Classification of Jets using Jet Morphology and Deep Learning

Amit Chakraborty
IISc, Bangalore

Anomalies 2020
Sep 11, 2020

Motivation

- **Post-Higgs discovery**: Non observation of (statistically) significant excess over SM expectation at LHC ... anomalies at several low/high energy expts!
- Severe constraint on well-motivated Beyond SM scenarios ...
- Machine Learning (Deep Learning): Outperformed traditional approach ... huge excitement within Particle Physics community!
- Many applications with success: Jet classification, Anomaly detection, Particle detection, Pileups, ...
- "Black box" models, famous for their performances, but not so trivial to extract specific physics knowledge(s) ...

Motivation

Can we achieve Convolutional Neural Network (CNN) level performance with calculable physics observables for Classifying Jets?

We find,

- Possible to obtain classification performance (comparable to CNN) e.g., Jet Spectrum!

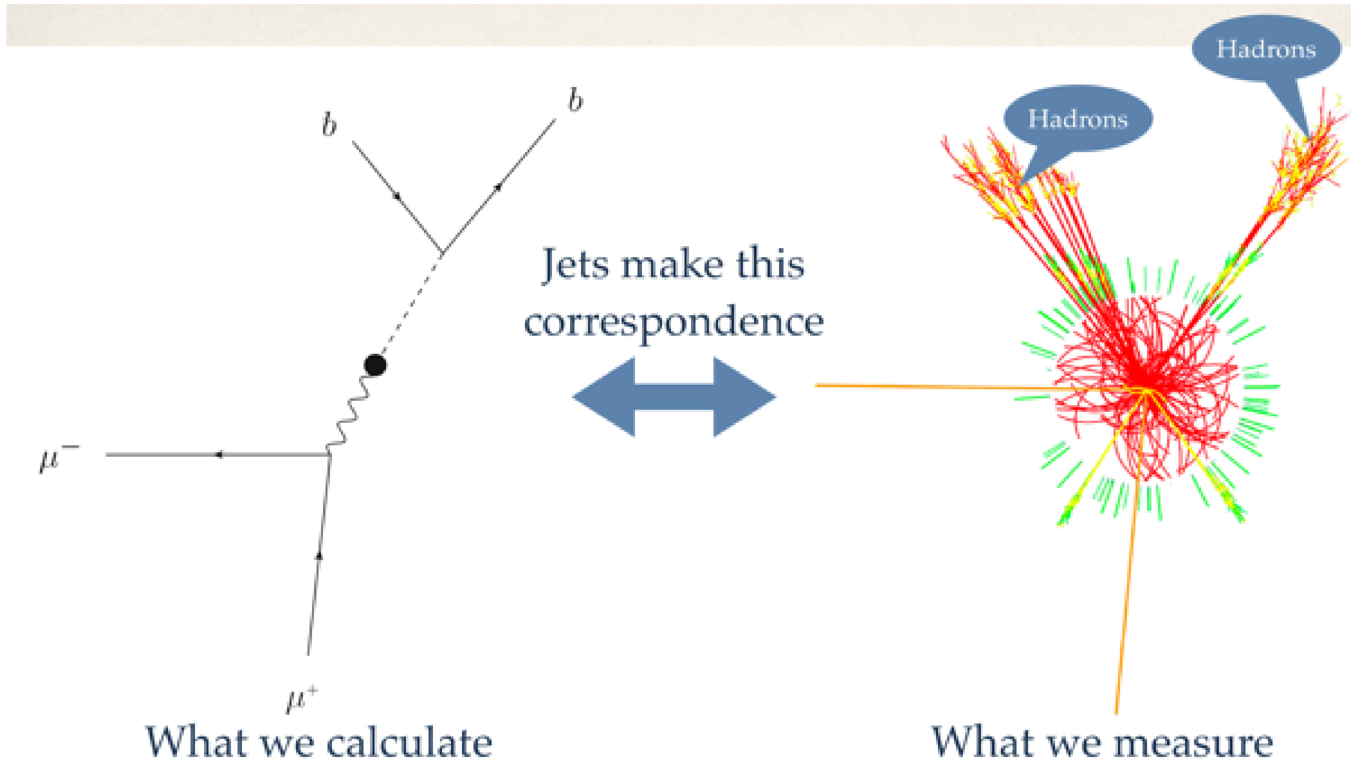
- Two examples:

    - Higgs jet vs QCD jet classification
    - Top jet vs QCD jet classification: Need to include additional inputs from Jet Morphology!

Based on:

AC, Lim and Nojiri, JHEP 07 135 (2019)
AC, Lim, Nojiri and Takeuchi, JHEP 07 111 (2020)

Jets



Calorimeter and Tracker Information clustered together
- Jet Radius (R) and jet algorithm (kT, anti-kT, C/A)

Map to the underlying physics!

Classification of Jets

Goal: To know the jets of SM particles, apply the knowledge to BSM Physics!

- **General strategy**:

  - Nature and Multiplicity of constituent particles, ratio of EM to hadronic energy deposits, Vertex information …

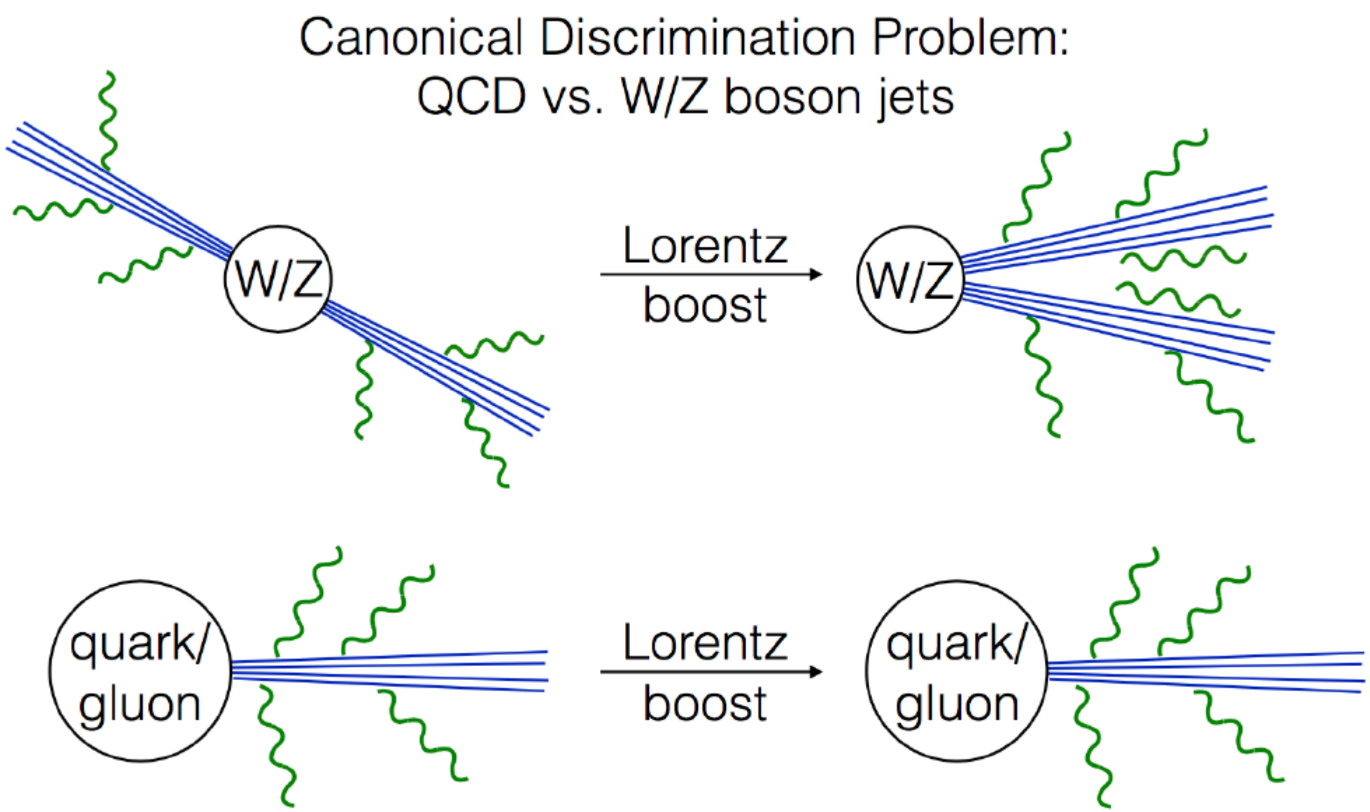
  - Distribution of energy deposits inside the Jet ... e.g., widely distributed or, prong-like structured

- **Boosted particles**: As centre-of-mass energy increases at LHC, particles with large transverse momentum, classification become a challenging task!

Boosted Jet Classification

W/Z Jet

QCD Jet



2-prong object

1-prong object

"Jet Substructure Technique"

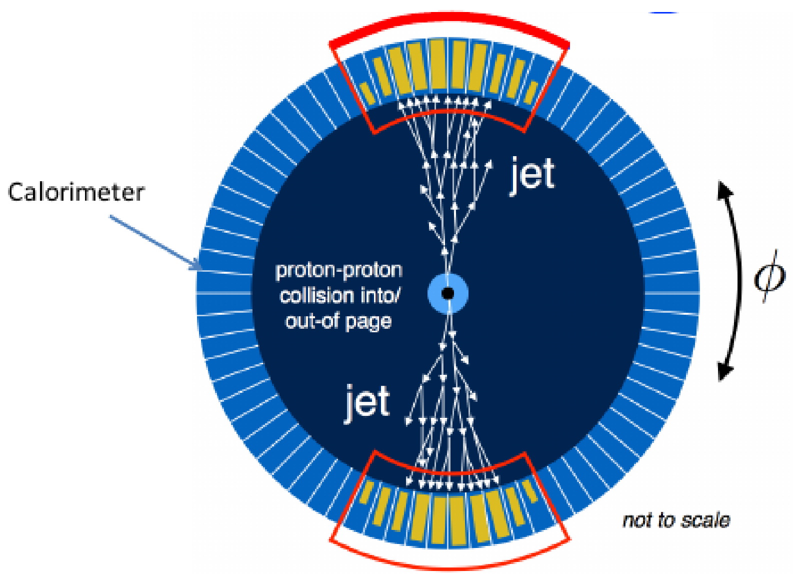
Buttherworth et. al. PRL 100 242001 (2008)

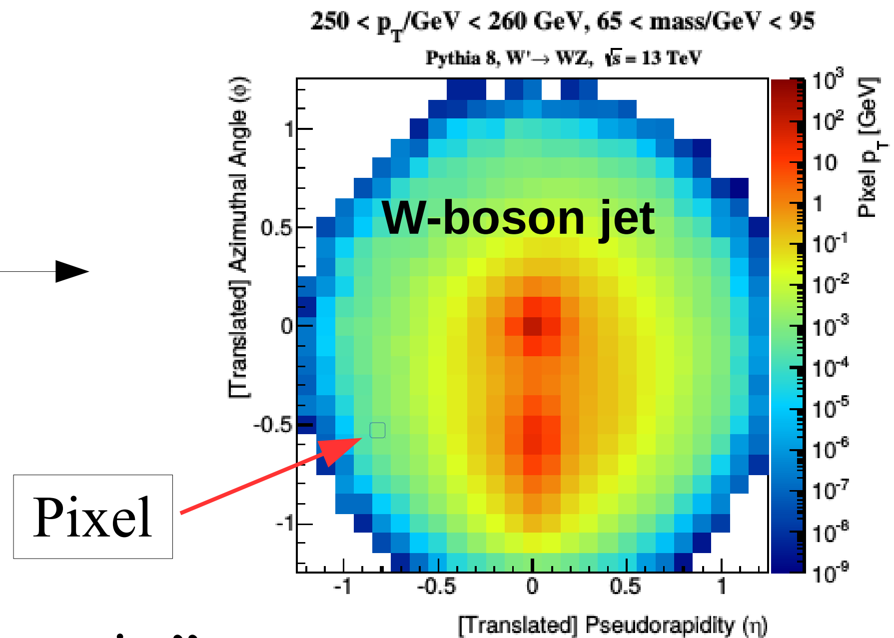
Look inside the "Fatjet", study the energy flow inside ...

Probe BSM particles using Fatjets (Higgs, Top, W/Z jets) ...

Jets as "Image"

Oliveria et. al., JHEP 07, 069 (2016)

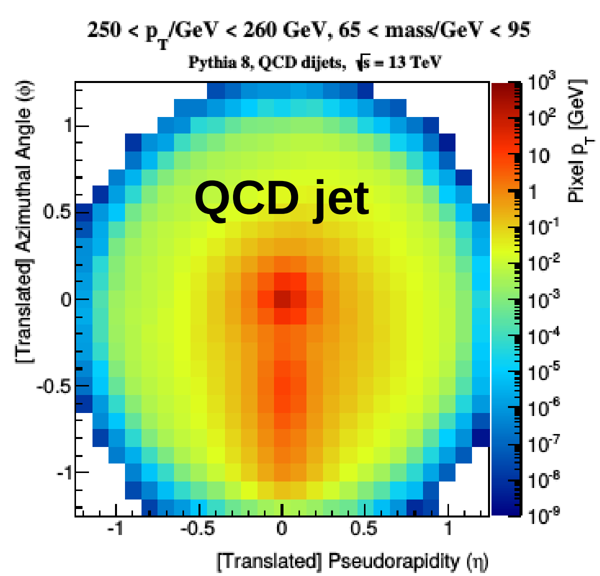




- Calorimeters "Camera", pixels "energy deposits"

- Paradigm shift for visualizing and classifying jets.

- Significant improvement using ML
@Experiment: real data combined with MC!
(e.g., DPS-2017-013, DP-2018/046 ... many more)

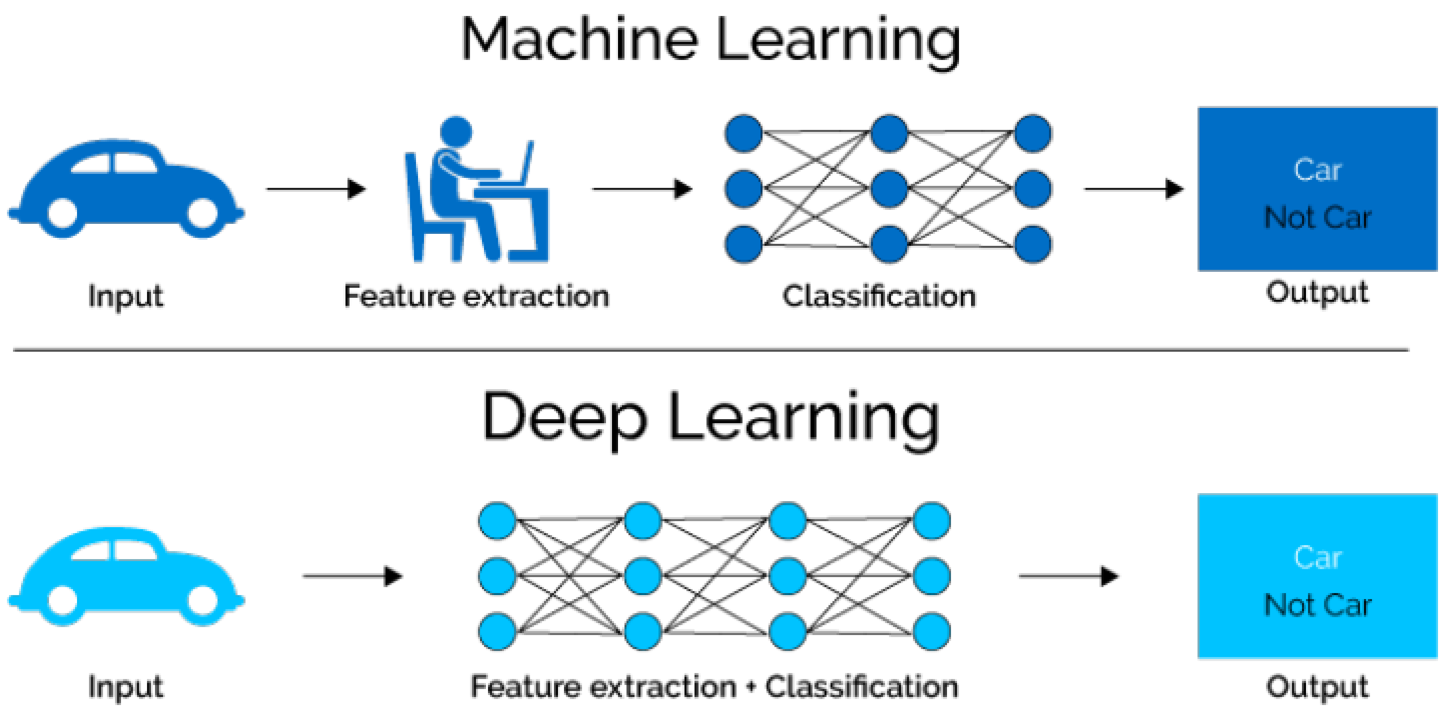


Machine Learning

"Giving computers the ability to learn without explicitly programming them" (Arthur Samuel, 1959)

Standard approach

Modern approach



From towardsdatascience.com

Replace "Car" with "Jet":

e.g., Jet → **mass, pT, njets, ...** → Cuts → Higgs / QCD

Jet → **Image/4-mom** → Higgs / QCD

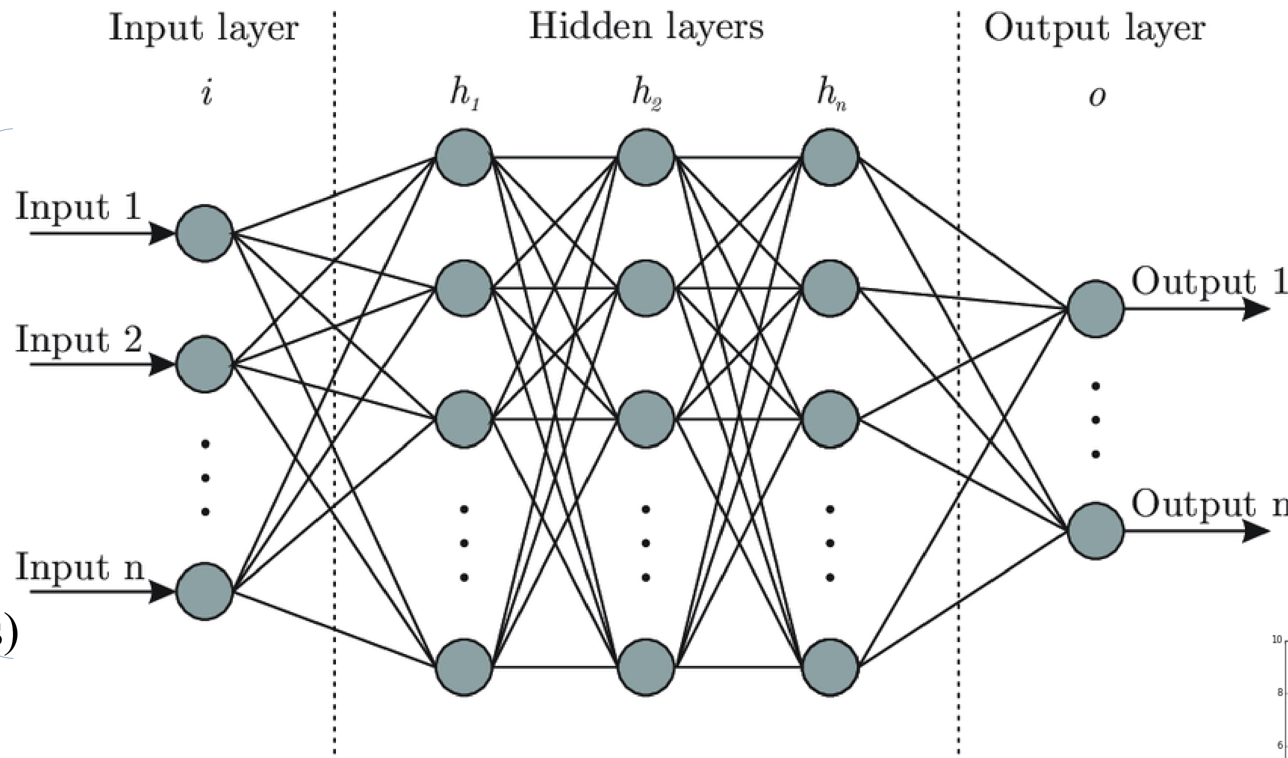
- **<u>Types</u>**: Supervised learning, Unsupervised leaning ...

