Published for SISSA by Springer

Received: *April 29, 2019*
Accepted: *July 11, 2019*
Published: *July 23, 2019*

Interpretable deep learning for two-prong jet classification with jet spectra

Amit Chakraborty,[a] **Sung Hak Lim**[a] **and Mihoko M. Nojiri**[a,b,c]

[a] *Theory Center, IPNS, KEK,*
1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan
[b] *The Graduate University of Advanced Studies (Sokendai),*
1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan
[c] *Kavli IPMU (WPI), University of Tokyo,*
5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8583, Japan

E-mail: amit@post.kek.jp, sunghak.lim@kek.jp, nojiri@post.kek.jp

Abstract: Classification of jets with deep learning has gained significant attention in recent times. However, the performance of deep neural networks is often achieved at the cost of interpretability. Here we propose an interpretable network trained on the jet spectrum $S_2(R)$ which is a two-point correlation function of the jet constituents. The spectrum can be derived from a functional Taylor series of an arbitrary jet classifier function of energy flows. An interpretable network can be obtained by truncating the series. The intermediate feature of the network is an infrared and collinear safe C-correlator which allows us to estimate the importance of an $S_2(R)$ deposit at an angular scale R in the classification. The performance of the architecture is comparable to that of a convolutional neural network (CNN) trained on jet images, although the number of inputs and complexity of the architecture is significantly simpler than the CNN classifier. We consider two examples: one is the classification of two-prong jets which differ in color charge of the mother particle, and the other is a comparison between `Pythia 8` and `Herwig 7` generated jets.

Keywords: Jets, QCD Phenomenology

ArXiv ePrint: 1904.02092

https://doi.org/10.1007/JHEP07(2019)135

JHEP07(2019)135

Contents

1 Introduction **1**

2 Two-point correlation spectrum and two-prong jets **3**
- 2.1 Jet spectra 3
- 2.2 Derivation of classifiers based on energy flows and jet spectra 6
- 2.3 Relation between two-point correlation spectra and energy flow polynomials 7
- 2.4 Spectra of two-prong jets 8

3 Classifying Higgs jet, sgluon jet, and QCD jet **11**
- 3.1 Basic kinematics 11
- 3.2 Multilayer perceptron of spectra 13
- 3.3 Event generator dependence 16

4 Interpretable two-level architecture **17**

5 Summary and outlook **24**

A Event generation and reconstruction **27**

B Oversampling and $p_{T,\mathrm{J}}$-bias removal **28**

C Jet image and convolution neural network **28**

JHEP07(2019)135

1 Introduction

Deep learning is gaining significant interest recently in the field of collider data analysis. One of the primary motivations is to extract the maximum information from the complex collision events. The deep learning in collider physics takes advantage of a large influx of data from experiments, more precise theoretical predictions, significant improvement in computing power, and ongoing progress in the field of machine learning itself. Such techniques offer advances in areas ranging from event selection to particle identification.

The large center-of-mass energy at the Large Hadron Collider (LHC) enables the production of boosted particles whose decay products are highly collimated. These collimated objects are reconstructed as a jet, and it is often misidentified as a QCD jet originated from light quarks or gluons. Many jet substructure techniques using the information of subjets [1–10] and the distribution of jet constituents [11–17] have been developed in order to improve the sensitivity of tagging and to classify these boosted particle jets. The deep learning methods [18–34] have provided useful insight into the internal structure of the

jets and, thereby, shown better performances than those jet substructure techniques.[1] The flexibility of deep learning also enables us to solve problems beyond supervised classifications, such as weakly supervised learning [37–39], adversarial learning to suppress learning from unwanted information [40, 41], and unsupervised learning for finding anomalous signatures [42–47]. The neural network can also be useful to new physics searches with deep learning at the LHC [48–56].

The output of a neural network is, in general, a highly non-linear function of the inputs. A neural network classifier often acts like a "black box". One may consider architectures with post-hoc interpretability [57], which allows us to extract information other than its prediction from the learned model after training. A simple strategy is using a predefined functional form to restrict the representation power of the neural network [31, 58]. Then the network is interpreted in terms of the functional form. The aim of this paper is also to construct an interpretable neural network architecture that allows us not only to interpret the predictions of the network but also to visualize it in terms of trained weights connected to physical variables.

In [28], a multilayer perceptron (MLP) trained on two-point correlation functions S_2 and $S_{2,\mathrm{trim}}$ of angular scale R was introduced. The $S_2(R)$ and $S_{2,\mathrm{trim}}(R)$ spectra are constructed from the constituents of a jet before and after the trimming [59] respectively. The angular scale R is an important parameter for describing the kinematics of a decaying particle and parton shower (PS); hence, these spectra efficiently encode the radiation pattern inside a jet. The MLP trained on these inputs learns relevant features for the classification among the Higgs boson jet (Higgs jet) and QCD jet.

