ N_0 ■ dN_1 ■ dN_2 ■ dN_3

We add the N_0 and N_1 to the RN inputs.

We use two jet images:

- Original jet image
- Jet image with pixels whose pt > 4 GeV

A top tagger architecture with S_2 and Minkowski Functionals

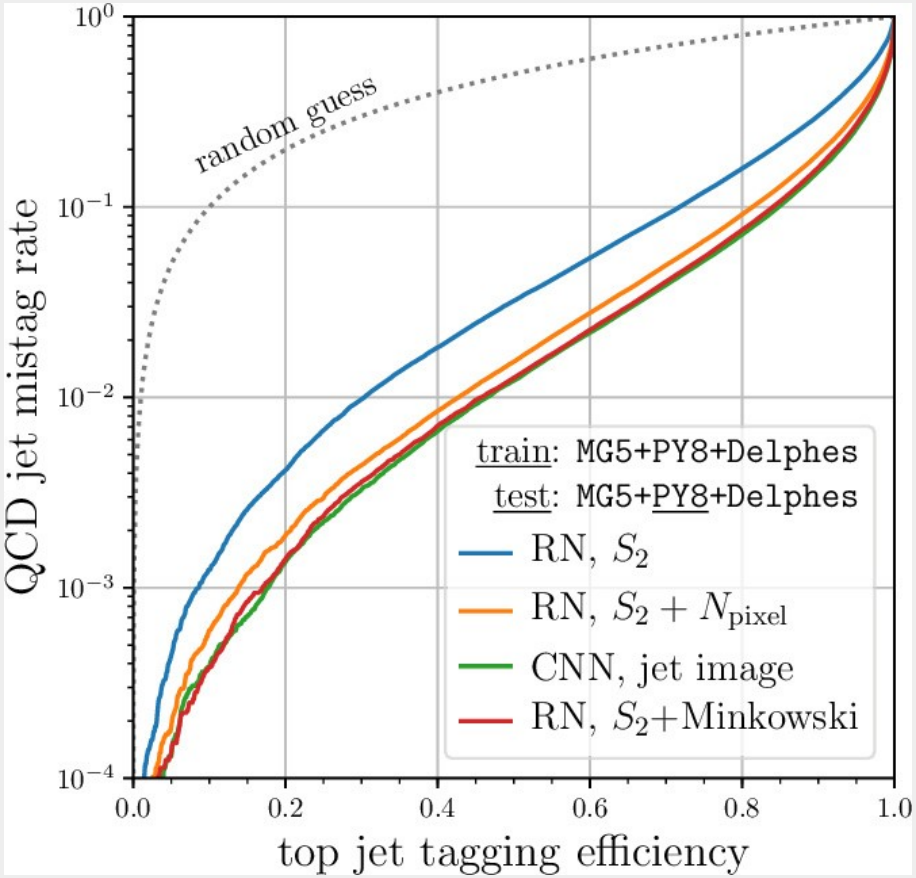


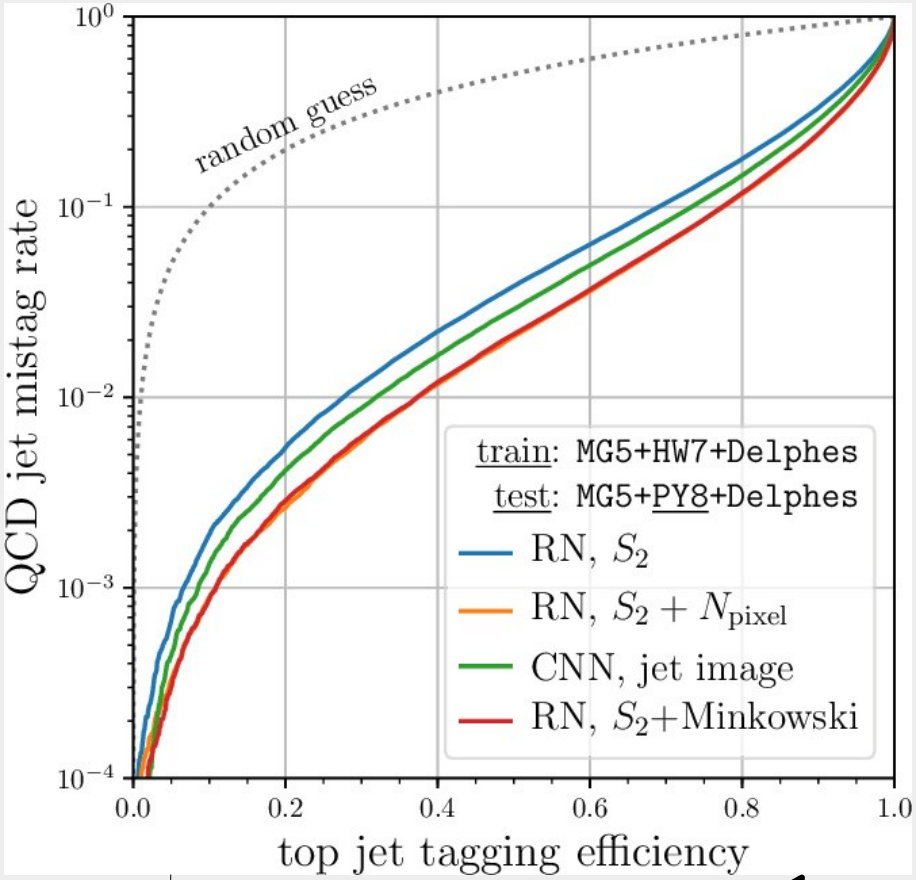
$$\hat{y}(x) = q(\mathrm{top}|x)$$

Train the network output with the cross-entropy.

$$\mathcal{L} = \sum_{i=0}^{N_{\mathrm{event}}} w^{(i)}(x|y^{(i)})\, y^{(i)}\, \log \hat{y}^{(i)}(x)$$

ROCs





The gap between ROC of RN with S2 and ROC of CNN is filled by the Minkowski functional. The CNN may find out the same solution, but it might require more events to get it due to larger functional space coverage of the CNN.