[Courtesy to D. Shih, Ikaho talk]

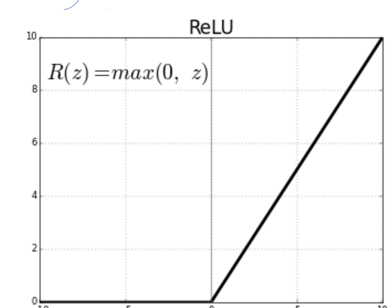
Neural Network Architecture

Image Or, Observables

(Activation fns, Bias, hyperparameters)

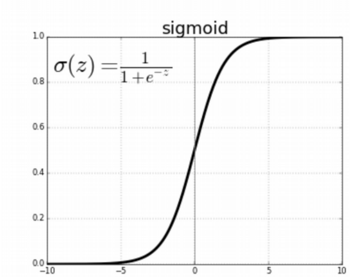


Output



- Representation of the Input data using multiple "weights"!
  (more "hidden" layers, more finer features are learned!)

- <u>Convolutional NN</u>:
  (In general) One of the best Classifiers till date!



What are these "Black-box Models" Learning?

Jet Spectra

Spectral Analysis: Jet → Constituents
(Energy Deposits in Trackers and/or Calorimeters)

$$S_2(R;\Delta R) = \frac{1}{\Delta R} \sum_{\substack{i,j\in \text{jet} \\ R_{ij}\in[R,R+\Delta R)}} p_{T,i}\, p_{T,j},$$

$$R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$$

Resolution parameter : $\Delta R = 0.1$

- 2-point energy correlation function among the Jet constituents
(derivable from a General classifier with jet constituents)

Spectrum: distribution binned in R = [0, 2*jet radius]

- Jet as a "Graph" with Vertices and Edges!

Similar proposals,
Tkachov Int. J. Mod. Phy A12 (1997), Jankowiak et al JHEP 06 057(2011), Thaler et al JHEP 04 013 (2018)

Jet Spectra

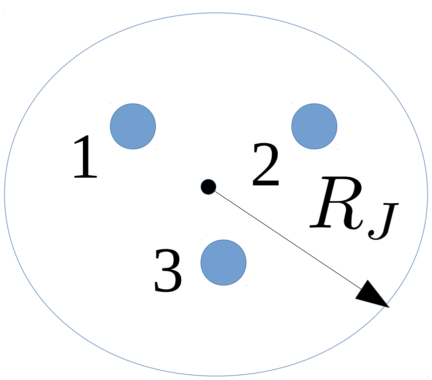
Spectral Analysis: Jet → Constituents
(Energy Deposits in Trackers & Calorimeters)

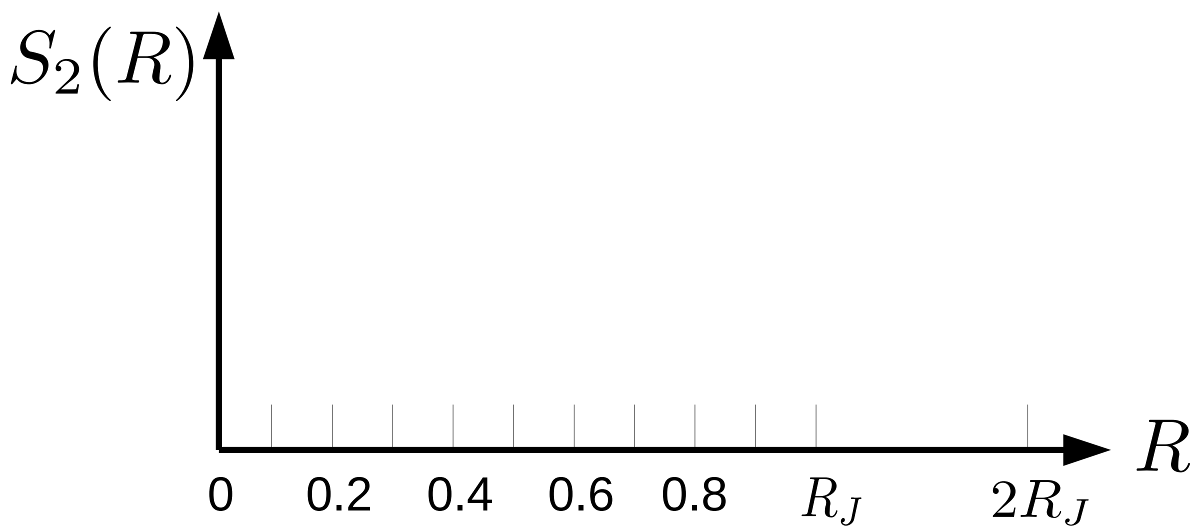
$$S_2(R;\Delta R) = \frac{1}{\Delta R} \sum_{\substack{i,j\in\text{jet} \\ R_{ij}\in[R,R+\Delta R)}} p_{T,i}p_{T,j},$$

$$R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$$

$\Delta R = 0.1$ (Resolution parameter)

An Event





Jet Spectra

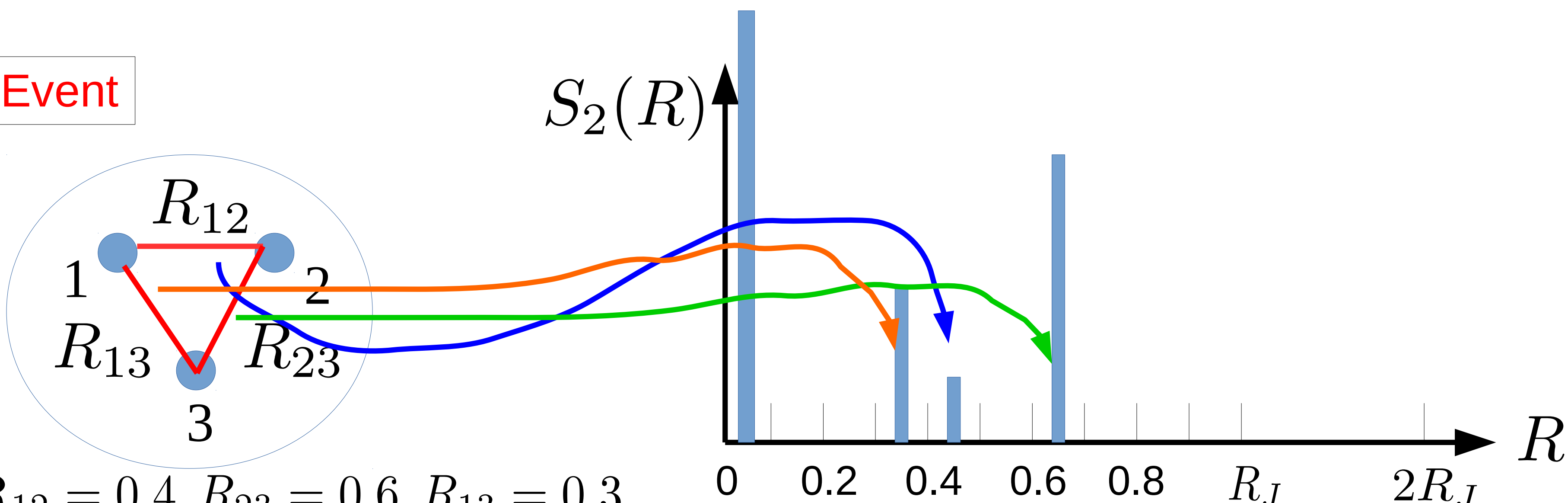
<u>Spectral Analysis</u>: Jet $\rightarrow$ Constituents
(Energy Deposits in Trackers and/or Calorimeters)

$$S_2(R;\Delta R) = \frac{1}{\Delta R} \sum_{\substack{i,j\in \text{jet} \\ R_{ij}\in[R,R+\Delta R)}} p_{T,i}p_{T,j},$$

$$R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$$

$\Delta R = 0.1$ (Resolution parameter)

An Event



e.g.,

$R_{12} = 0.4, R_{23} = 0.6, R_{13} = 0.3$

$p_{T,1} = 2$ GeV

$p_{T,2} = 3$ GeV

$p_{T,3} = 5$ GeV

$S_2(R) = (38, 0, 0, 10, 6, 0, 15, \ldots)$

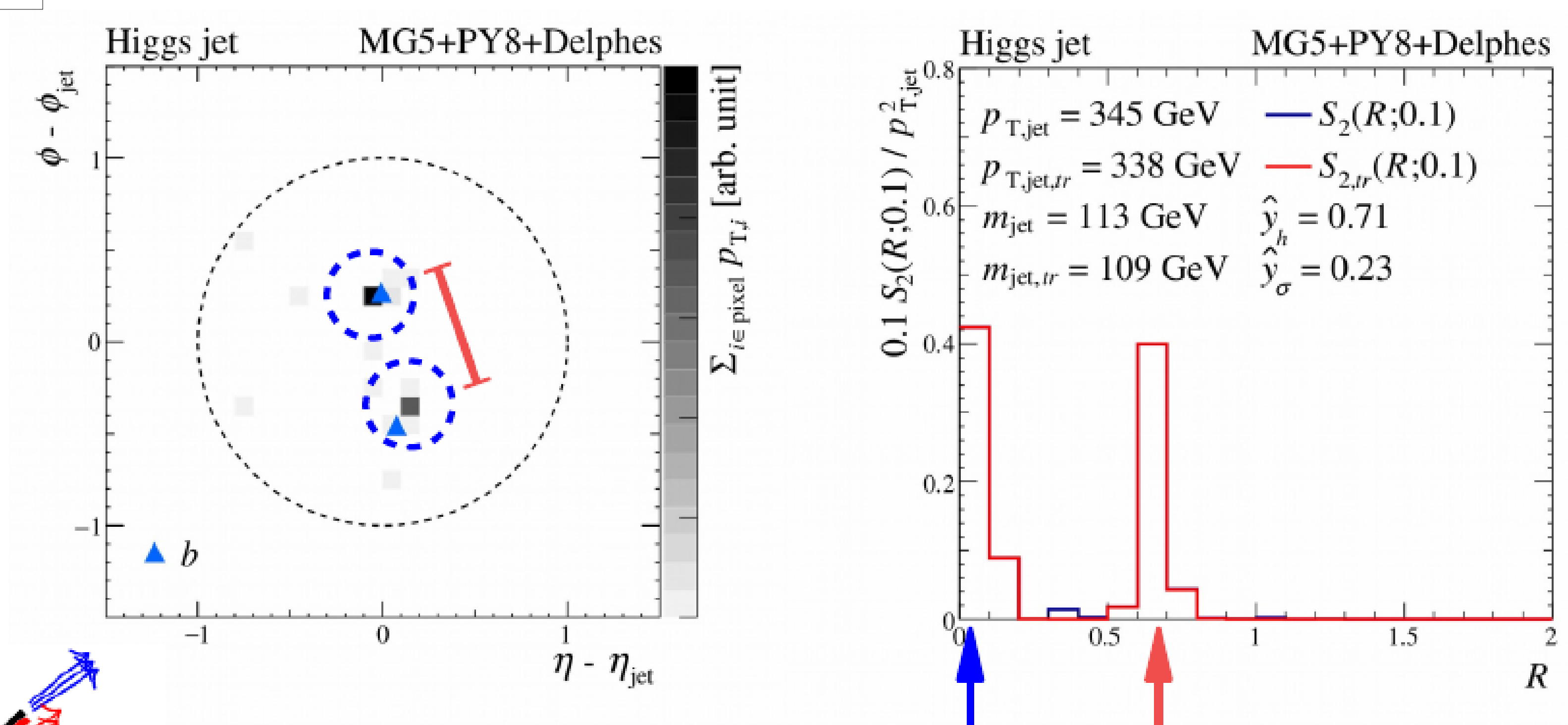
<u>Peak</u> at R values where most of the energetic particles are present ...

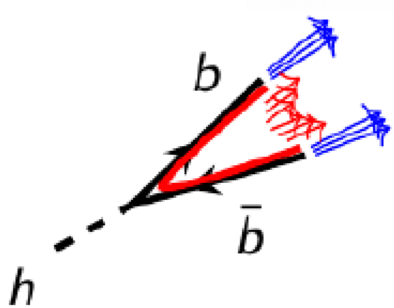
Jet Spectrum

$$S_2(R;\Delta R) = \frac{1}{\Delta R} \sum_{\substack{i,j\in \text{jet} \\ R_{ij}\in[R,R+\Delta R)}} p_{T,i}\,p_{T,j},$$

Higgs Jet

$$R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$$





Characteristic Higgs peaks observed!

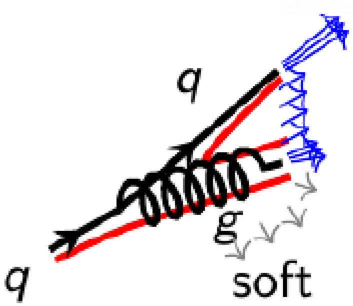
[courtesy to S.H.Lim]

Jet Spectrum

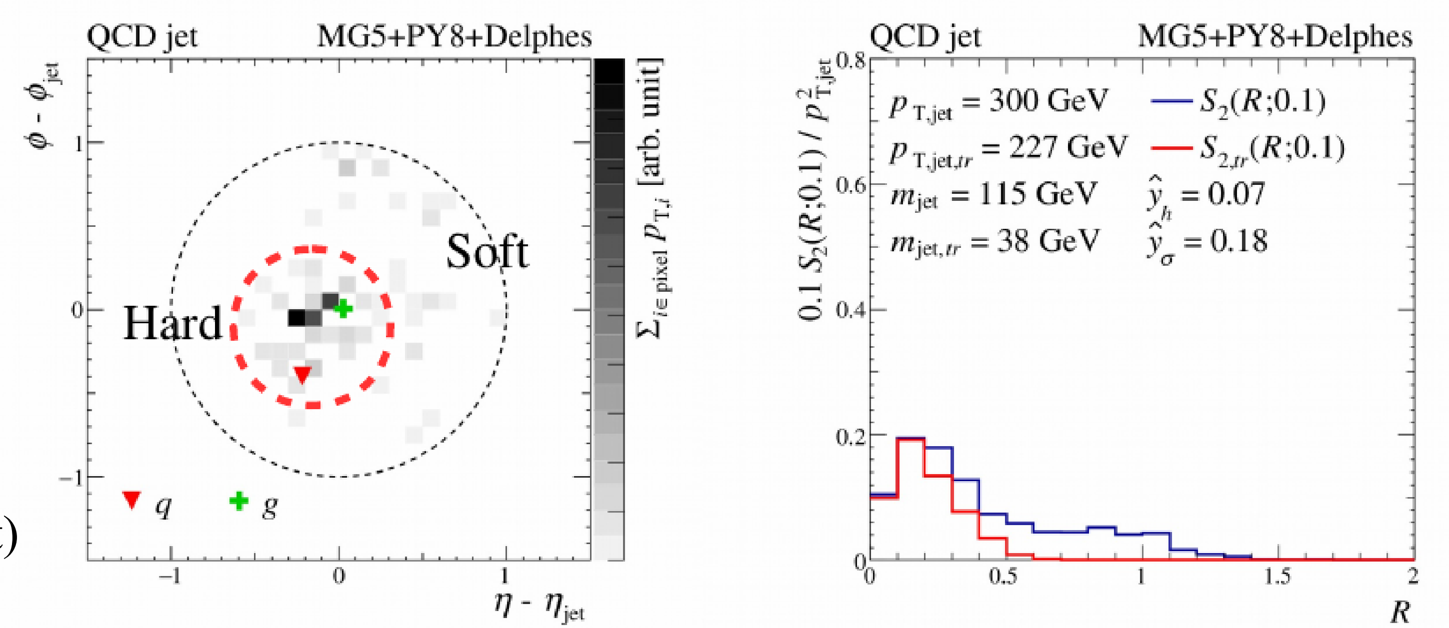
$$S_2(R;\Delta R) = \frac{1}{\Delta R} \sum_{\substack{i,j\in\text{jet} \\ R_{ij}\in[R,R+\Delta R)}} p_{T,i}p_{T,j},$$

QCD Jet

$$R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$$



(same for Gluon jet)

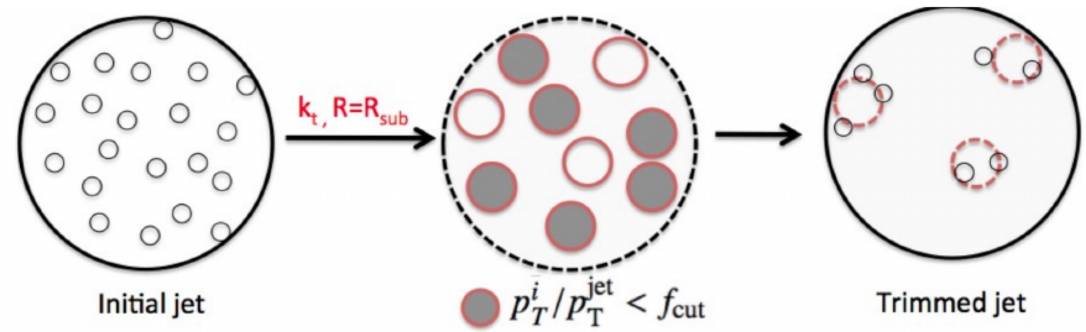


"Hard" center surrounded by soft particles, smoothly falling distribution ...