In this paper, we connect the spectra to energy flow functionals $P_T(\vec{R})$ [60], i.e., we consider transverse energy of a jet constituent as particle-specific information at $\vec{R}$ in the $\eta - \phi$ plane [61]. The spectra are basis vectors of infrared and collinear (IRC) safe variables called bilinear C-correlators [60] whose angular weighting function depends only on the relative distance between two constituents. Those correlators naturally appear in the functional Taylor series of a classifier of $P_T(\vec{R})$, and the MLP can be considered as a subseries of the Taylor series. We show that the performance of the MLP and neural networks trained on jet images [18, 19, 21, 62] are comparable. This strongly suggests that S_2 and $S_{2,\mathrm{trim}}$ contain sufficient information for jet classification. Encouraged by this feature, we construct an interpretable architecture by truncating the series. Namely, $\int dR\, S_2(R) w(R; \vec{x}_{\mathrm{kin}})$ can be implemented in a classifier after proper discretization in R, where $\vec{x}_{\mathrm{kin}}$ is a set of kinematic variables of the jet and w is a smooth function. By reading the weights $w(R; \vec{x}_{\mathrm{kin}})$, we could quantify important features for the given classification problem.

Jet substructure studies often suffer from systematic uncertainties of soft activities. The soft radiations generated by a Monte Carlo program are strongly model dependent. While this mismodeling could be corrected by using real data, it is certainly useful to use input variables with less systematic uncertainties. When hard substructures are important

[1] For a review on the recent theoretical and machine learning developments in jet substructure techniques at the LHC, we refer [35, 36].

JHEP07(2019)135

for solving the problem, we may use jet grooming techniques [1, 10, 59, 63, 64] to remove the soft activity. Instead of throwing this soft activity away, we encode it in $S_{2,\text{soft}}(R)$, which is $S_2(R) - S_{2,\text{trim}}(R)$. Then, the inputs $S_{2,\text{trim}}$ and $S_{2,\text{soft}}$ include hard and soft substructure information, respectively. The interpretable architecture trained on $S_{2,\text{trim}}$ and $S_{2,\text{soft}}$ is able to quantify these features. We study two classification problems: one is a classification of two-prong jets to understand their hard substructures and color coherence, and the other is a comparison of `Pythia 8` [65] and `Herwig 7` [66, 67] events to quantify the differences.

The paper is organized as follows. In section 2, we review S_2 and $S_{2,\text{trim}}$ and show its relation to energy flow and C-correlators. We also show S_2 and $S_{2,\text{trim}}$ distributions of typical Higgs jet and QCD jet. A hypothetical color octet scalar particle, sgluon, decaying to $b\bar{b}$ is considered to study the color connection in two-prong jets. In section 3, we first discuss the capability of S_2 and $S_{2,\text{trim}}$ for the classification of two-prong jets and show the result of an MLP trained on those inputs. The results are then compared with that of a CNN trained on jet images. In section 4, we introduce a two-level architecture consists of a softmax classifier and an MLP trained on S_2 and $S_{2,\text{trim}}$. The intermediate feature of this architecture is the bilinear C-correlator whose basis vectors are $S_{2,\text{trim}}$ and $S_{2,\text{soft}}$, and the MLP generates its components. We visualize and interpret the weights of the given classification problem. Finally, the summary and outlook are given in section 5.

2 Two-point correlation spectrum and two-prong jets

2.1 Jet spectra

In [28], we introduced a two-point correlation spectral function $S_2(R)$ which maps a jet to a function of angular scale R,

$$S_2(R) = \int d\vec{R}_1\, d\vec{R}_2\, P_T(\vec{R}_1)\, P_T(\vec{R}_2)\, \delta(R - R_{12}), \tag{2.1}$$

$$P_T(\vec{R}) = \sum_{i \in \mathbf{J}} p_{T,i}\, \delta(\vec{R} - \vec{R}_i), \tag{2.2}$$

where $\mathbf{J}$ is a set of jet constituents, $\vec{R}_i = (\eta_i, \phi_i)$ is the position of the i-th jet constituent in the pseudorapidity-azimuth plane, $R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$ is the angular distance between the two jet constituents i and j, and $P_T(\vec{R})$ is an energy flow functional [60] of $\mathbf{J}$. For practical purpose, $S_2(R)$ is discretized as below,

$$\begin{aligned} S_2(R; \Delta R) &= \frac{1}{\Delta R} \int_R^{R+\Delta R} dR\, S_2(R) \\ &= \frac{1}{\Delta R} \sum_{i,j \in \mathbf{J}} p_{T,i}\, p_{T,j}\, I_{[R,R+\Delta R)}(R_{ij}), \end{aligned} \tag{2.3}$$

where $I_A(R_{ij})$ is an indicator function of the angular distance R_{ij} of the domain A,

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

JHEP07(2019)135

The spectral function $S_2(R;\Delta R)$ is, therefore, the sum of the product of p_T's of the two jet constituents with an angular distance R_{ij} lying between R and $R+\Delta R$.