Good performance

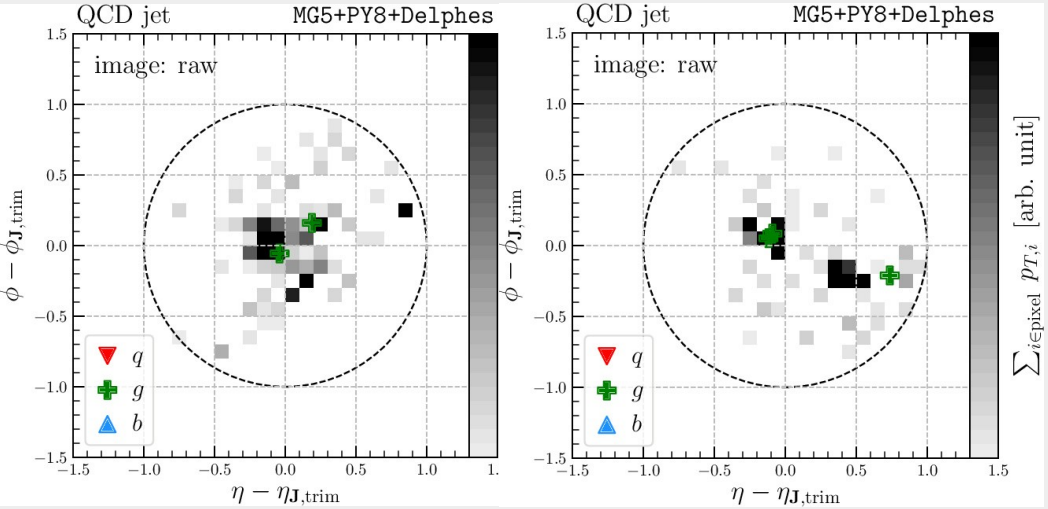
Summary

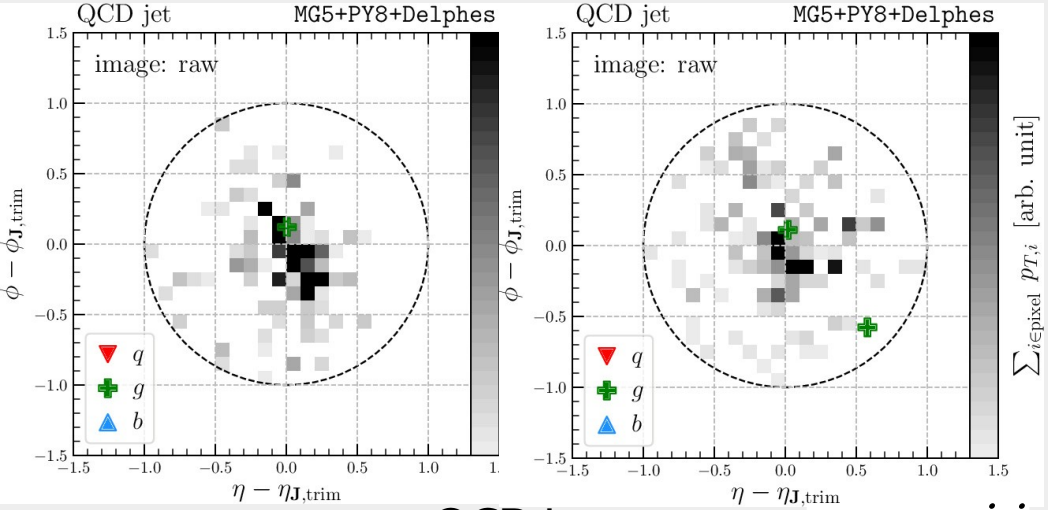
- We introduced a top tagger using **two-point correlation spectra** for analyzing the correlation between subjets and **Minkowski functionals** for analyzing the geometry of soft radiation.
- The number of inputs and computation complexity of this setup is small, but the performance is similar to that of CNN with jet images.
- Analysis with all Minkowski functionals for uncovering hidden geometry of jet, and linear analysis is for the interpretable setup are also ongoing.

Please stay tuned!