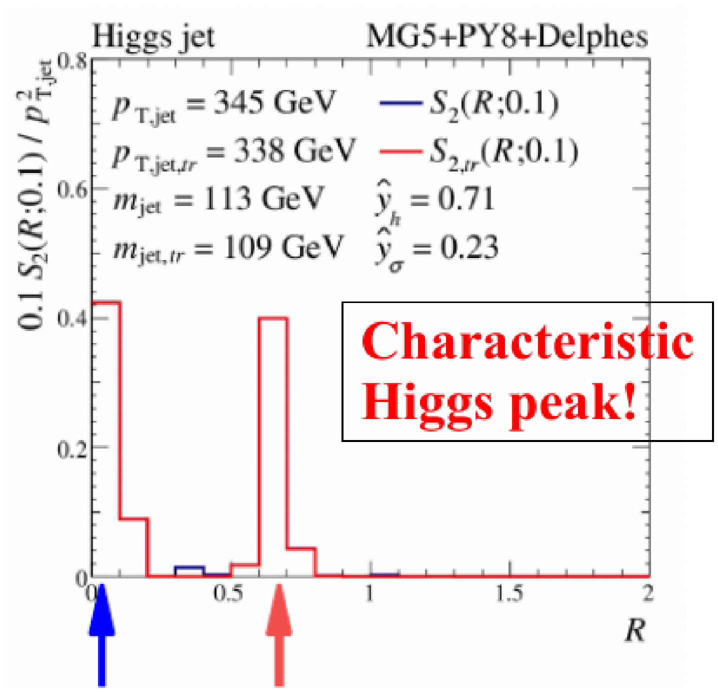
[courtesy to S.H.Lim]

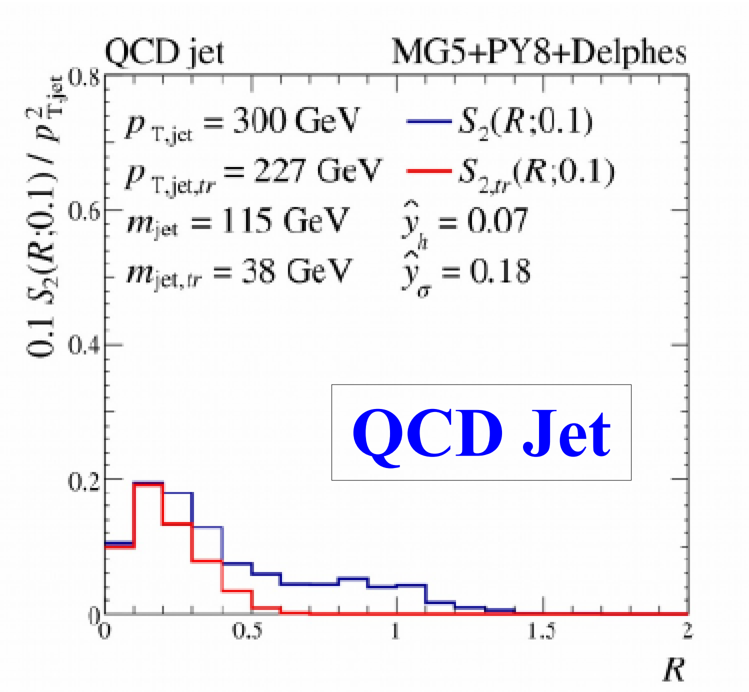
Jet Trimming ...

Krohn, Thaler and Wang, JHEP 02, 084 (2010)



Robust under
UE & MI events!





Cuts the long tail ... Removes "soft" components, keeps the interesting parts!

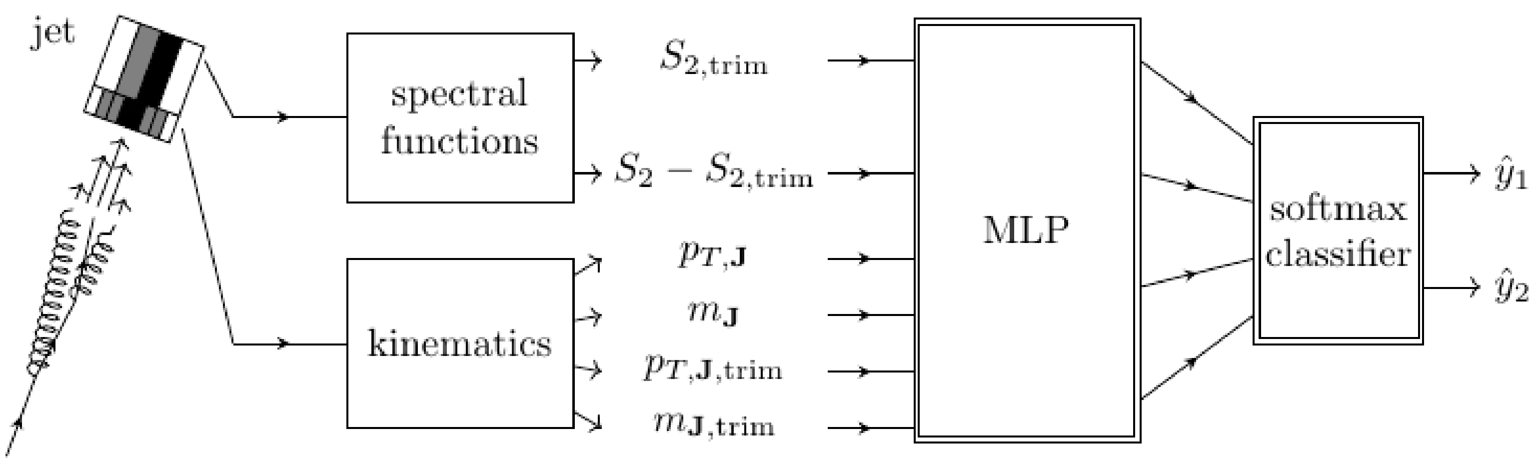
Our Network

We define a quantity,

$$S_{2,\mathrm{soft}}(R;\Delta R) = S_2(R;\Delta R) - S_{2,\mathrm{trim}}(R;\Delta R)$$

We keep the "soft part"!!

$$\begin{aligned} S_{2,\mathrm{trim}}(R;\Delta R) &= p_{T,\mathbf{J}}^2 \cdot \mathcal{O}\left[1\right], \\ S_{2,\mathrm{soft}}(R;\Delta R) &= p_{T,\mathbf{J}}^2 \cdot \left(\mathcal{O}\left[f_{\mathrm{trim}}\right] + \mathcal{O}\left[f_{\mathrm{trim}}^2\right]\right) \end{aligned}$$



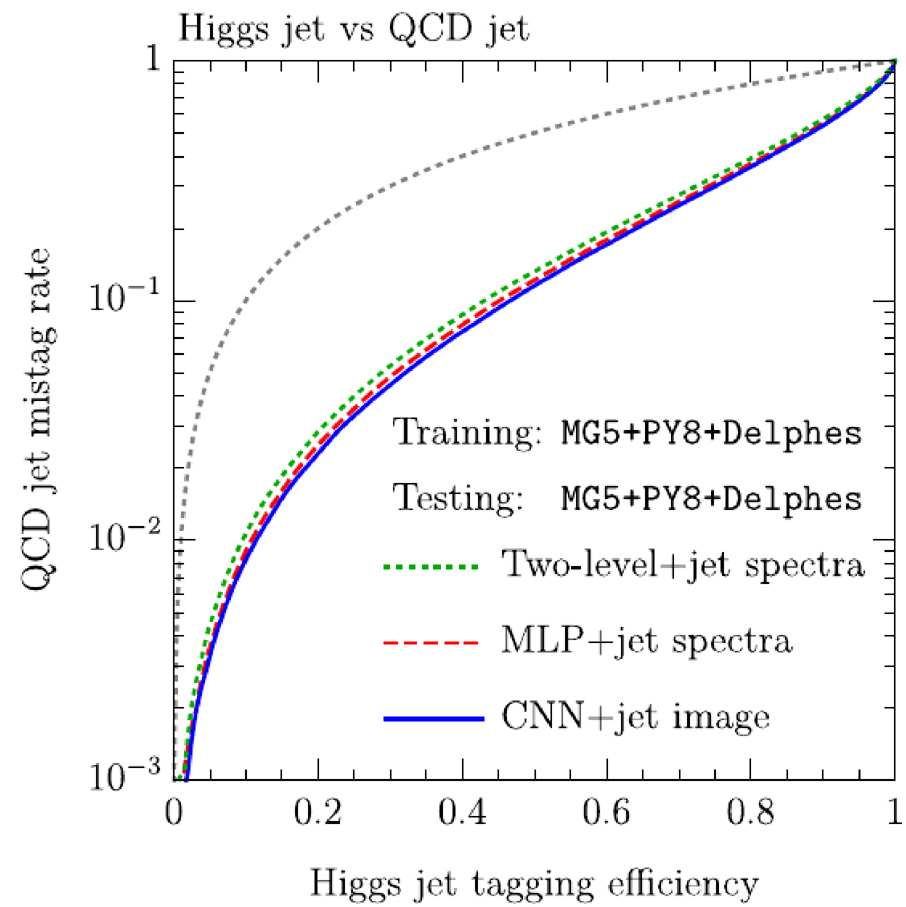
Train the network using S2 spectra and compare with CNN classification!

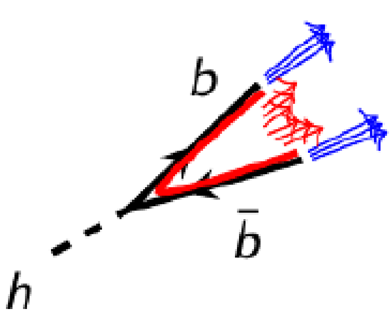
Higgs jet vs QCD jet

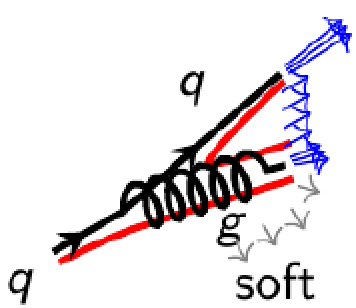
AC, Lim and Nojiri, JHEP 07 135 (2019)

- Performance comparable to CNN
  (Also, similar to D2)

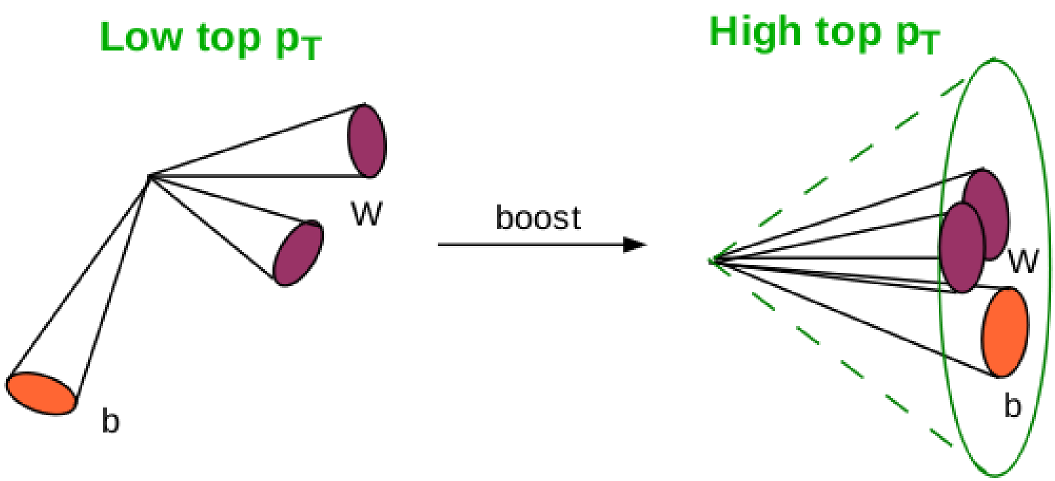
- No Information loss,
  Smaller no of Inputs:
  CNN (~20 * 20), DNN (2 * 20)





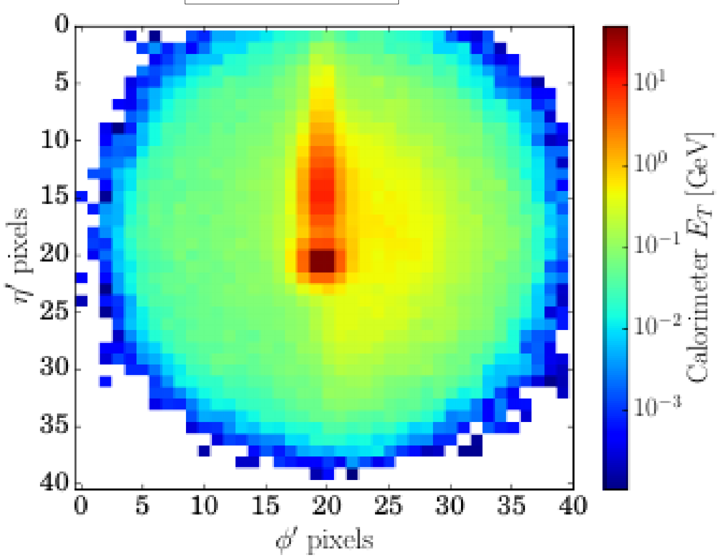


Top Jet

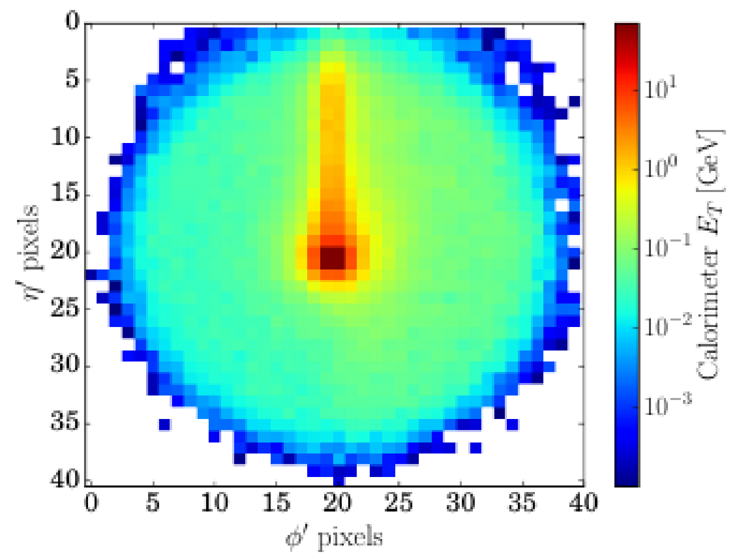


R = 1.5 Top fatjet

Top jet



QCD jet



Top jet has More activity Away from the Centre!

How about the Jet Spectrum?

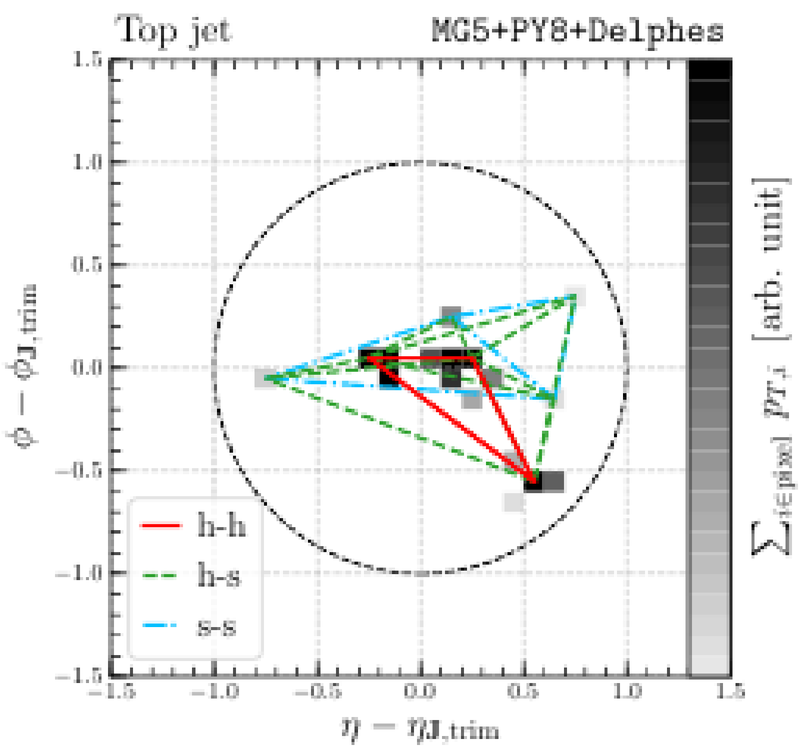
[hep-ph 1701.08784 Plehn et. al.]

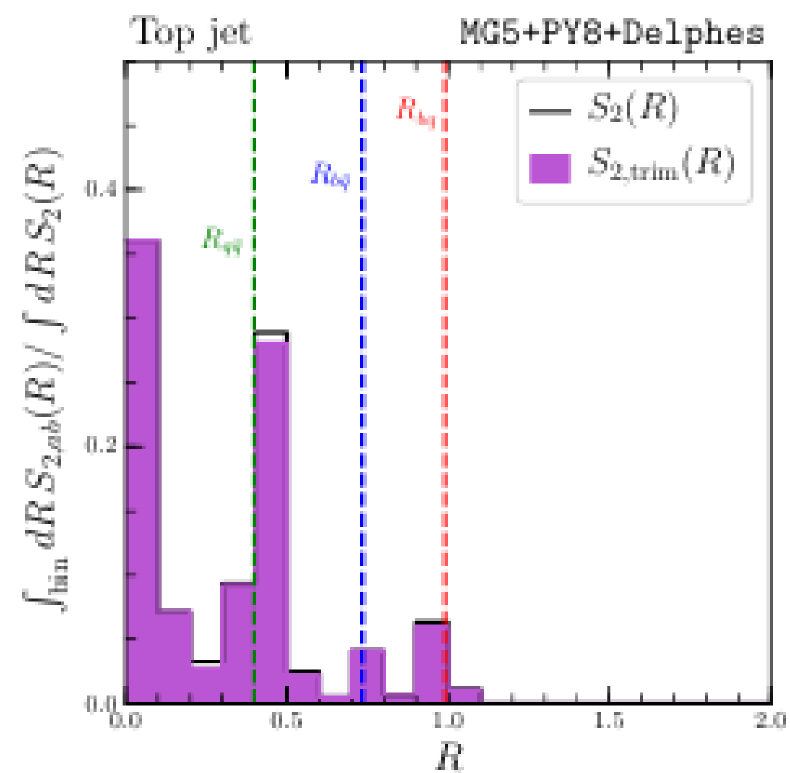
Top Jet

Trimmed (3 prong) Top jet must have 4 peaks in the S2 spectrum!