We obtain IRC safe quantities by multiplying smooth functions[2] $w(\vec{R})$ and $P_T(\vec{R})$ (or $S_2(R)$), and integrating over $\vec{R}$. To understand the IRC safety of $P_T(\vec{R})$, let us consider splitting of a given constituent i_0 in $\mathbf{J}$ into two constituents, $i_0 \to i_1 i_2$. The inner product of $w(\vec{R})$ and the difference of the energy flow before and after the splitting, $\delta P_T(\vec{R})$, is given as follows,

$$\begin{aligned}\int d\vec{R}\,\delta P_T(\vec{R}) w(\vec{R}) &= p_{T,i_1} w(\vec{R}_{i_1}) + p_{T,i_2} w(\vec{R}_{i_2}) - p_{T,i_0} w(\vec{R}_{i_0}) \\ &= \left[\delta p_{T,i_0} - p_{T,i_1}(\delta\vec{R}_{i_1}\cdot\nabla_{\vec{R}}) - p_{T,i_2}(\delta\vec{R}_{i_2}\cdot\nabla_{\vec{R}}) + \cdots\right] w(\vec{R}_{i_0}),\end{aligned} \tag{2.4}$$

where $\delta p_{T,i_0} = p_{T,i_1} + p_{T,i_2} - p_{T,i_0}$, and $\delta\vec{R}_{i_1(i_2)} = \vec{R}_{i_1(i_2)} - \vec{R}_{i_0}$. The soft limit, where i_2 carries a small momentum, corresponds to $\delta p_{T,i_0},\ \delta\vec{R}_{i_1},\ p_{T,i_2} \to 0$, while $\delta p_{T,i_0},\ \delta\vec{R}_{i_1},\ \delta\vec{R}_{i_2} \to 0$ in the collinear limit. The integral vanishes in these limits, namely the energy flow after parton splitting converges weakly [60] to the one before splitting.

The spectrum $S_2(R)$ inherits the same property. The inner product of the smooth function $w(R)$ and the difference of the spectrum, $\delta S_2(R)$, before and after the splitting $i_0 \to i_1 i_2$ is given as follows,

$$\int dR\,\delta S_2(R) w(R) = 2\sum_{j\in\mathbf{J}} \left[\delta p_{T,i_0} + p_{T,i_1}(\delta\vec{R}_{i_1}\cdot\nabla_{\vec{R}}) + p_{T,i_2}(\delta\vec{R}_{i_2}\cdot\nabla_{\vec{R}}) + \cdots\right] p_{T,j}\, w(R_{i_0 j}). \tag{2.5}$$

Again, this integral vanishes in the IRC limits. Note that the binned spectrum $S_2(R;\Delta R)$ is not completely IRC safe because of the discontinuity of the indicator function at the bin boundaries. Nevertheless, when the domain is discretized into small sections $[R_i, R_i+\Delta R_i)$, the IRC unsafe terms cancel in the sum, $\sum_i S_2(R_i;\Delta R_i)\, w(R_i)$, and it is approximately IRC safe up to binning errors.

The resulting IRC safe observables belong to C-correlators [60], which are multilinear forms of the energy flow. An n-linear C-correlator is expressed as follows,

$$\int d\vec{R}_1 \cdots d\vec{R}_n\, P_T(\vec{R}_1)\cdots P_T(\vec{R}_n)\, w(\vec{R}_1,\cdots,\vec{R}_n), \tag{2.6}$$

where w is a continuous function of $\vec{R}_1,\cdots,\vec{R}_n$. For example, an inner product of $P_T(\vec{R})$ and $w(\vec{R})$ is a linear C-correlator, and an inner product of $S_2(R)$ and $w(R)$ is a bilinear C-correlator with w depending only on the relative distance R_{12},

$$\int dR\, S_2(R)\, w(R) = \int d\vec{R}_1 d\vec{R}_2\, P_T(\vec{R}_1) P_T(\vec{R}_2)\, w(R_{12}). \tag{2.7}$$

Many well-known jet observables belong to the C-correlator, for example, a jet transverse momentum $p_{T,\mathbf{J}}$ is a linear C-correlator with $w(\vec{R}_1) \approx 1$, a jet mass $m_{\mathbf{J}}$ is a bilinear C-correlator with $w(\vec{R}_1,\vec{R}_2) \approx R_{12}^2/2$.