Backup

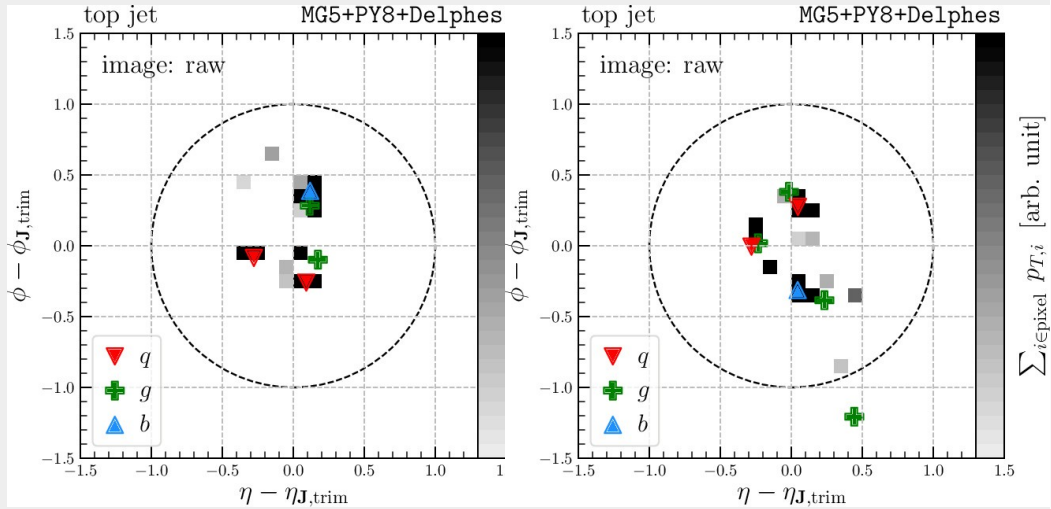
More images

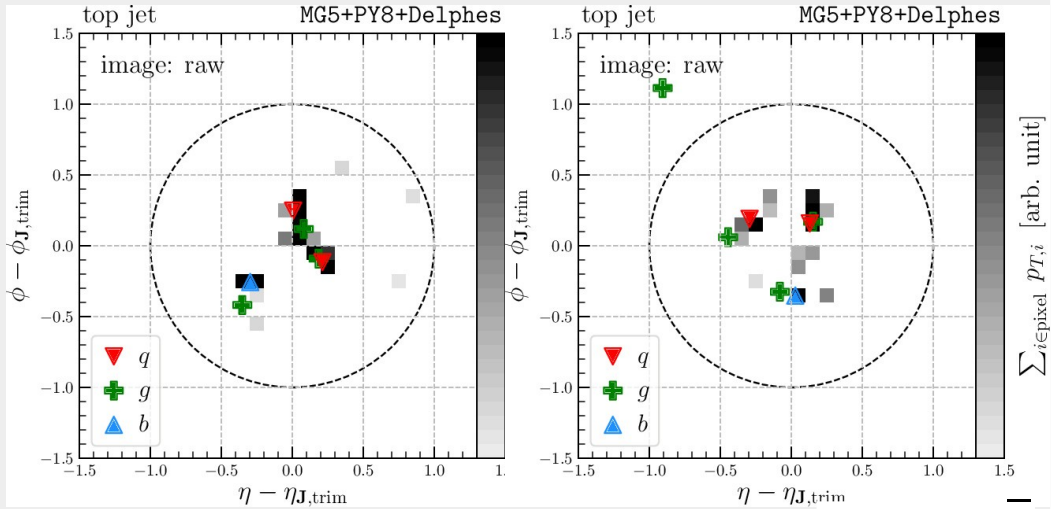




QCD jets
(figures are gluon jets)

$pp \to jj$

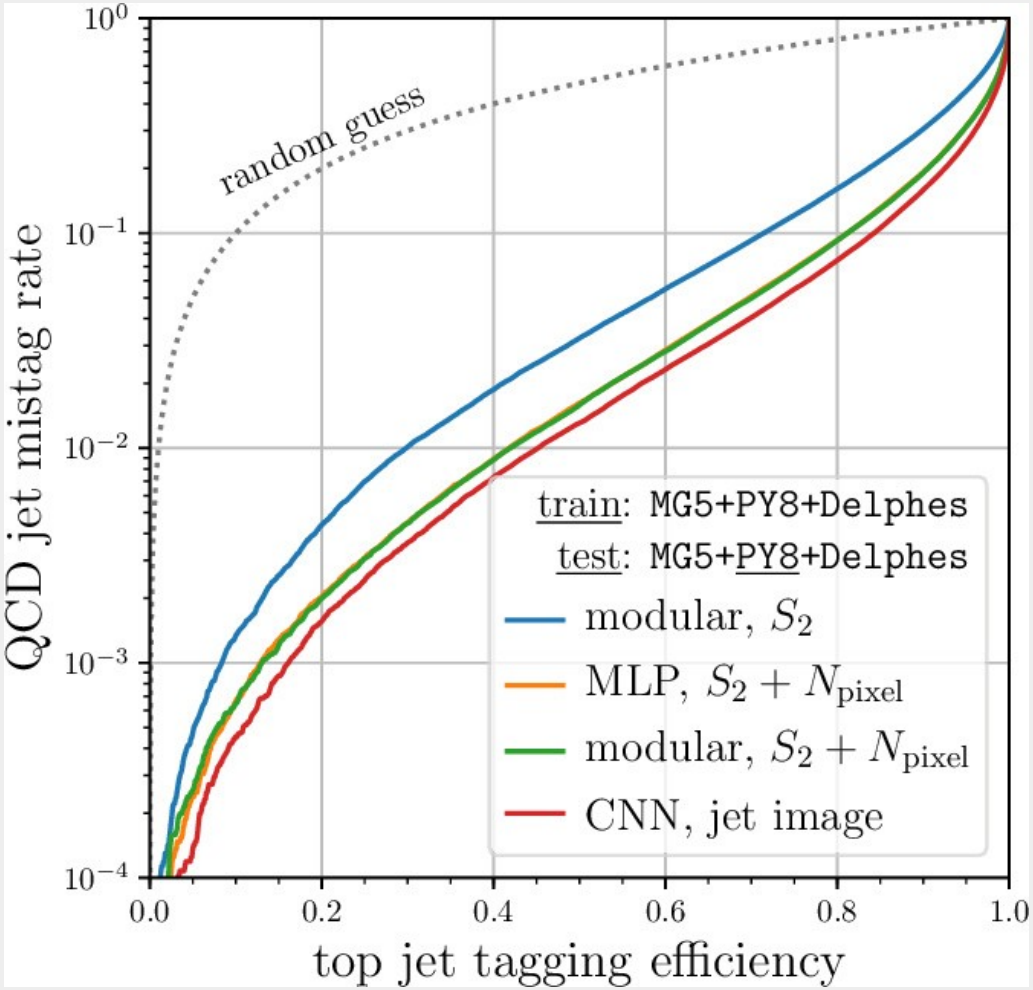




Top jets

$pp \to t\bar{t}$

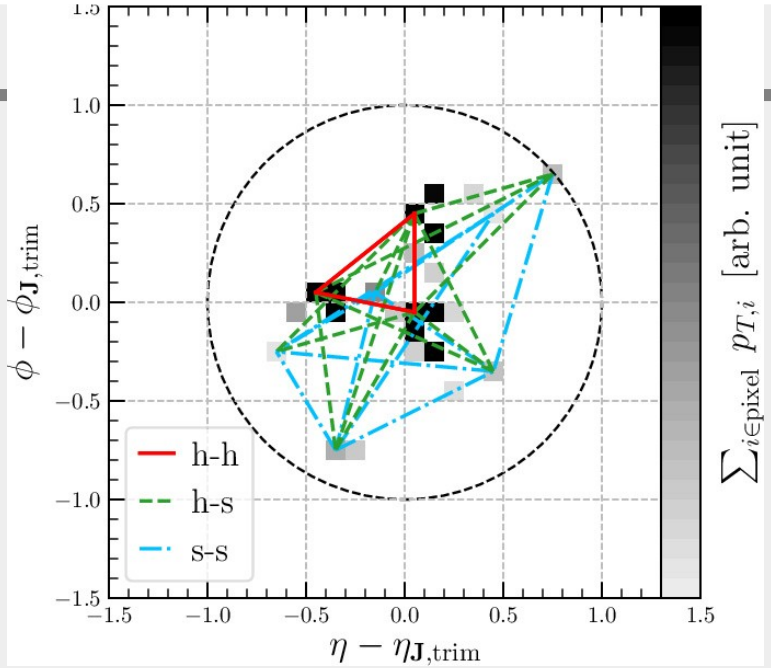
$$p_{T,\mathbf{J}} \in [500, 600]\,\mathrm{GeV}, \quad m_{\mathbf{J}} \in [150, 200]\,\mathrm{GeV}$$