Parton level

$$S_{2,\mathrm{trim}}(R) = (p_{T,b}^2 + p_{T,q}^2 + p_{T,\bar{q}}^2)\,\delta(R) + 2p_{T,b}p_{T,q}\delta(R - R_{bq}) + 2p_{T,b}p_{T,\bar{q}}\delta(R - R_{b\bar{q}}) + 2p_{T,q}p_{T,\bar{q}}\delta(R - R_{q\bar{q}})$$

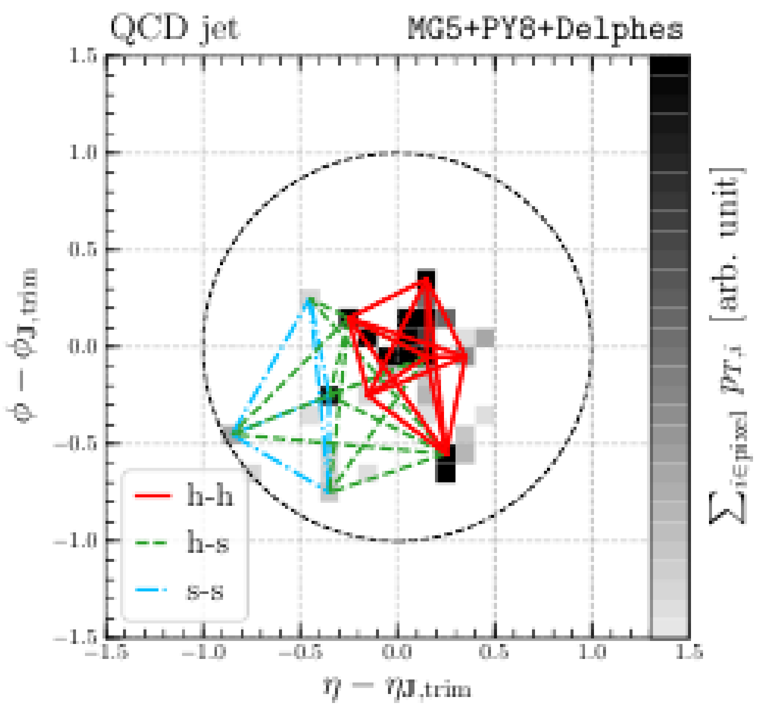


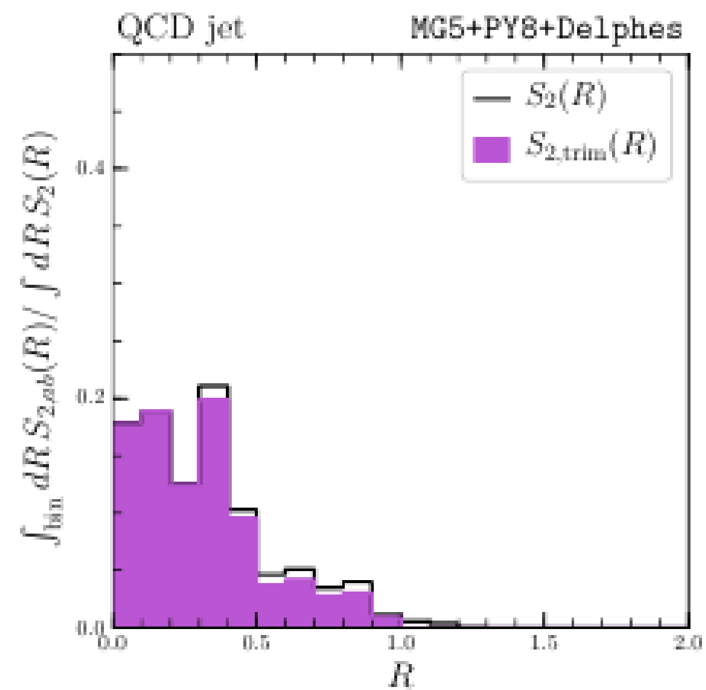


AC, Lim, Nojiri and Takeuchi, JHEP 07 111 (2020)

QCD Jet

Trimmed jet spectrum peaks at smaller values of R!

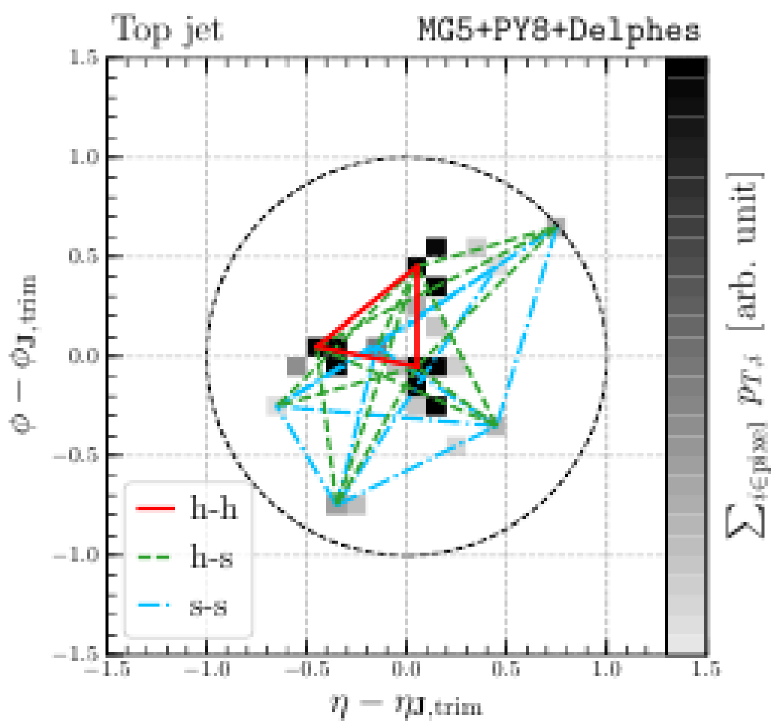


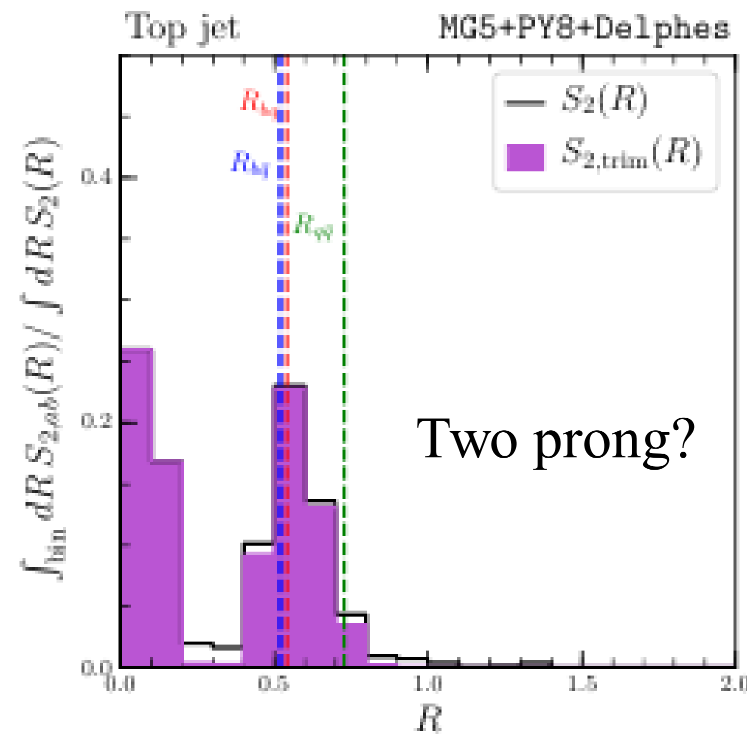


Depending on the transverse momentum,

Top jet spectrum may also show 1/2 peaks!

Overlapping subjets





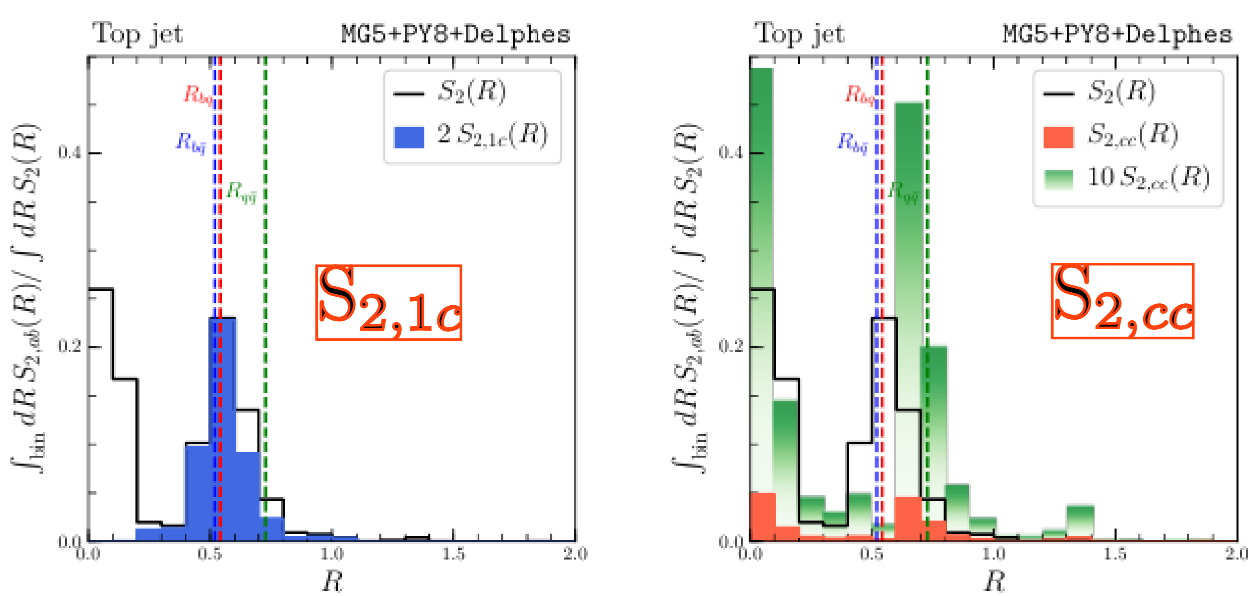
- Additional correlations may help!

- Like Trimmed-Soft components,
  how about calculating correlations at the <u>subjet level</u>?

Correlation with the Leading p_T subjet

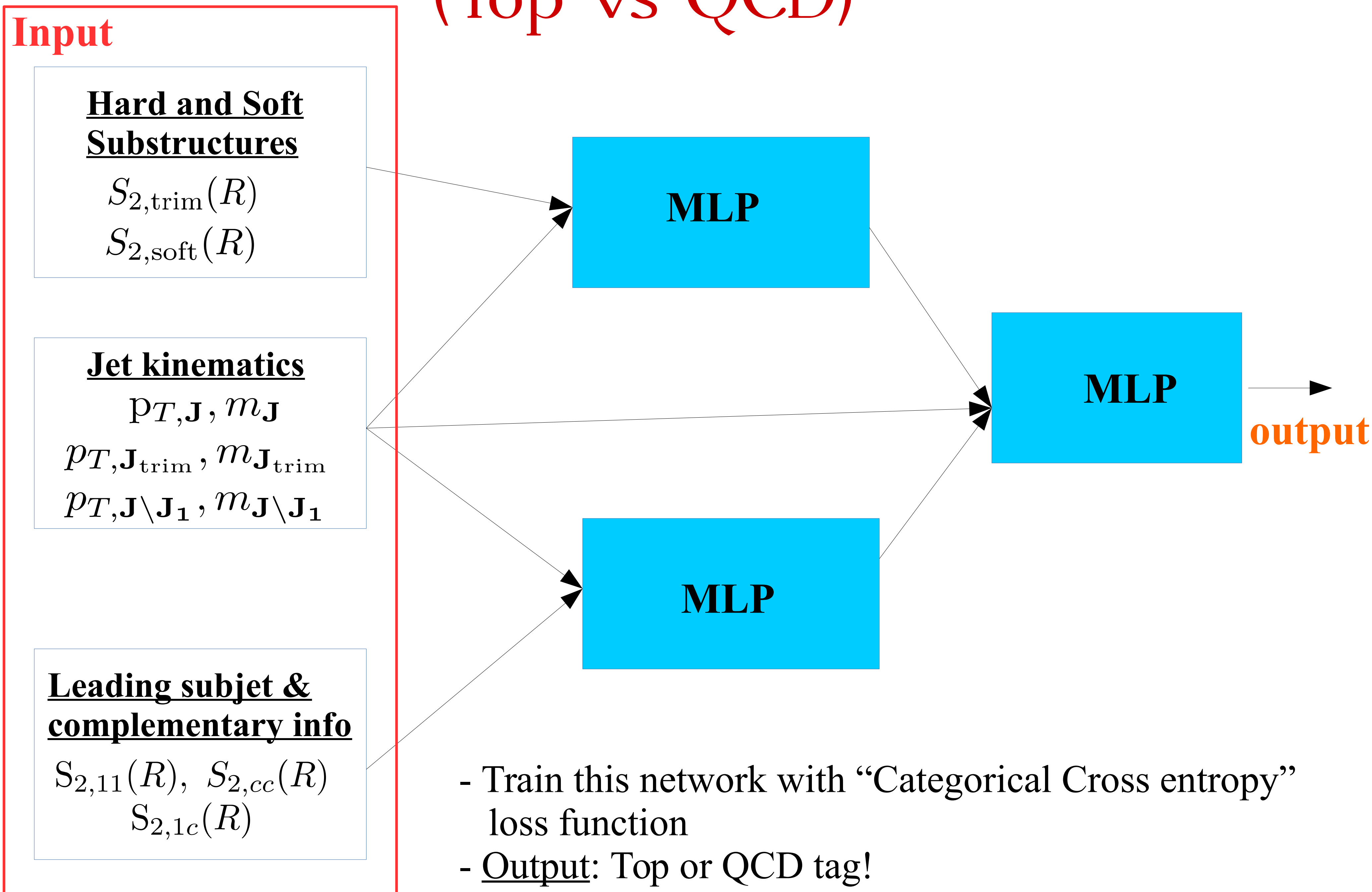
- the leading p_T subjet, $\mathbf{J}_1$, denoted by 1,
- the compliment set of $\mathbf{J}_1$, $\mathbf{J} \setminus \mathbf{J}_1$, denoted by c,

$$\begin{aligned} S_{2,11}(R) &= p_{T,i_1}^2 \delta(R), \\ 2\, S_{2,1c}(R) &= 2p_{T,i_1} p_{T,i_2} \delta(R - R_{i_1 i_2}) + 2p_{T,i_1} p_{T,i_3} \delta(R - R_{i_1 i_3}), \\ S_{2,cc}(R) &= (p_{T,i_2}^2 + p_{T,i_3}^2)\delta(R) + 2p_{T,i_2} p_{T,i_3} \delta(R - R_{i_2 i_3}), \end{aligned}$$



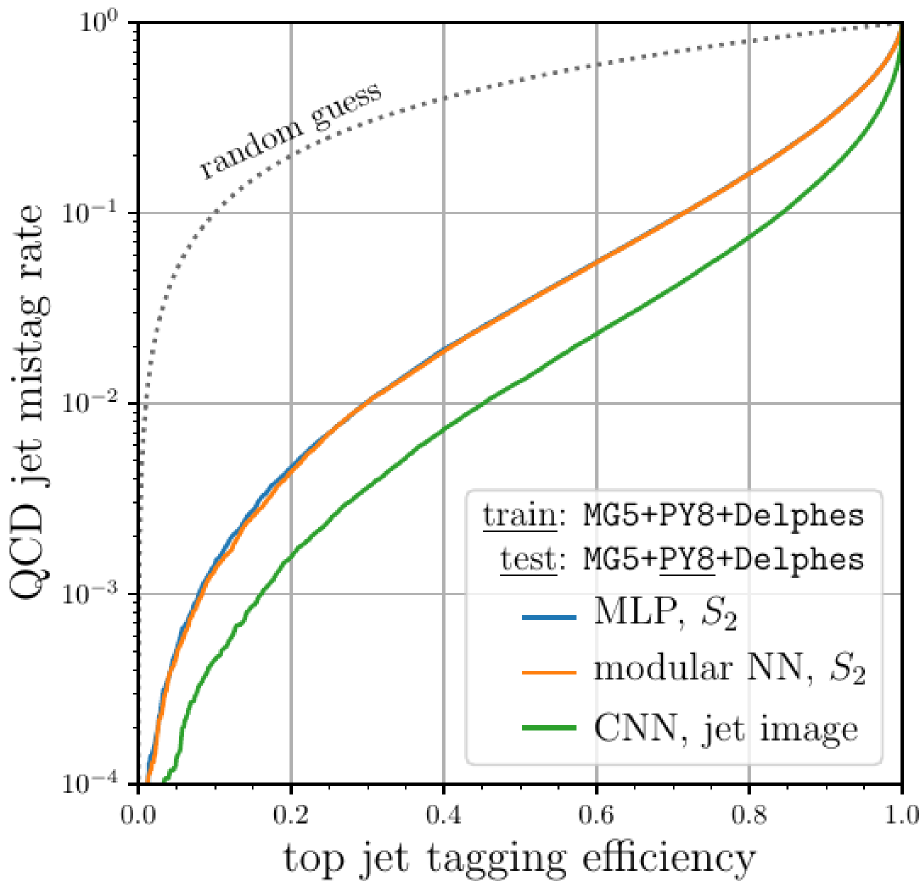
Two prong Structure, Overlap effect Amplified!!

The revised Architecture (Top vs QCD)



- Train this network with "Categorical Cross entropy" loss function
- Output: Top or QCD tag!

Top Jet



- **A Gap observed!**

- CNN is doing better in Background rejection!

- Expected?
  S2 is just 2-point correlations, CNN has more complex pixel Correlations ...