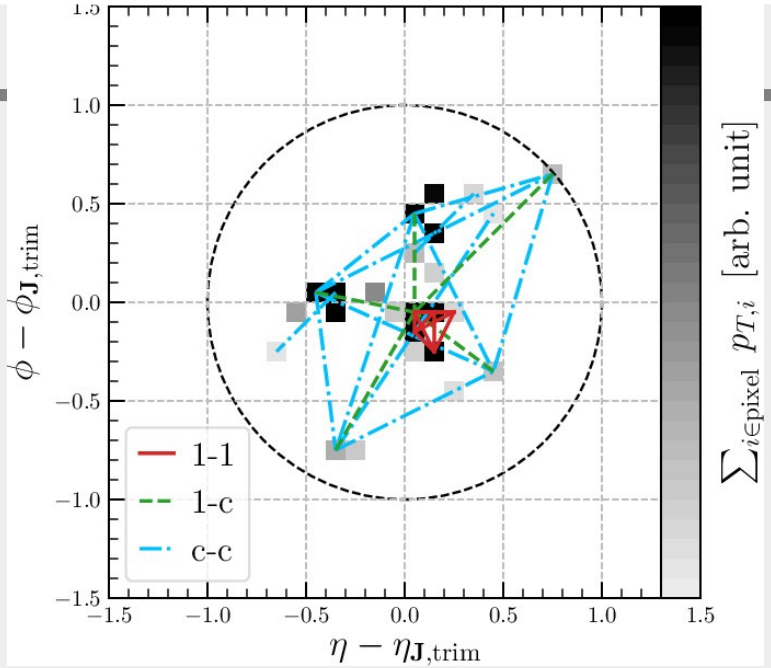

random guess
QCD jet mistag rate
top jet tagging efficiency
train: MG5+PY8+Delphes
test: MG5+PY8+Delphes
modular, S_2
MLP, $S_2 + N_{\text{pixel}}$
modular, $S_2 + N_{\text{pixel}}$
CNN, jet image
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0.0
0.2
0.4
0.6
0.8
1.0

Trimmed constituent label: h
Complementary set label: s

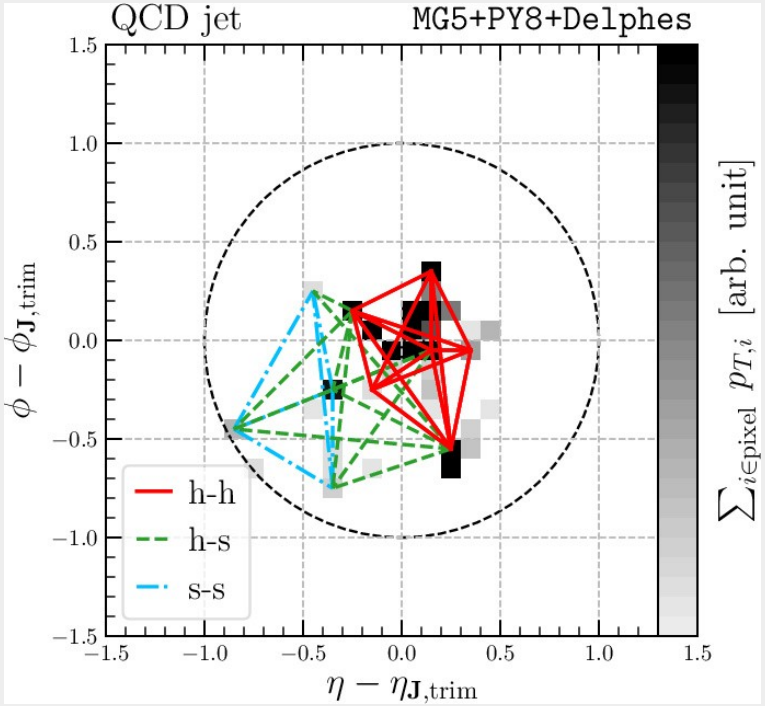
Leading subjet label: 1
Complementary set label: c