Complete info is missing in S2 Spectra!

AC, Lim, Nojiri and Takeuchi, JHEP 07 111 (2020)

Soft Activity

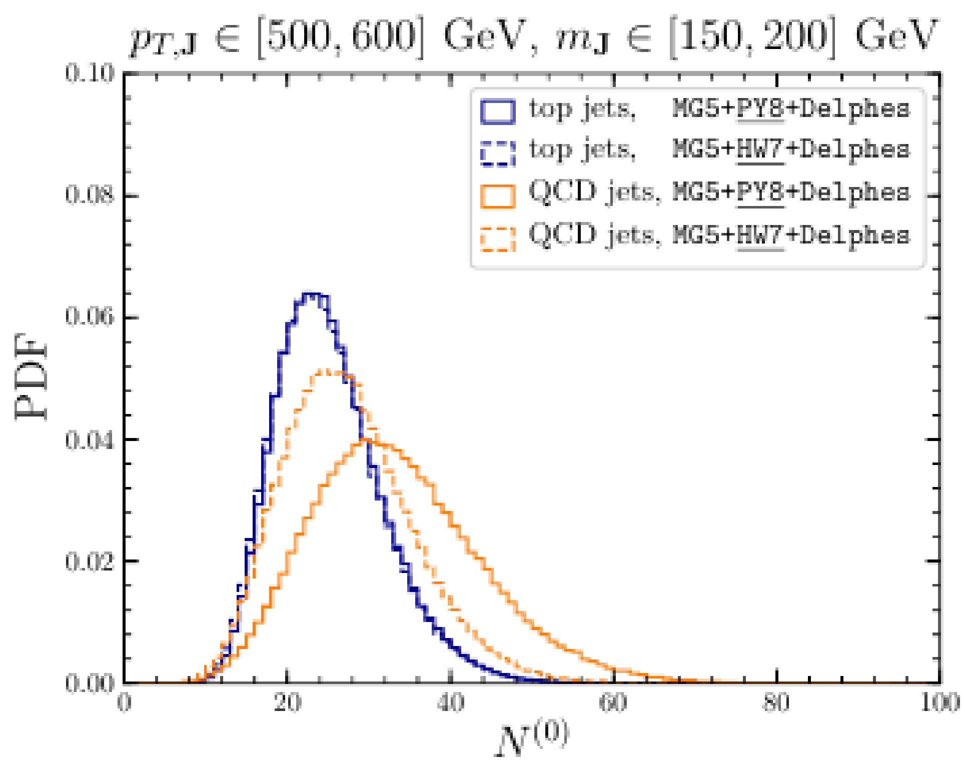
- QCD jets (Quark/gluon jets):
  More soft activity all around the Jet Image

- Higgs Jet:
  Color singlet object, activity mostly centered around the b-jets ...

- Top jet:
  Colored object, more activity than the Higgs,
  will have large angle soft radiations too ...

Can we quantify these effects to see the discrimination power of these soft activities?

Distribution of pixels



- More pixel hits for Gluon jets!

- Similar to Quark/Gluon discrimination (# of charged tracks)

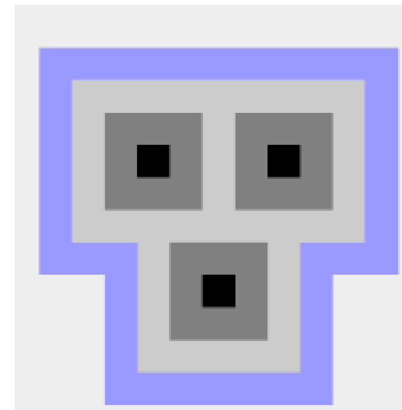
- But, an IRC unsafe quantity! (sensitive to soft/ Colinear splittings!)

- **NOT used directly,** maybe important for classification!

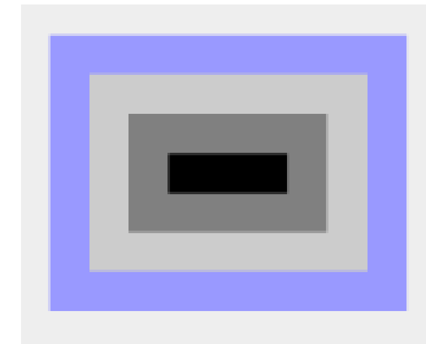
- What about a "geometric description" of the pixel hits?

Geometry of pixel hits

- N_0 : # of active pixels in the Jet
- dN_n : # of pixels surrounding the pixels used in N_{n-1}
- N_n : sum of # of pixels N_0, ..., dN_n



$$N_0 = 3$$
$$N_1 = 9 * N_0$$
$$N_1/N_0 = 9$$



$$N_0 = 3$$
$$N_1 = 3 * N_0 + 6$$
$$N_1/N_0 = 5$$

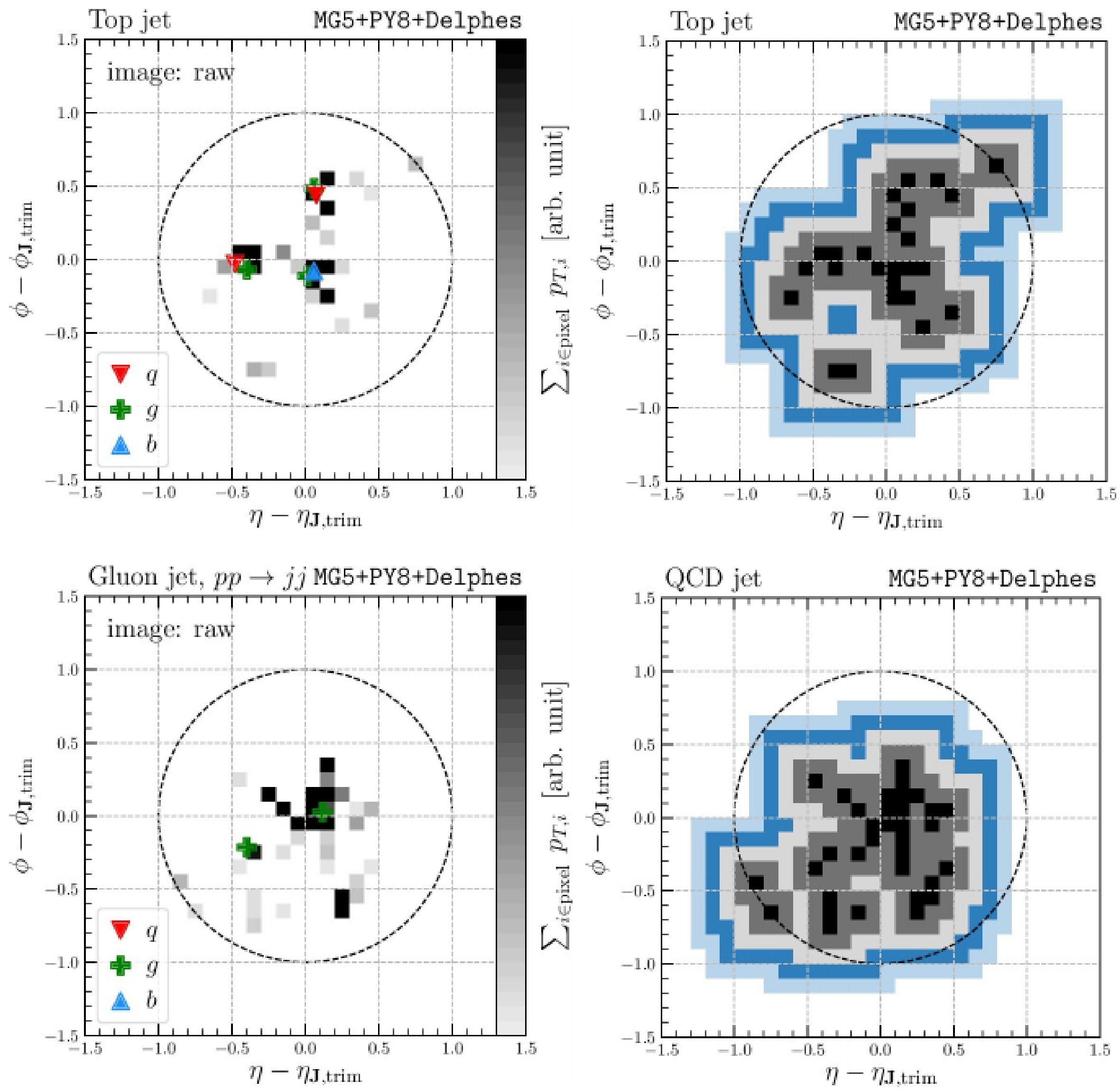
Minkowski Sequence: [Hermann Minkowski et al Mathematische Annalen 57 (1903), 447-495].

A sequence of numbers describing the spatial distribution of pixels!

More connected (isolated) the pixels, Smaller (larger) the ratio!

A notion of the geometrical size of the objects!

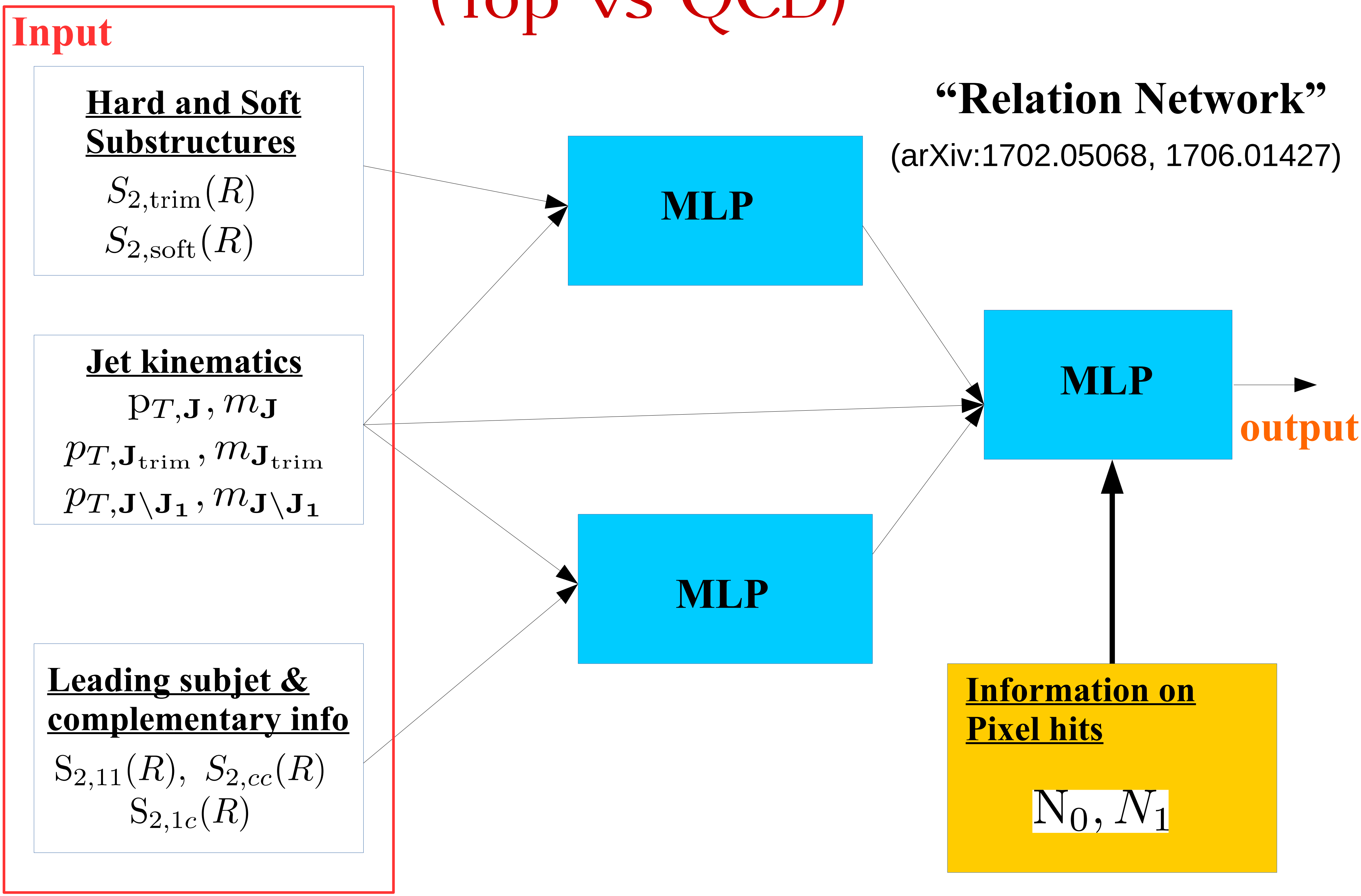
Minkowski Seq for Jet Image



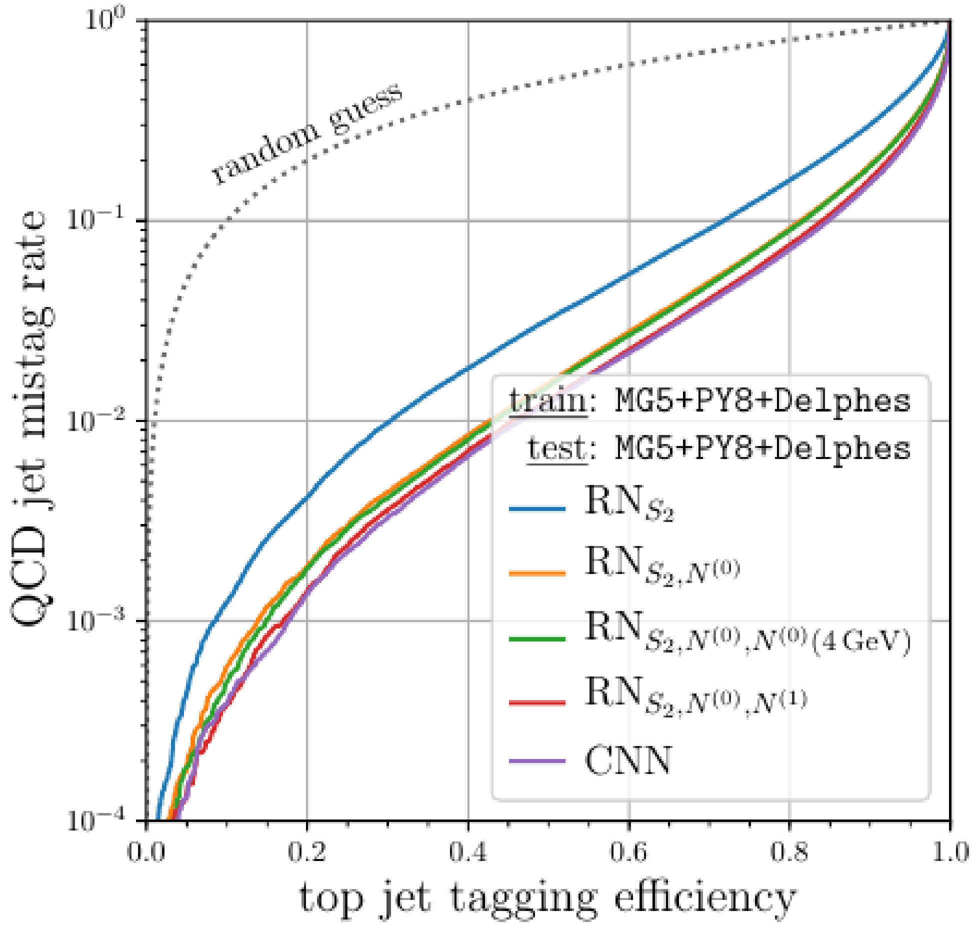
- Lower orders are important, higher terms may not show much difference ...

- We include first two terms, namely N_0 and N_1, as input to Neural Network!

The revised Architecture (Top vs QCD)



Comparable performance to CNN



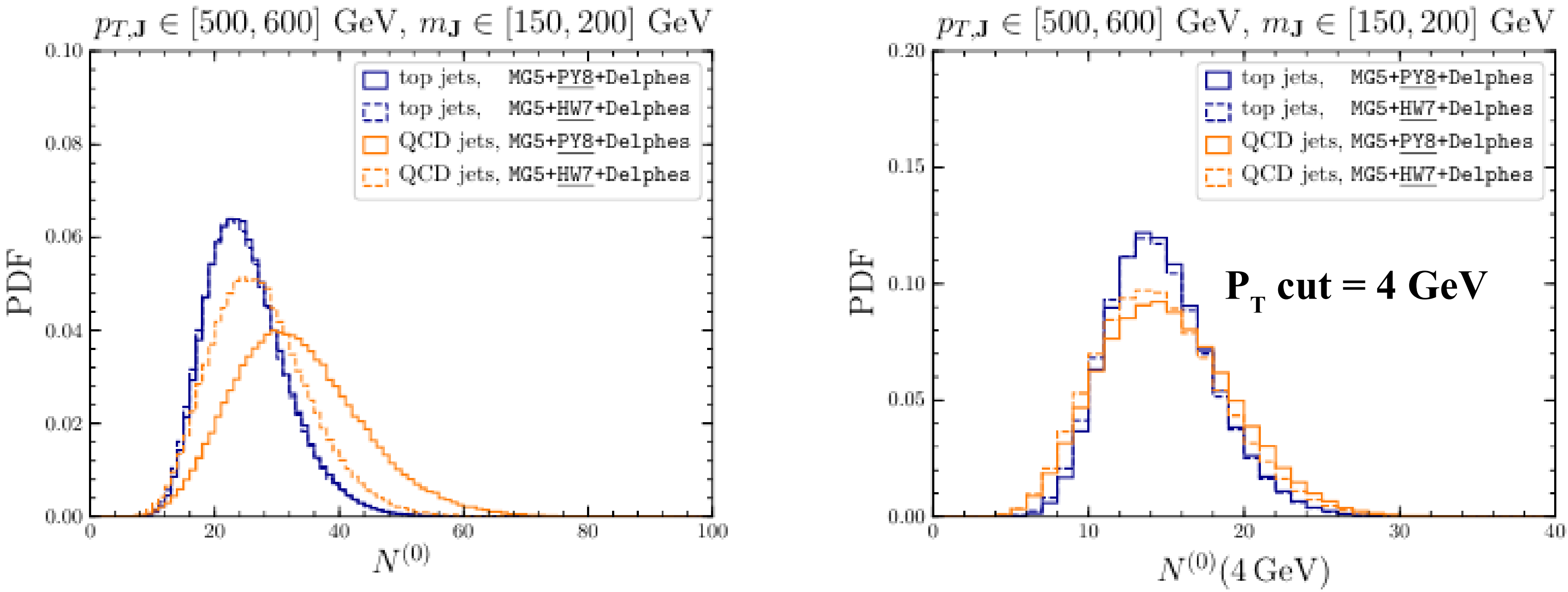
- **The Gap is closed now!!**

- Wider functional space
  Coverage by CNN ...
  Morphology helps to
  Probe these phase spaces ...