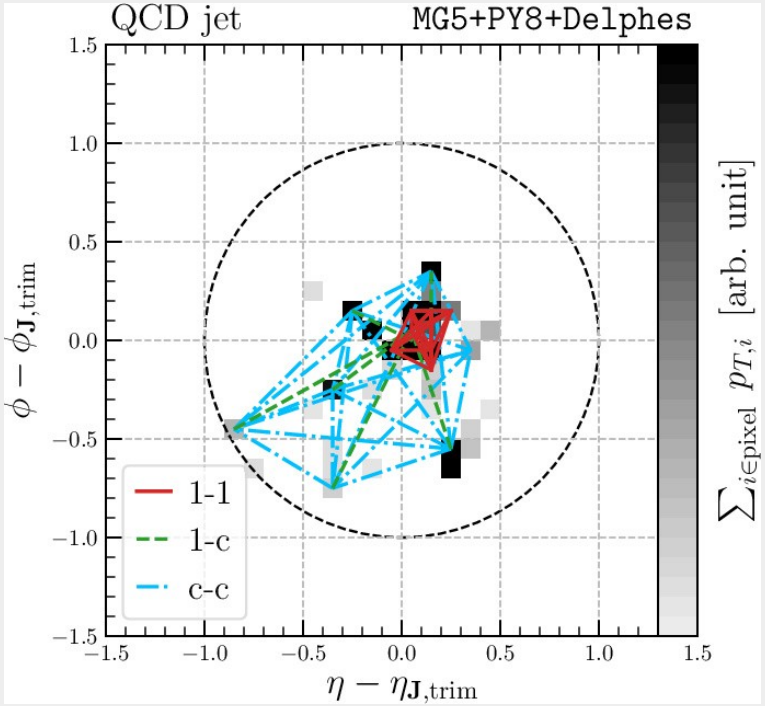
Top jet





QCD jet

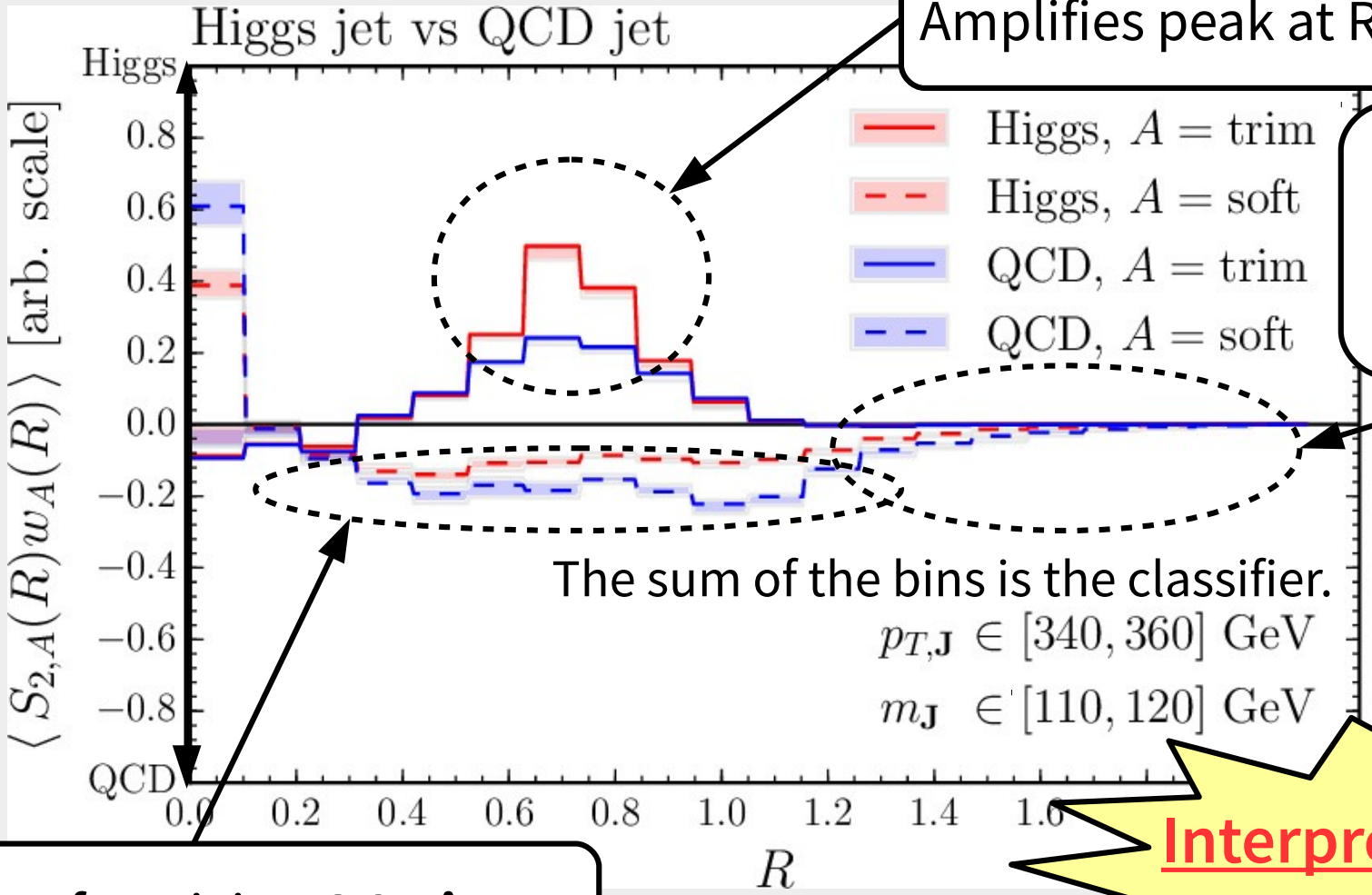




Relational Reasoning of the Classifier Output

arXiv: 1904.02092

$$\Phi[S_{2,ab}] = \int dR\, S_{2,\mathrm{trim}}(R) w^{(2)}_{\mathrm{trim}}(R) + \int dR\, S_{2,\mathrm{soft}}(R) w^{(2)}_{\mathrm{soft}}(R)$$



Amplifies peak at Rbb: **Higgs jet**

Large angle radiation: **QCD jet**

More soft activity: **QCD jet**

Interpretable

Training setup

- The model is implemented with Keras with backend tensorflow.
- Optimizer: ADAM, mimimize the weighted cross-entropy.

$$\mathcal{L} = \sum_{i=0}^{N_{\text{event}}} w^{(i)}(x|y^{(i)})\, y^{(i)}\, \log \hat{y}^{(i)}(x)$$

- $p_{T,J}$ distribution is reweighted to be flat.

$$w(x|y) = \frac{1}{f_{p_{T,\mathbf{J}}}(p_{T,\mathbf{J}}|y)}$$

  The marginal distribution is approximated by the kernel density estimation.
- Weight initialization: He uniform
- L2 regularization: weight decay constant: 0.001
- Early stopping: patience = 50
- Use moving average of weights and bias for the validation and test.
  Ignore early t_0=50 epochs.

- Batch size: modular NN: 20, 50, 100, CNN: 100, 200, 500
- Tested two random seeds
- Select a network with the smallest validation AUC
- Cross-validate the trained model with the model with a focal loss.

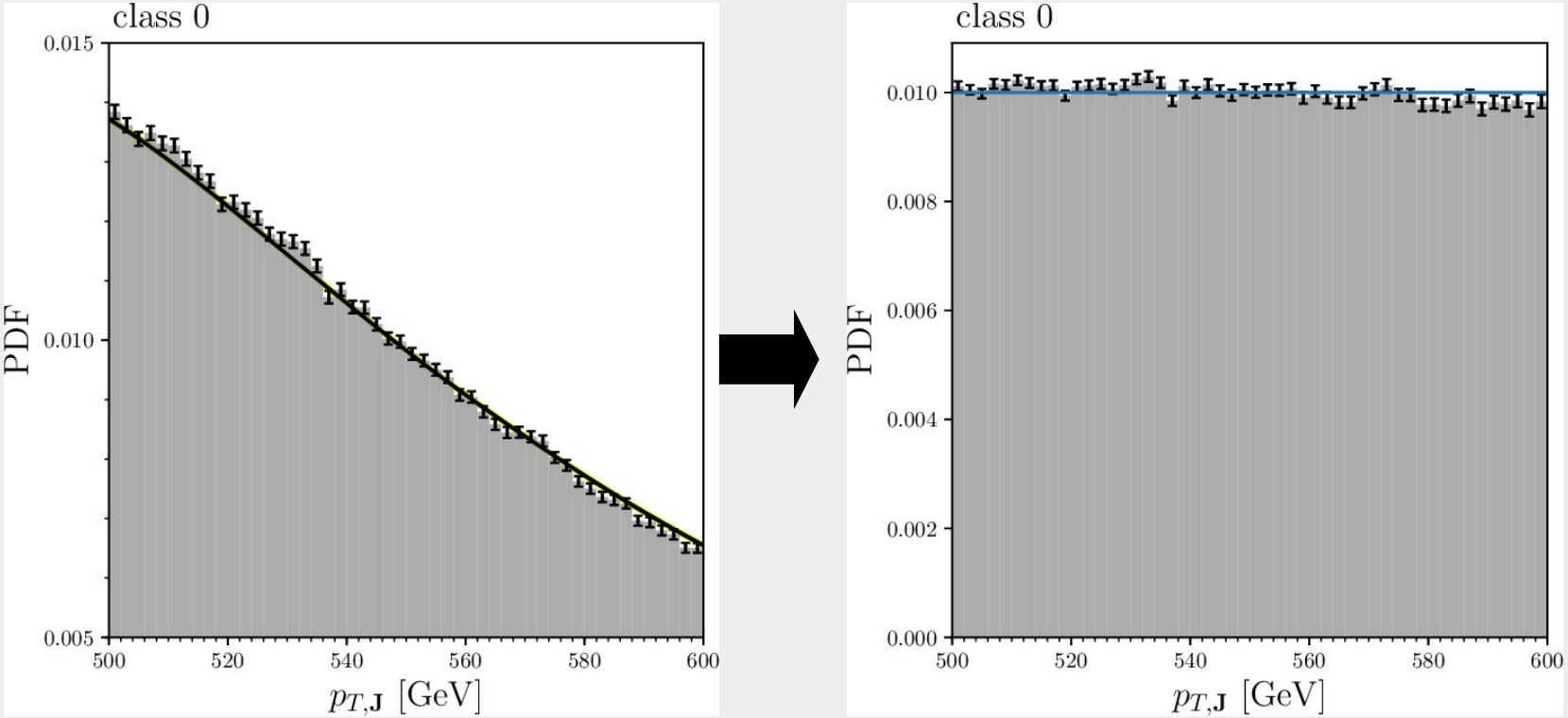
Training setup

- The model is implemented with Keras with backend tensorflow.
- Optimizer: ADAM, mimimize the weighted cross-entropy.

$$\mathcal{L} = \sum_{i=0}^{N_{\text{event}}} w^{(i)}(x)\, y^{(i)}\, \log \hat{y}^{(i)}(x)$$

- $p_{T,J}$ distribution is reweighted to be flat.
  The marginal distribution is approximated by the kernel density estimation.

$$w(x) = \frac{1}{f_{p_{T,\mathbf{J}}}(p_{T,\mathbf{J}})}$$



Training setup

- Weight initialization: He uniform
- L2 regularization: weight decay constant: 0.001
- Early stopping: patience = 50
- Use moving average of weights and bias for the validation and test. Ignore early t_0=50 epochs.

$$\bar{\theta}^{(t)} = \alpha\bar{\theta}^{(t-1)} + (1-\alpha)\theta^{(t)}$$

$$\hat{\theta}^{(t)} = \frac{1}{1-\alpha^{t-t_0+1}}\bar{\theta}^{(t)}$$

For training: $q(\mathrm{top}|x;\theta^{(t)})$ For validation and test: $q(\mathrm{top}|x;\hat{\theta}^{(t)})$

- Batch size: modular NN: 20, 50, 100, CNN: 100, 200, 500
- Tested two random seeds
- Select a network with the smallest validation AUC

Training setup

- Weight initialization: He uniform
- L2 regularization: weight decay constant: 0.001
- Early stopping: patience = 50
- Use moving average of weights and bias for the validation and test. Ignore early t_0=50 epochs.

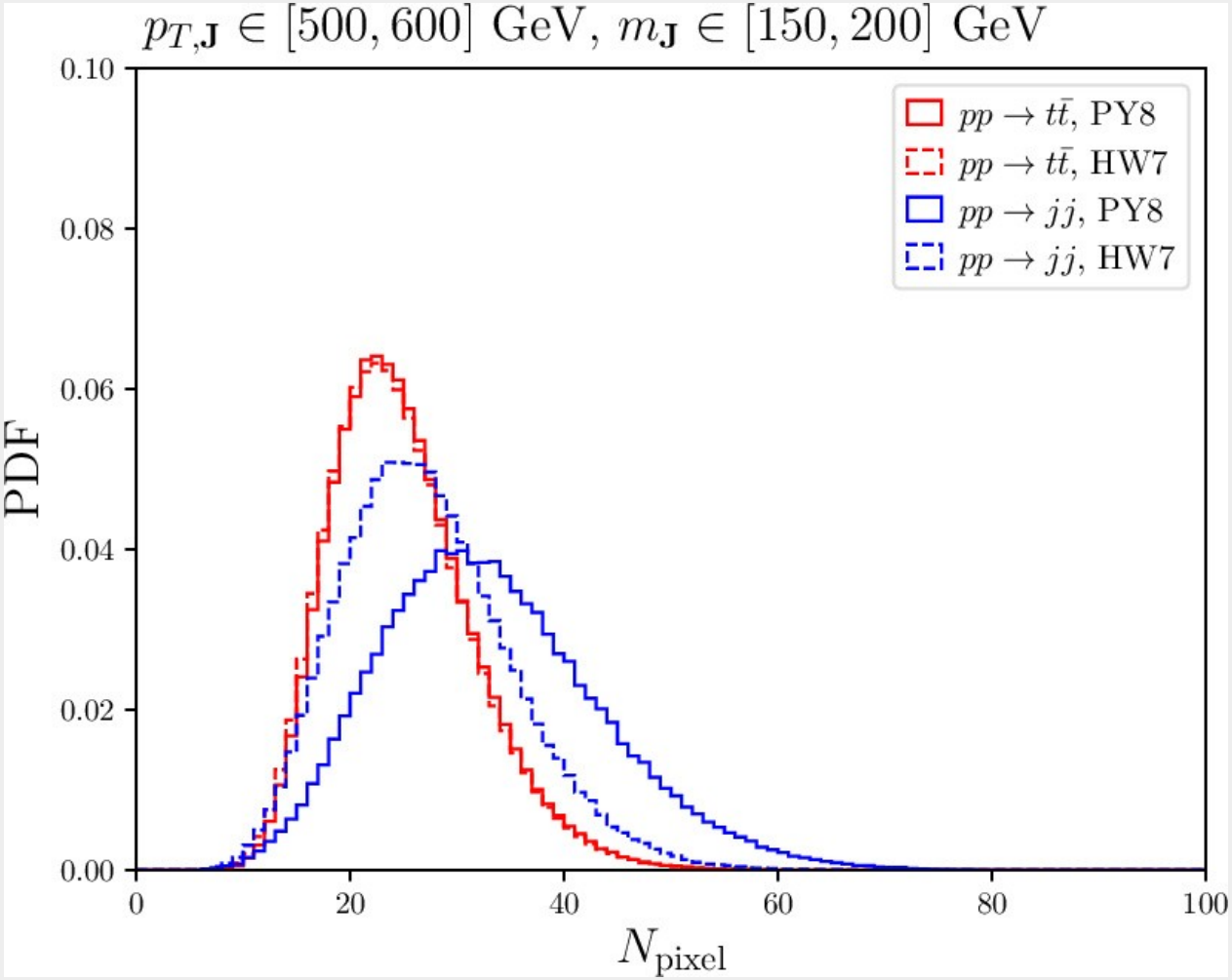
$$\bar{\theta}^{(t)} = \alpha\bar{\theta}^{(t-1)} + (1-\alpha)\theta^{(t)}$$