- Training is more controled
  (seed variation) than CNN!

AC, Lim, Nojiri and Takeuchi, JHEP 07 111 (2020)

Calibration



- To model CNN, need to estimate S2, N_0 and N_1 distributions properly!

- We compare distributions from two different PSMC (e.g., Pythia vs Herwig)
  Good <u>agreement</u> for "Trimmed" components of S2, but S2 (soft),
  N_0 and N_1 are <u>highly sensitive</u> to PS algorithm, as expected!

- <u>Reweighting</u> performed, Soft distributions (partially) improved, more work needed!

(For q/g case, see Larkoski et al JHEP (2013, 2014), Bhattacherjee et al JHEP (2015) + more)

Summary & Outlook

- **Jet Spectrum**: Higgs and Top tagger based on 2-point energy correlations among the Jet constituents and the geometry of the Soft radiations

- With smaller set of inputs and better controlled training, we obtain classification performance comparable to the CNN

- IRC unsafe plays some significant role, less controlled in Theory, need to tune with experimental data!

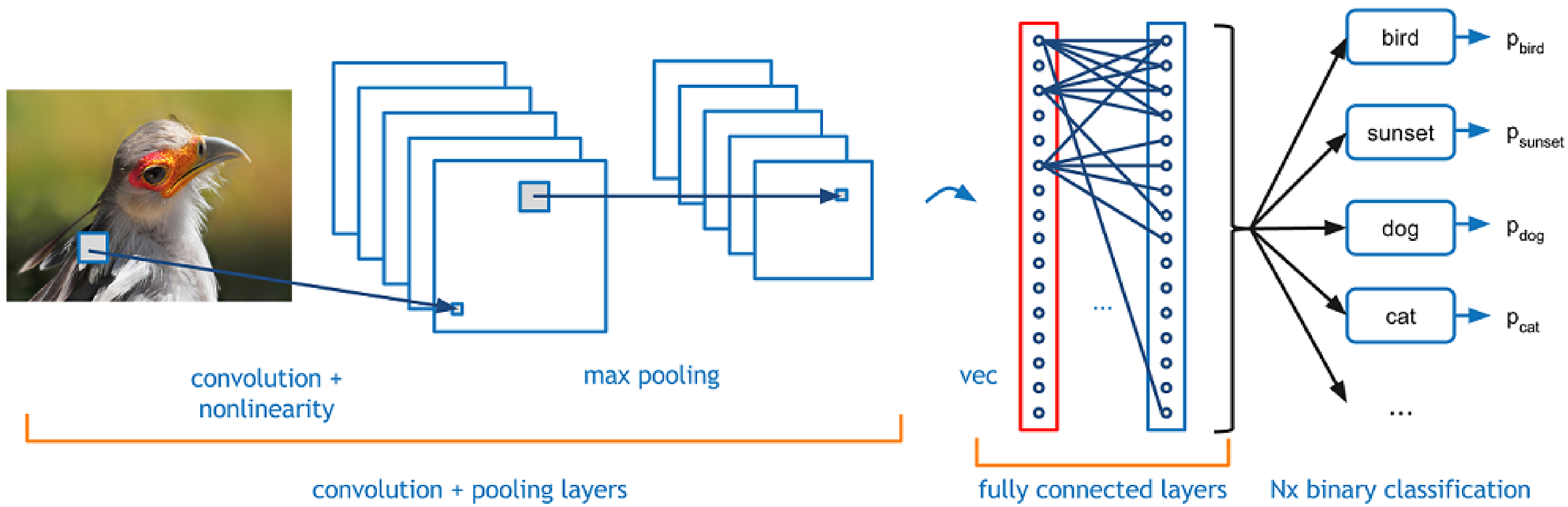
- Time to make use of "<u>Interpretable</u>" Deep Learning frameworks to devise new proposals testable at ongoing/future colliders! Improve & extend traditional taggers for better sensitivity!

- Jet Clustering algorithms need to be revisited and improved, if possible!
  [Ref: AC, Dasmahapatra et. al. 2008.02499, Nachman et. al. 2008.06064]

Thank you!

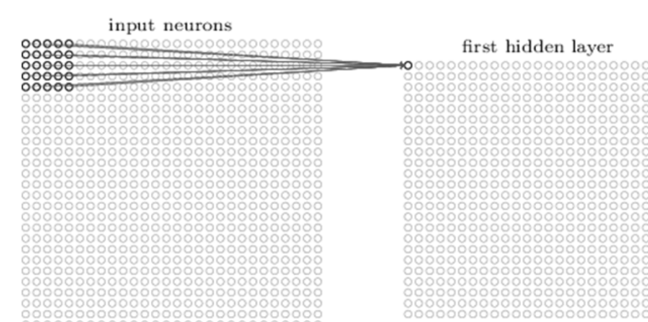
Back ups

Convolutional Neural Network (CNN)



- Large number of free parameters (Hyperparameters) to be optimized

- Computationally very expensive!



Visualization of 5 x 5 filter convolving around an input volume and producing an activation map

[Courtesy to A. Deshpande github]

Higgs jet vs QCD jet

ROC Curve : Signal efficiency Vs Background rejection rates

S2 performs better compared to D2

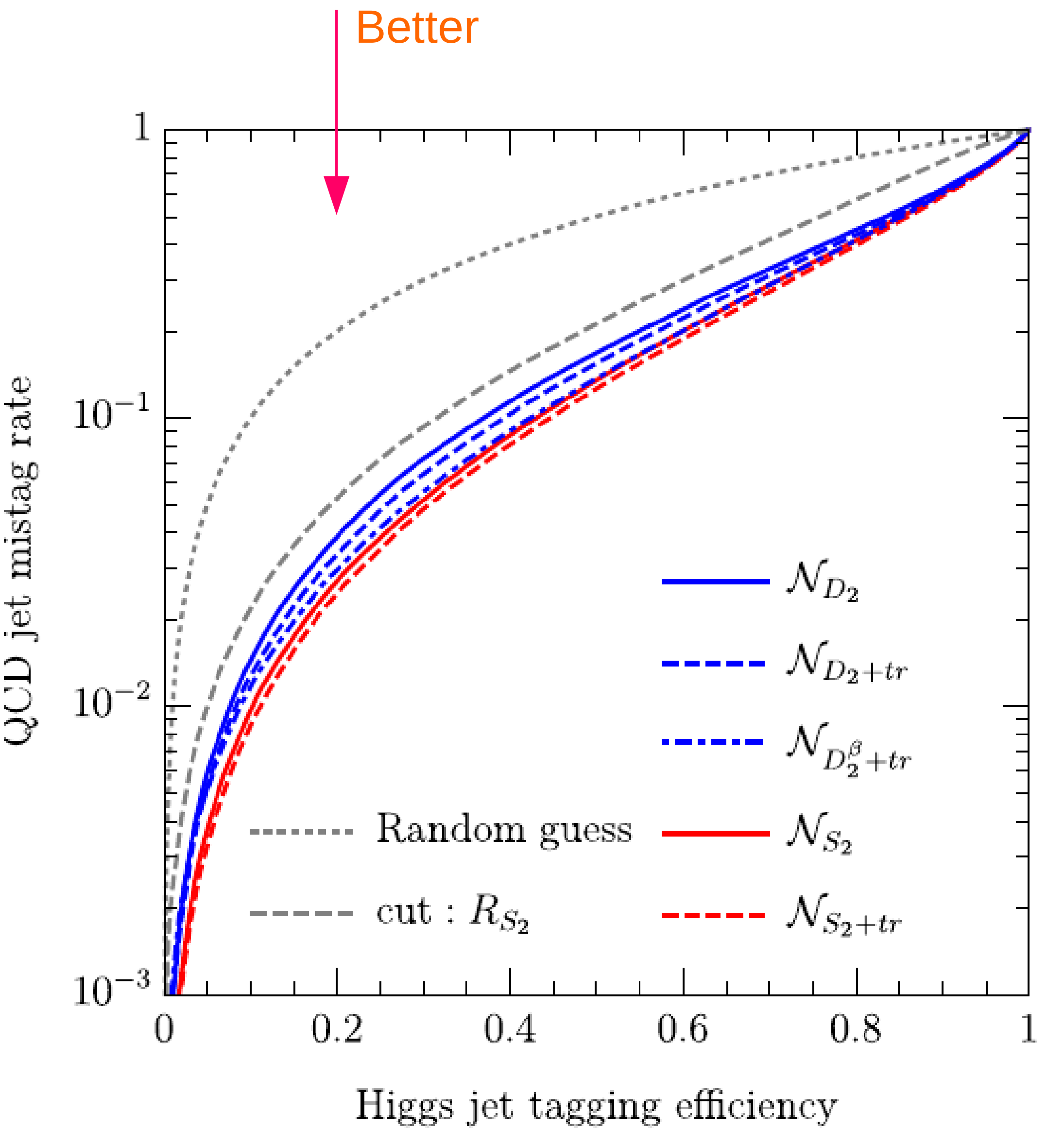
Currently,
D2 default choice for 2-prong object identification!

[Larkoski et. al., JHEP 06, 108 (2013)]

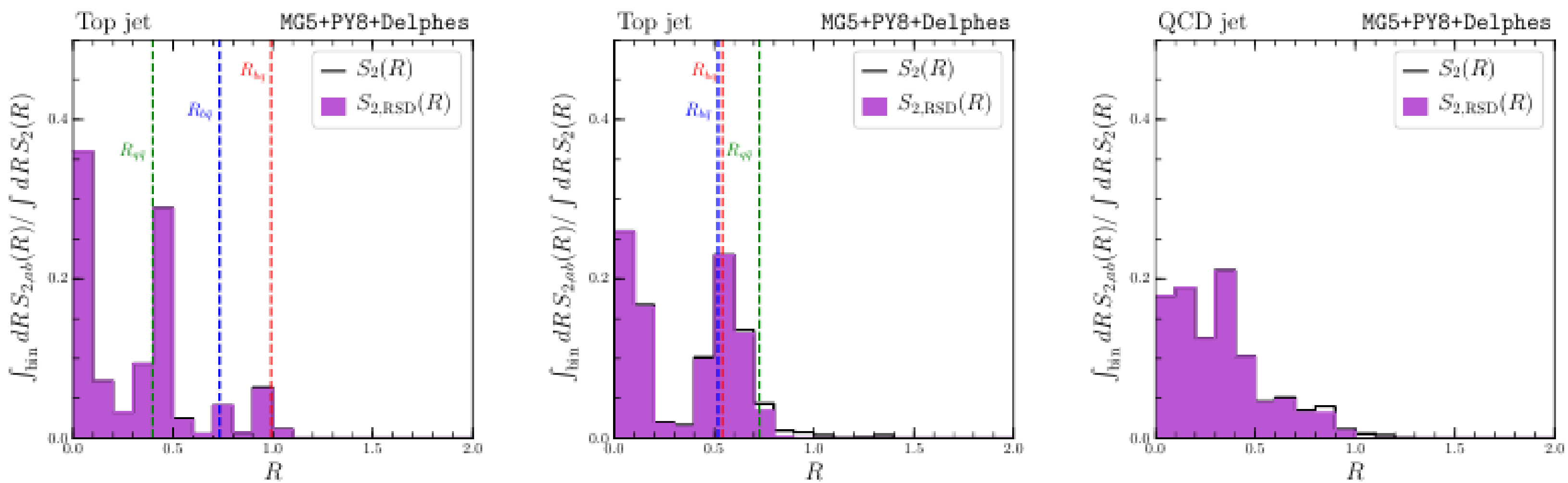
$$e_2^\beta = \frac{1}{p_{T,\mathrm{jet}}^2} \sum_{\substack{i,j\in \mathrm{jet} \\ i<j}} p_{T,i}\, p_{T,j}\, R_{ij}^\beta ,$$

$$e_3^\beta = \frac{1}{p_{T,\mathrm{jet}}^3} \sum_{\substack{i,j,k\in \mathrm{jet} \\ i<j<k}} p_{T,i}\, p_{T,j}\, p_{T,k}\, R_{ij}^\beta R_{jk}^\beta R_{ki}^\beta ,$$

$$D_2^\beta = \frac{e_3^\beta}{(e_2^\beta)^3} \sim \frac{\triangle}{(-)^3},$$



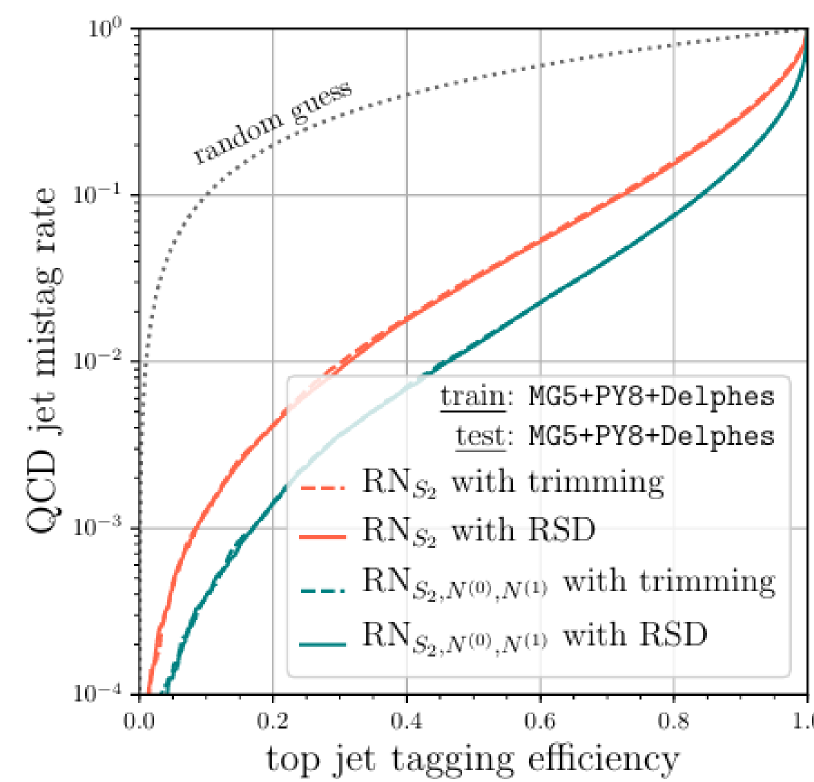
SoftDrop Effect



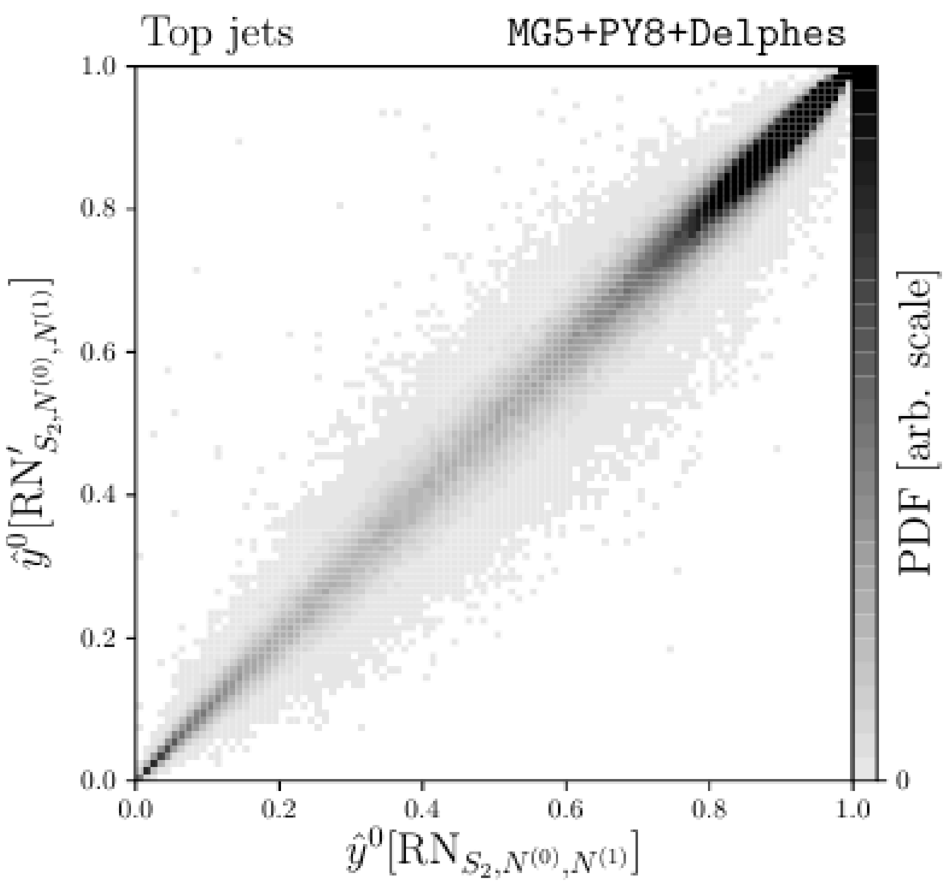
(a)

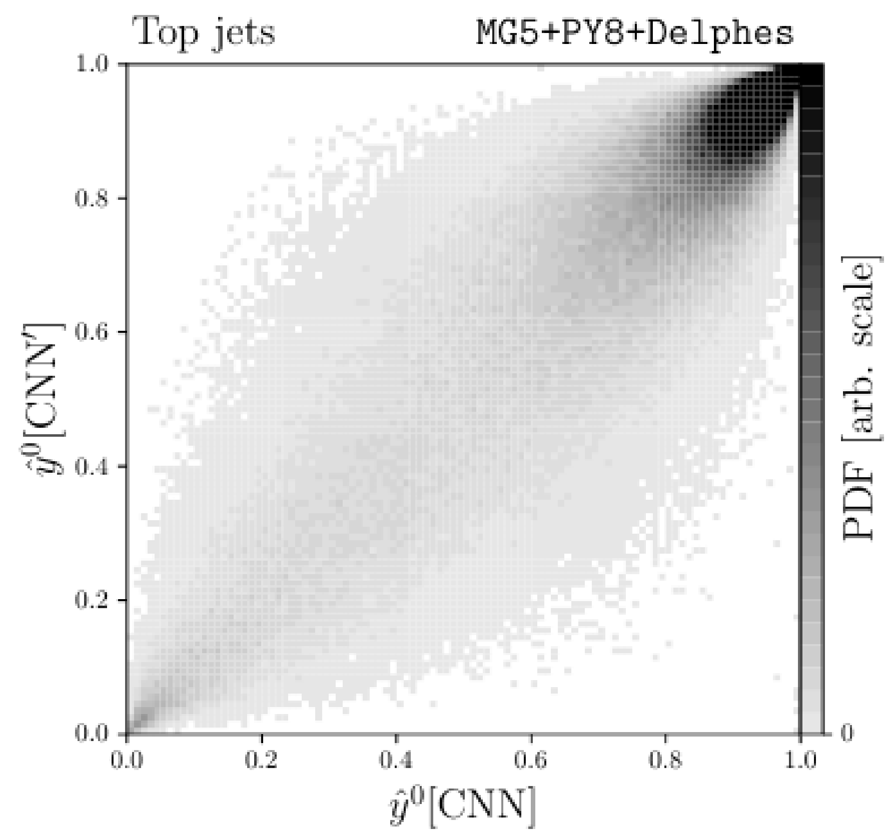
(b)

- The impact on the top jet classification performance due to the change of groomer is small!