$$\hat{\theta}^{(t)} = \frac{1}{1-\alpha^{t-t_0+1}}\bar{\theta}^{(t)}$$

For training: $q(\mathrm{top}|x;\theta^{(t)})$ For validation and test: $q(\mathrm{top}|x;\hat{\theta}^{(t)})$

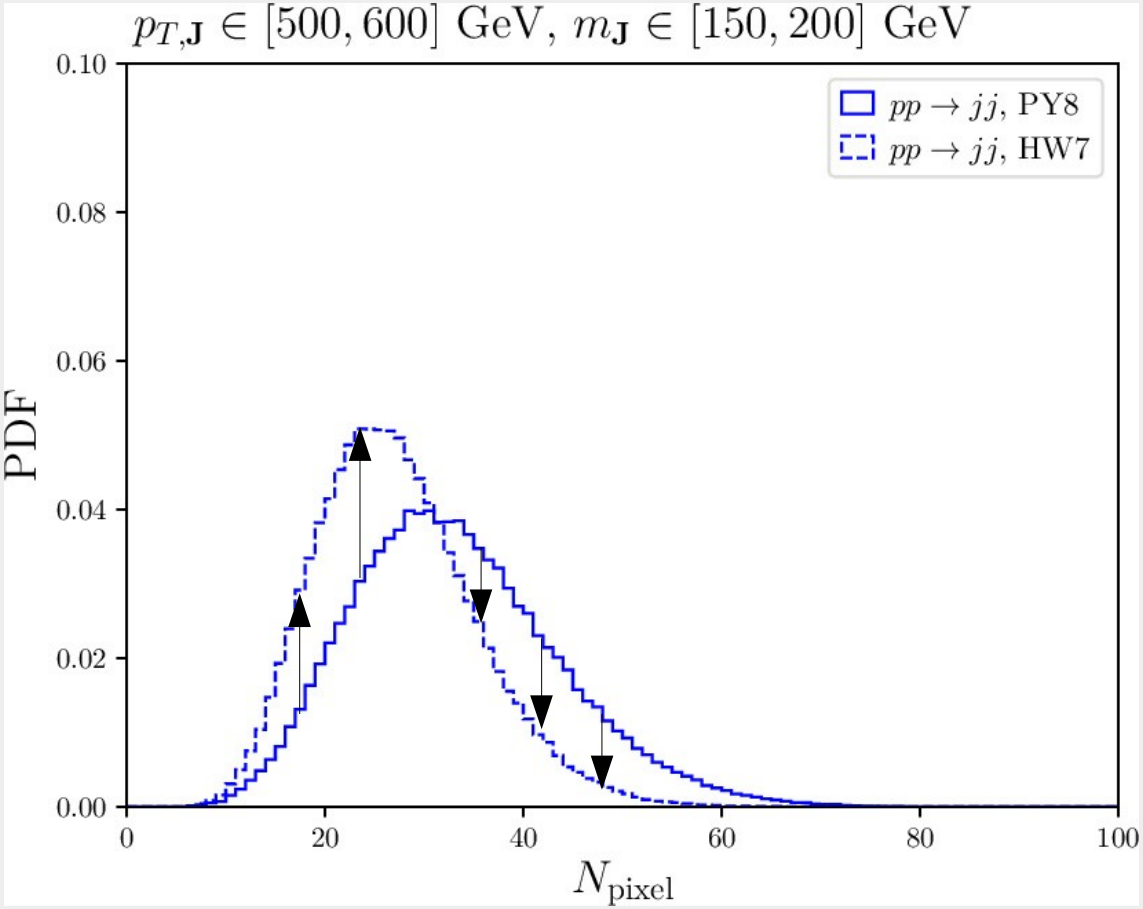
- Batch size: modular NN: 20, 50, 100, CNN: 100, 200, 500
- Tested two random seeds
- Select a network with the smallest validation AUC