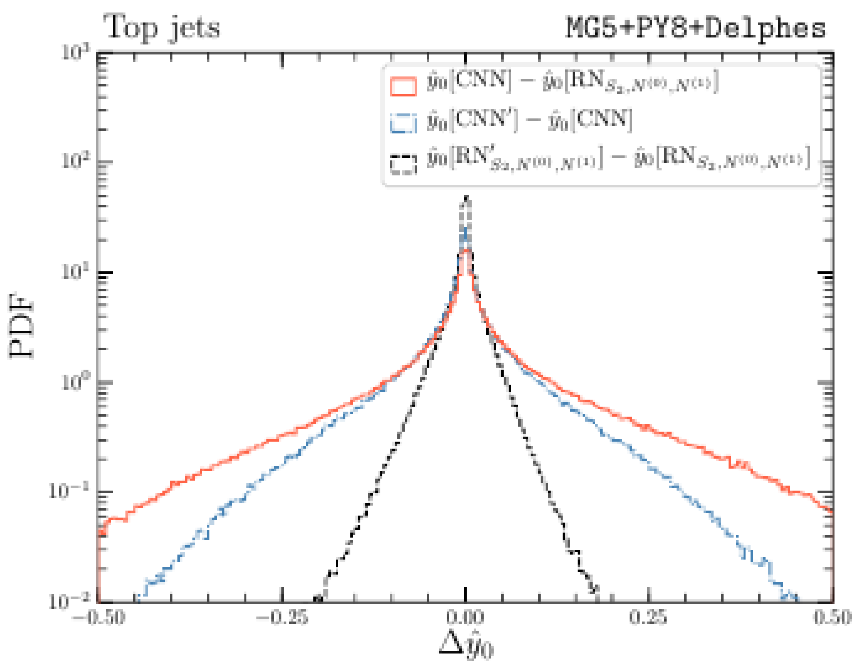
Training uncertainty

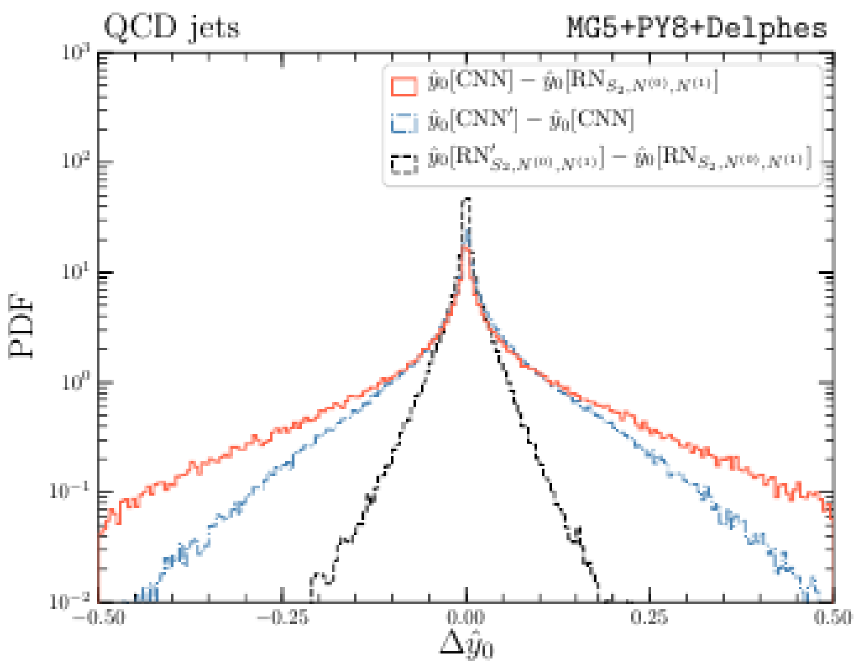




- Variation wrt to Seeds
- CNN has more complexity, so predictions vary widely!
- RN seems more robust under the variation of seeds ...

Training uncertainty

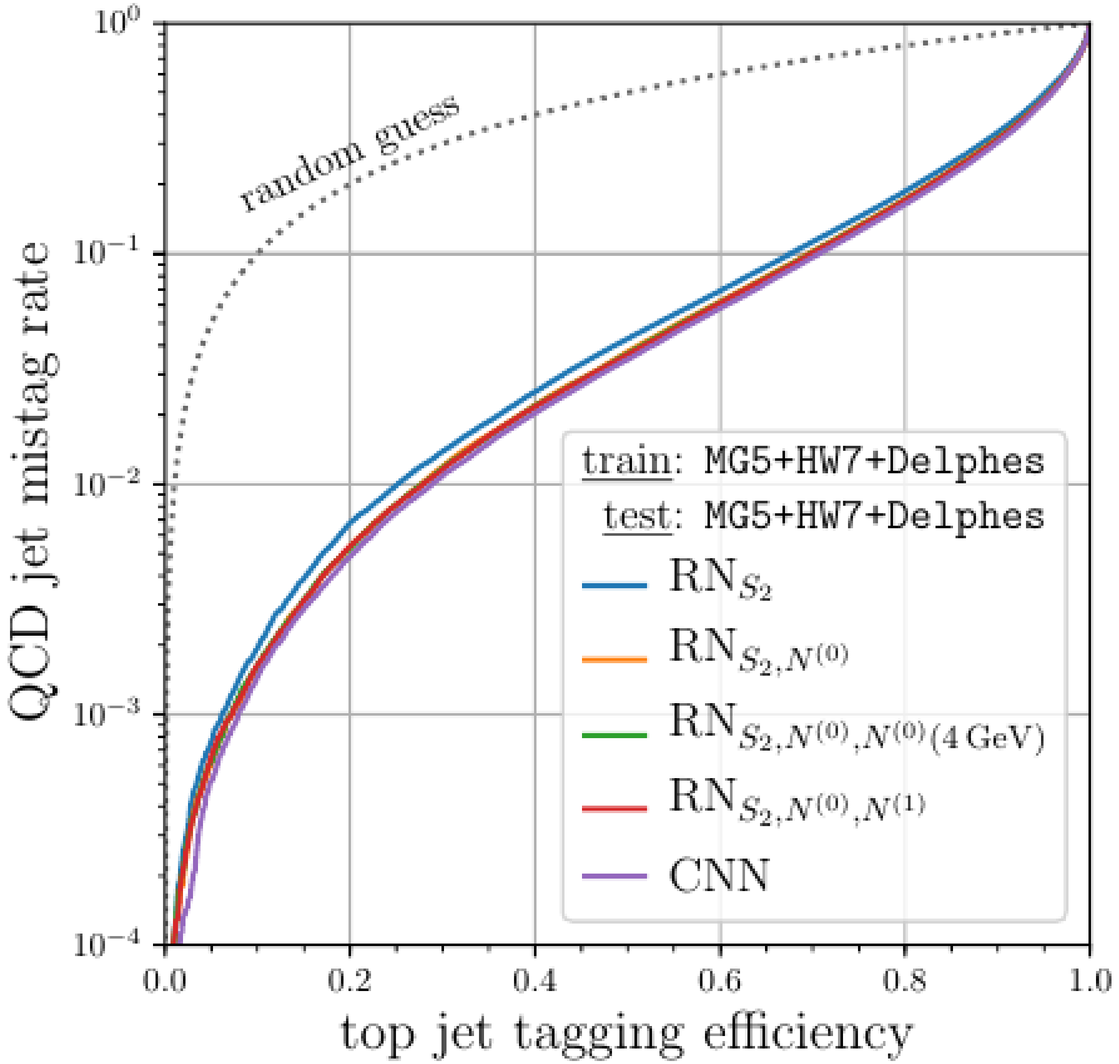




- Variation wrt to Seeds
- CNN has more complexity, so predictions vary widely!
- RN seems more robust under the variation of seeds ...

Herwig samples

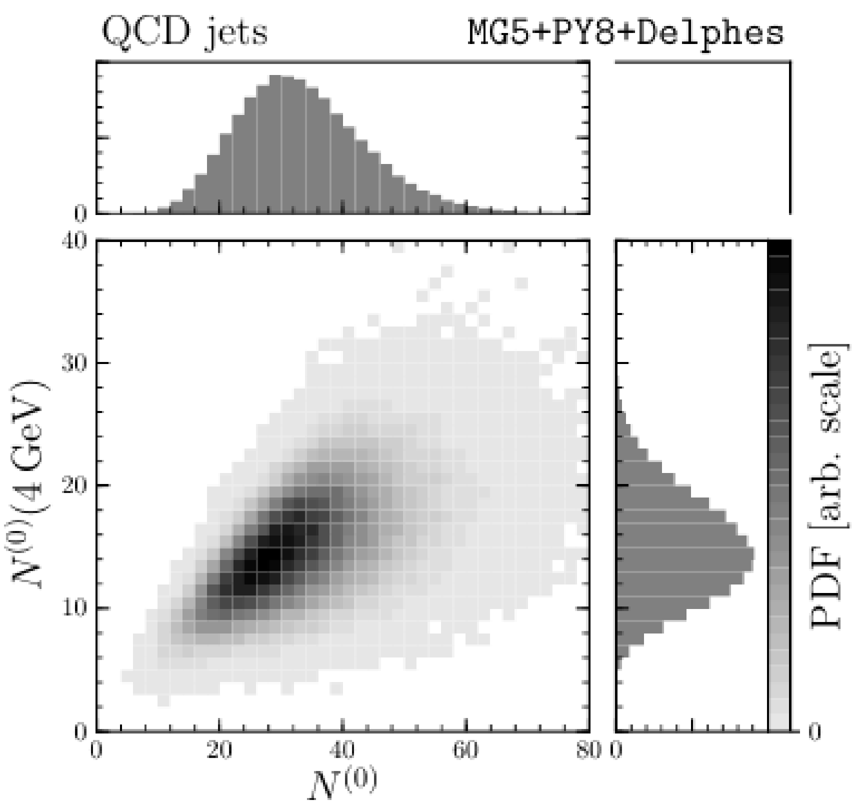
Herwig



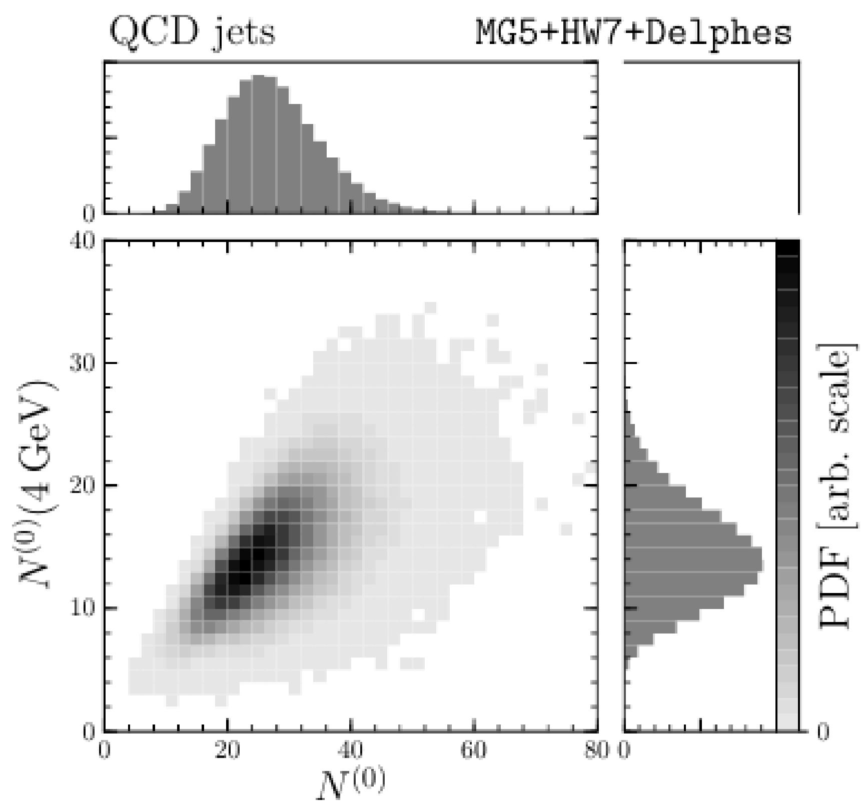
N_0 distribution helps to close the (small) Gap!

Re-weighting

Pythia

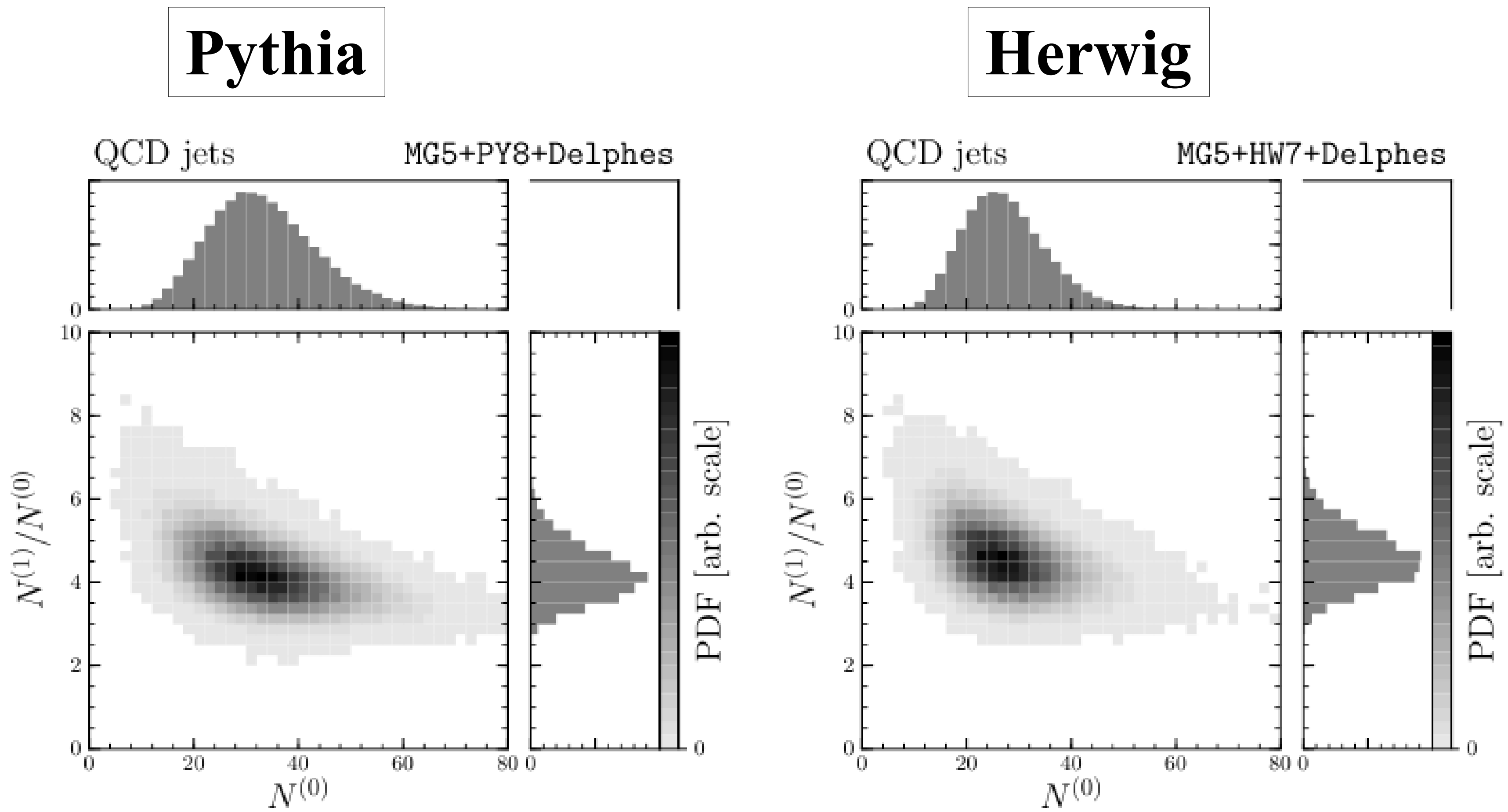


Herwig



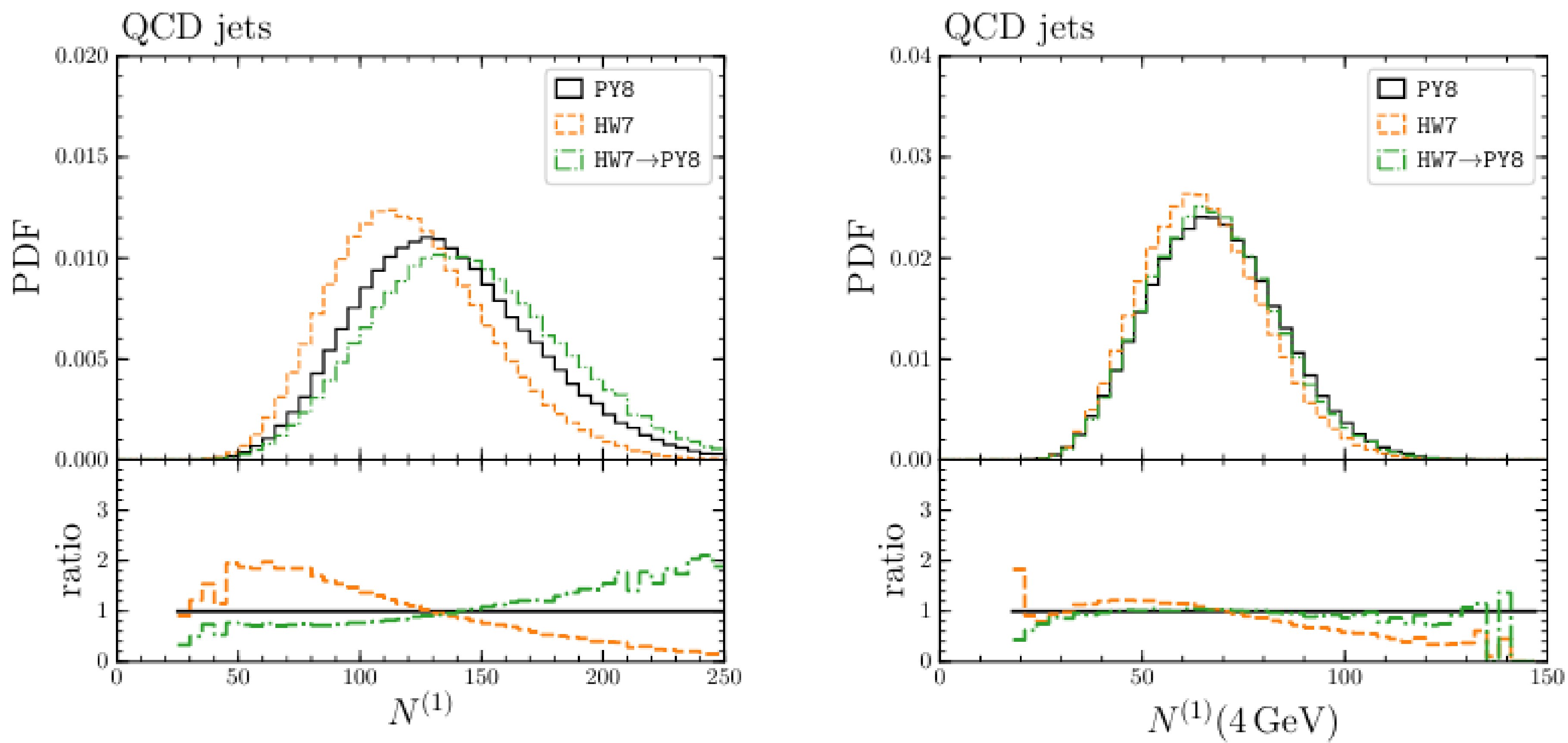
Wider N_0 distribution in PY8, gluons are more radiating ...