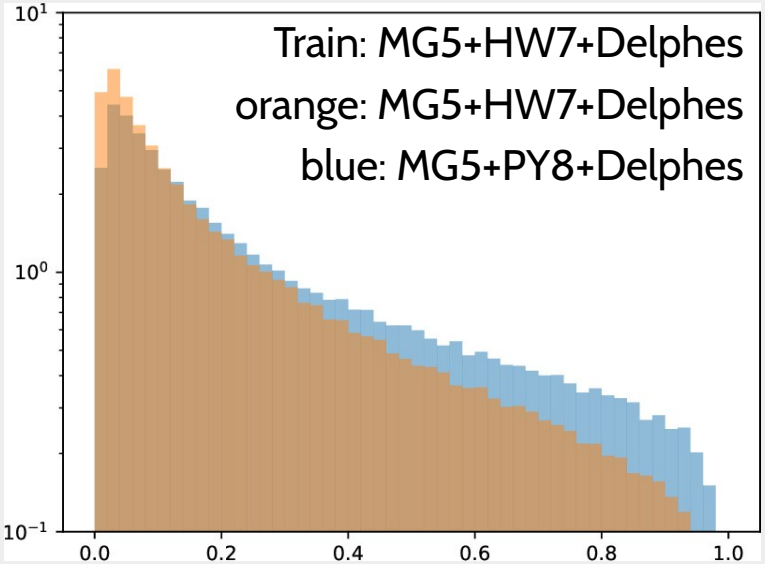
N_{pixel} distribution: top jet and QCD jet

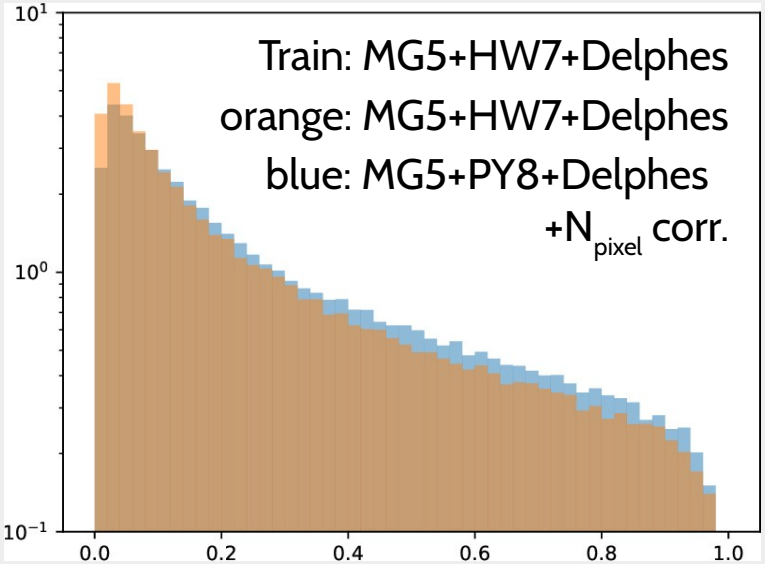


pp → jj samples are gluon jet rich, so that the deviation is large.

Correcting MC: reweighting PY8 to HW7



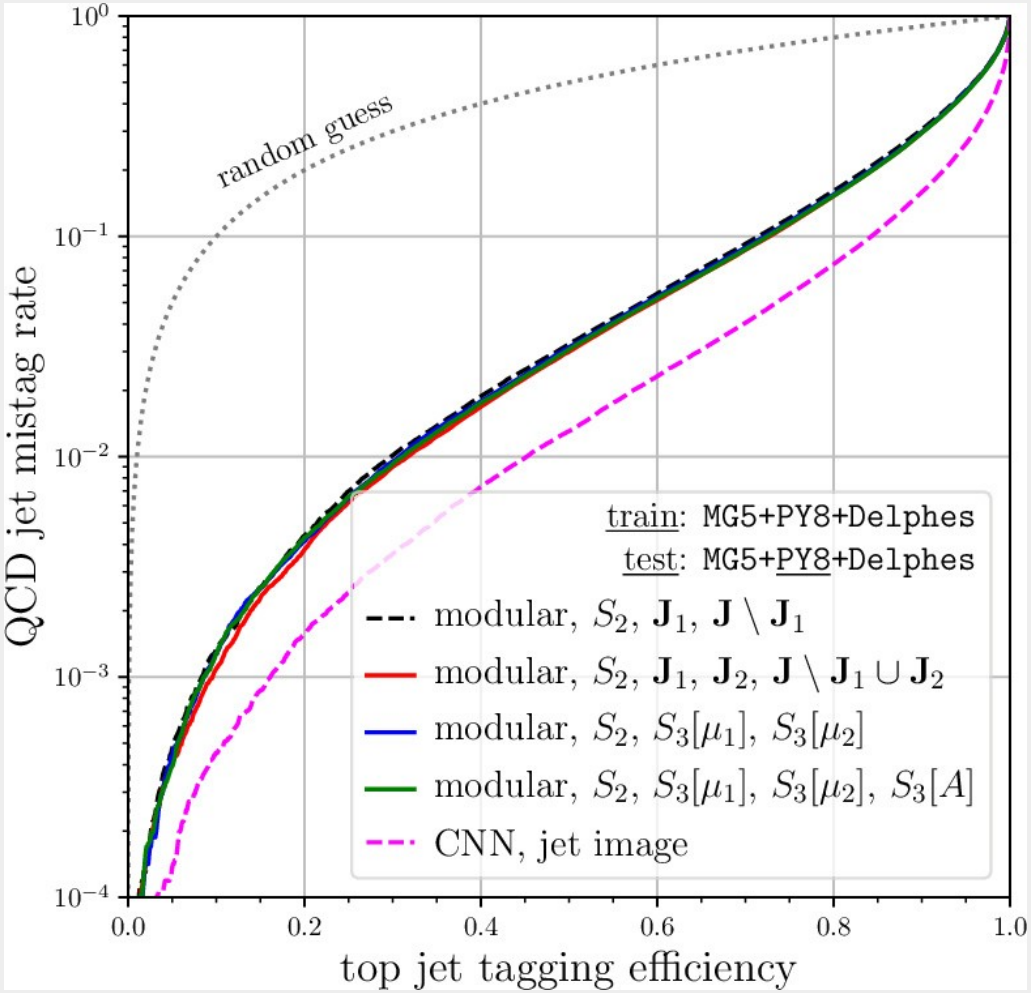




We rescale N_{pixel} distirbution of PY8 dijet samples to that of HW7 dijet samples.
The NN output distributions between PY8 and HW7 are more close.

$\hat{y} = q(\text{top}|x)$

Including higher order terms...



N_{pixel} distribution: Higgs jet and QCD jet