Re-weighting



Better control in ratio N_1/N_0 distribution ...

Re-weighting



- The disagreement between PY8 and HW7 remains after the reweighting!

- The difference is large enough to achieve perfect agreement simply by reweighting!

Interpretable Architecture

A general classifier,

$$h_i = \Psi_i[S_{2,A}; \vec{x}_{\text{kin}}],$$

h = input for predictions
A = 1 : "Hard"
A = 2 : "Soft"

Using a Functional Taylor Series Expansion around $S_{2,A}$(R) = 0 gives,

$$h_i = w_i^{(0)}(\vec{x}_{\text{kin}}) + \int dR\, S_{2,A}(R)\, \frac{w_{i,A}^{(2)}(R; \vec{x}_{\text{kin}})}{2} + \frac{1}{2}\int dR_1 dR_2\, S_{2,A}(R_1) S_{2,B}(R_2)\, \frac{w_{i,AB}^{(4)}(R_1, R_2; \vec{x}_{\text{kin}})}{12} + \cdots .$$

Consider the first non-trivial term with S2,

$$h_i = \frac{1}{2}\int dR\, S_{2,A}(R) w_{i,A}^{(2)}(R; \vec{x}_{\text{kin}})$$

In short,

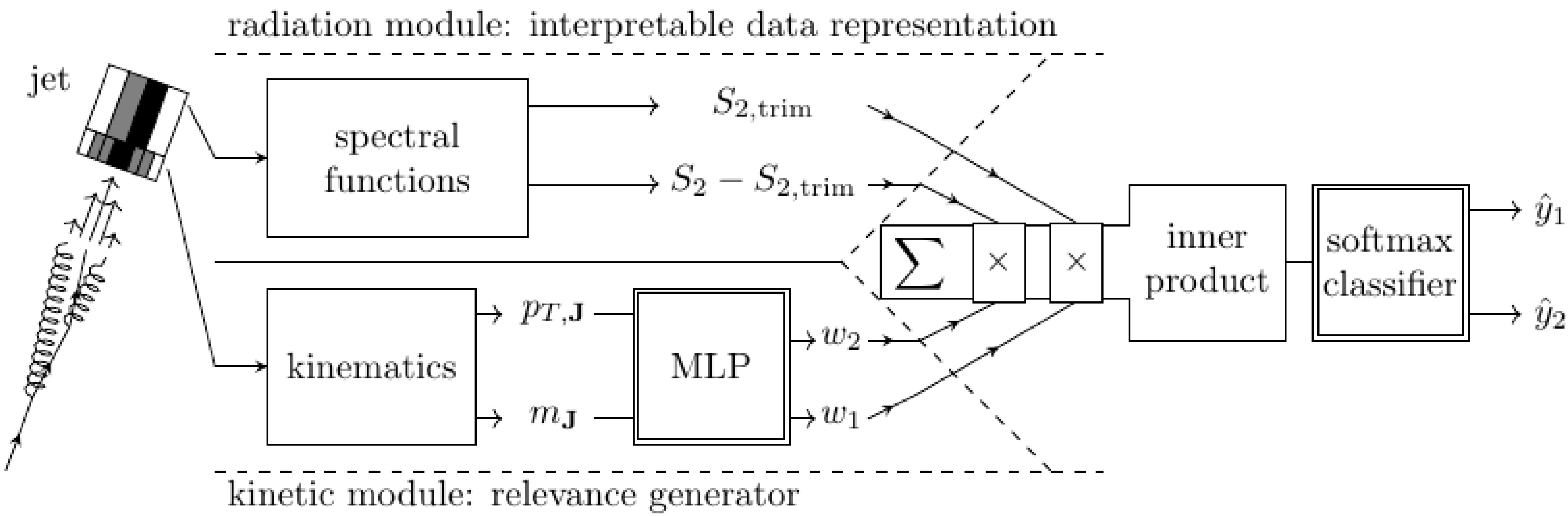
$$h = \sum_k S_{2,\text{trim}}^k\, w_1^k + \sum_k S_{2,\text{soft}}^k\, w_2^k,$$

$$\hat{y}_i = \exp[w_i^{(\text{out})}\, h] \Big/ \sum_i \exp[w_i^{(\text{out})}\, h],$$

Read of the "weights" to get the correlation between the Weights and S2

- Interpretability!

Interpretable Architecture



- An MLP trained on pT and mass of the jet, generates the weights w1 and w2 (MLP has 3 hidden layers with nodes 400, 100 and 40 respectively!)

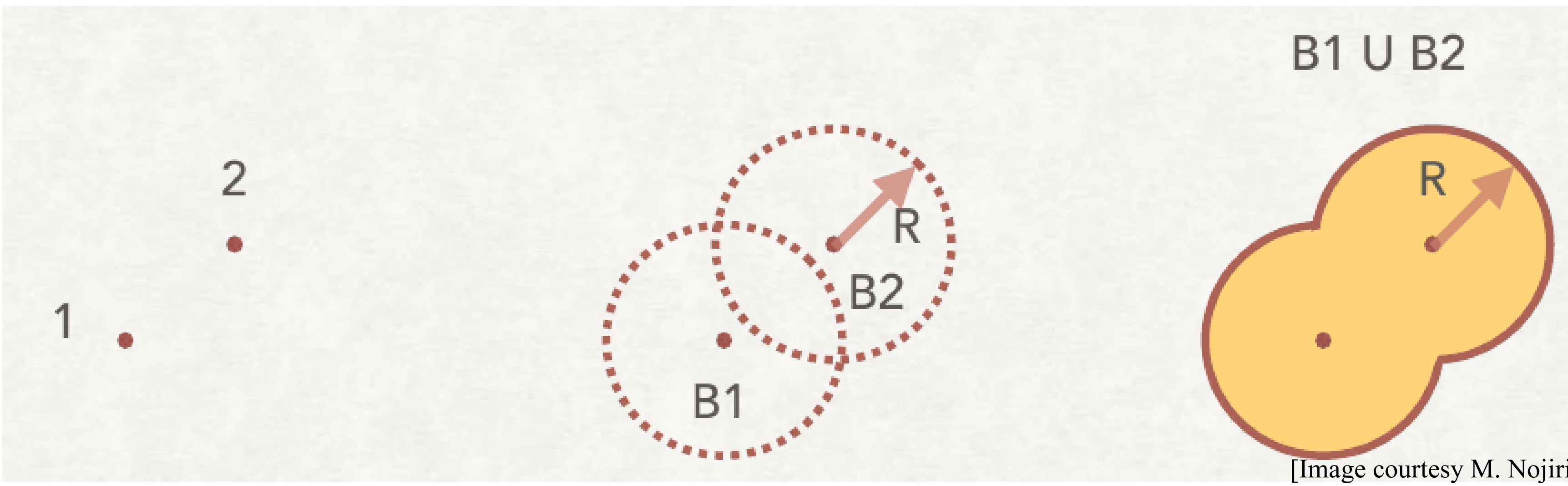
- “Softmax” classifier combines the “Radiation module” with weights!

- Performance of the classifier depends on the “correlation” of “weights” and “S2 spectra!

Minkowski Sequence/Functional

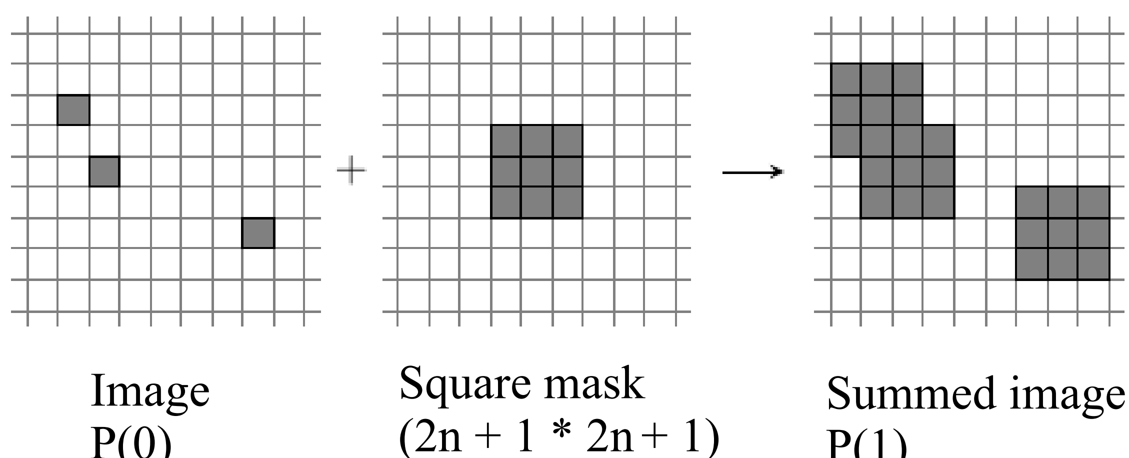
In Mathematics:

A notion of the geometrical size of the objects



[Image courtesy M. Nojiri]

Count the number of pixels in the Summed image



[In Cosmology:
Schmalzing et al, astro-ph/9508154]

MLP architecture

The relation networks used in this paper are implemented as follows. The module for analyzing the energy correlation with jet trimming, $\boldsymbol{h}_{\text{trim}} = \text{MLP}_{\text{trim}}(\boldsymbol{x}_{\text{trim}}, \boldsymbol{x}_{\text{kin}})$, consists of two hidden layers,

$$\begin{aligned}
\boldsymbol{h}^{(1)}_{\text{trim}} &= \text{FC}(\boldsymbol{z}_{\text{trim}}, \boldsymbol{z}_{\text{kin}}), && \text{size: } 200, && \text{activation: ELU} \\
\boldsymbol{h}^{(2)}_{\text{trim}} &= \text{FC}(\boldsymbol{h}^{(1)}_{\text{trim}}), && \text{size: } 200, && \text{activation: ELU} \\
\boldsymbol{h}_{\text{trim}} &= \text{FC}(\boldsymbol{h}^{(2)}_{\text{trim}}), && \text{size: } 5, && \text{activation: linear}
\end{aligned} \tag{C.1}$$

where $\boldsymbol{z}_i$ is the standardized inputs of $\boldsymbol{x}_i$, and FC is a fully-connected layer with a given output size and activation function. Note that we do not apply L_2 regularization for the FCs with linear activation. The module for analyzing the energy correlation of $\mathbf{J}_1$ and $\mathbf{J} \setminus \mathbf{J}_1$ is as follows.

$$\begin{aligned}
\boldsymbol{h}^{(1)}_{\mathbf{J}_1} &= \text{FC}(\boldsymbol{z}_{\mathbf{J}_1}, \boldsymbol{z}_{\text{kin}}), && \text{size: } 200, && \text{activation: ELU} \\
\boldsymbol{h}^{(2)}_{\mathbf{J}_1} &= \text{FC}(\boldsymbol{h}^{(1)}_{\mathbf{J}_1}), && \text{size: } 200, && \text{activation: ELU} \\
\boldsymbol{h}_{\mathbf{J}_1} &= \text{FC}(\boldsymbol{h}^{(2)}_{\mathbf{J}_1}), && \text{size: } 5, && \text{activation: linear}
\end{aligned} \tag{C.2}$$

The logits $\boldsymbol{u}'$ for the binary classification is implemented as follows.

$$\begin{aligned}
\boldsymbol{h}^{(1)}_{\text{logit}} &= \text{FC}(\boldsymbol{h}_{\text{trim}}, \boldsymbol{h}_{\mathbf{J}_1}, \boldsymbol{z}_{\text{kin}}), && \text{size: } 200, && \text{activation: ELU} \\
\boldsymbol{h}^{(2)}_{\text{logit}} &= \text{FC}(\boldsymbol{h}^{(1)}_{\text{logit}}), && \text{size: } 200, && \text{activation: ELU} \\
\boldsymbol{u}' &= \text{FC}(\boldsymbol{h}^{(2)}_{\text{logit}}), && \text{size: } 2, && \text{activation: linear}
\end{aligned} \tag{C.3}$$

For the relation networks with inputs $\boldsymbol{x}_{\text{geometry}}$, we replace $\boldsymbol{h}^{(1)}_{\text{logit}}$ of eq. (C.3) as follows.

$$\boldsymbol{h}^{(1)}_{\text{logit}} = \text{FC}(\boldsymbol{h}_{\text{trim}}, \boldsymbol{h}_{\mathbf{J}_1}, \boldsymbol{z}_{\text{geometry}}), \quad \text{size: } 200, \quad \text{activation: ELU}, \tag{C.4}$$

CNN architecture

The vanilla CNN of this paper consists of six convolutional layers with a filter size 3×3. The standardized image $\boldsymbol{z}_{\mathrm{image}}$ of $\boldsymbol{x}_{\mathrm{image}}$ is fed into a chain of convolutional layers as follows.

$$
\begin{aligned}
\boldsymbol{h}_{\mathrm{CNN}}^{(1)} &= \mathrm{CONV}(\boldsymbol{z}_{\mathrm{image}}), && \text{size: } 30 \times 30 \times 16, && \text{filter size: } 3 \times 3, && \text{activation: ELU}, \\
\boldsymbol{h}_{\mathrm{CNN}}^{(2)} &= \mathrm{CONV}(\boldsymbol{h}_{\mathrm{CNN}}^{(1)}), && \text{size: } 30 \times 30 \times 16, && \text{filter size: } 3 \times 3, && \text{activation: ELU}, \\
\boldsymbol{h}_{\mathrm{CNN}}^{(3)} &= \mathrm{CONV}(\boldsymbol{h}_{\mathrm{CNN}}^{(2)}), && \text{size: } 30 \times 30 \times 16, && \text{filter size: } 3 \times 3, && \text{activation: ELU}, \\
\boldsymbol{h}_{\mathrm{CNN}}^{(3,\mathrm{POOL})} &= \mathrm{POOL}(\boldsymbol{h}_{\mathrm{CNN}}^{(3)}), && \text{size: } 15 \times 15 \times 16, && \text{pool size: } 2 \times 2, && \\
\boldsymbol{h}_{\mathrm{CNN}}^{(4)} &= \mathrm{CONV}(\boldsymbol{h}_{\mathrm{CNN}}^{(3,\mathrm{POOL})}), && \text{size: } 15 \times 15 \times 8, && \text{filter size: } 3 \times 3, && \text{activation: ELU}, \\
\boldsymbol{h}_{\mathrm{CNN}}^{(5)} &= \mathrm{CONV}(\boldsymbol{h}_{\mathrm{CNN}}^{(4)}), && \text{size: } 15 \times 15 \times 8, && \text{filter size: } 3 \times 3, && \text{activation: ELU}, \\
\boldsymbol{h}_{\mathrm{CNN}}^{(6)} &= \mathrm{CONV}(\boldsymbol{h}_{\mathrm{CNN}}^{(5)}), && \text{size: } 15 \times 15 \times 8, && \text{filter size: } 3 \times 3, && \text{activation: ELU}, \\
\boldsymbol{h}_{\mathrm{CNN}}^{(6,\mathrm{POOL})} &= \mathrm{POOL}(\boldsymbol{h}_{\mathrm{CNN}}^{(6)}), && \text{size: } 7 \times 7 \times 8, && \text{pool size: } 2 \times 2, && \\
\boldsymbol{h}_{\mathrm{CNN}}^{(7)} &= \mathrm{FC}(\boldsymbol{h}_{\mathrm{CNN}}^{(6,\mathrm{POOL})}), && \text{size: } 200, && \text{activation: ELU}, && \\
\boldsymbol{h}_{\mathrm{CNN}} &= \mathrm{FC}(\boldsymbol{h}_{\mathrm{CNN}}^{(7)}), && \text{size: } 100, && \text{activation: linear}, &&
\end{aligned} \tag{C.5}
$$

where CONV is a two-dimensional convolutional layer with a given filter size and activation function, and POOL is a max-pooling layer with a given pool size. The output size consists of three numbers: the first two numbers represent output image width and height, and the third number is the number of filters. We simply put $\boldsymbol{h}_{\mathrm{CNN}}$ to $\mathrm{MLP}_{\mathrm{logit}}$ by replacing eq. (C.3) to the following.

$$
\boldsymbol{h}_{\mathrm{logit}}^{(1)} = \mathrm{FC}(\boldsymbol{h}_{\mathrm{CNN}}, \boldsymbol{z}_{\mathrm{kin}}), \qquad \text{size: } 200, \quad \text{activation: ELU} \tag{C.6}
$$