[2]Continuous functions are sufficient for the convergence and IRC safety [60], but we further restrict w's to smooth functions for perturbative calculations.

JHEP07(2019)135

The $S_2(R)$ spectra use all the jet constituents, but it is useful to separate the correlations of constituents of the hard subjets; we consider jet trimming for this purpose. We recluster the constituents of a jet of a radius parameter $R_{\mathbf{J}}$ to subjets with a smaller radius parameter $R_{\rm trim}$. A subjet $\mathbf{J}_a$ is discarded if $p_{T,\mathbf{J}_a} < f_{\rm trim}\, p_{T,\mathbf{J}}$, where $p_{T,\mathbf{J}}$ and $p_{T,\mathbf{J}_a}$ are the transverse momenta of the jet and a-th subjet respectively. The trimmed jet $\mathbf{J}_{\rm trim}$ is defined as a union of the remaining subjets,

$$\mathbf{J}_{\rm trim} = \bigcup_{\substack{a \\ \frac{p_{T,\mathbf{J}_a}}{p_{T,\mathbf{J}}} \geq f_{\rm trim}}} \mathbf{J}_a \,. \tag{2.8}$$

The jet trimming is beneficial because it does not introduce additional angular scale parameters other than $R_{\rm trim}$. The trimmed spectrum is then calculated using the constituents of the trimmed jet. We denote it as $S_{2,\rm trim}(R)$ and its binned version $S_{2,\rm trim}(R;\Delta R)$, which are defined as follows:

$$S_{2,\rm trim}(R) = \int d\vec{R}_1\, d\vec{R}_2\, P_{T,\rm trim}(\vec{R}_1)\, P_{T,\rm trim}(\vec{R}_2) \cdot \delta(R - R_{12}), \tag{2.9}$$

$$P_{T,\rm trim}(\vec{R}) = \sum_{i \in \mathbf{J}_{\rm trim}} p_{T,i}\, \delta(\vec{R} - \vec{R}_i), \tag{2.10}$$

$$S_{2,\rm trim}(R;\Delta R) = \frac{1}{\Delta R} \sum_{i,j \in \mathbf{J}_{\rm trim}} p_{T,i}\; p_{T,j} \cdot I_{[R,R+\Delta R)}(R_{ij}), \tag{2.11}$$

where $P_{T,\rm trim}(\vec{R})$ is the energy flow of $\mathbf{J}_{\rm trim}$.

In the limit of the constituents of each subjet $\mathbf{J}_a$ are localized, the energy flow and the jet spectrum can be approximated in terms of the subjet momenta. The energy flow of such a jet is decomposed into a sum of energy flows of all the subjets,

$$P_T(\vec{R}) = \sum_a P_{T,a}(\vec{R}), \quad P_{T,a}(\vec{R}) = \sum_{i \in \mathbf{J}_a} p_{T,i} \delta(\vec{R} - \vec{R}_i). \tag{2.12}$$

The energy flow of each subjet converges weakly to $p_{T,\mathbf{J}_a}\delta(\vec{R} - \vec{R}_{\mathbf{J}_a})$. The $S_2(R)$ spectrum can be approximated by the momenta of the subjets, i.e.,

$$S_2(R;\Delta R) \approx \sum_{\substack{a,b \\ \mathbf{J}_a,\mathbf{J}_b \subset \mathbf{J}}} p_{T,\mathbf{J}_a}\, p_{T,\mathbf{J}_b} \cdot I_{[R,R+\Delta R)}(R_{ab}). \tag{2.13}$$

The jet trimming also introduces a p_T scale hierarchy among the subjets, and so their pairwise contributions to $S_2(R;\Delta R)$ can be classified by the scale. We define a quantity $S_{2,\rm soft}(R;\Delta R)$ where

$$S_{2,\rm soft}(R;\Delta R) = S_2(R;\Delta R) - S_{2,\rm trim}(R;\Delta R). \tag{2.14}$$

In the r.h.s. of the above equation, the correlations among the constituents of the hard subjets are canceled, and we have

$$S_{2,\rm trim}(R;\Delta R) = p_{T,\mathbf{J}}^2 \cdot \mathcal{O}\left[1\right], \tag{2.15}$$

$$S_{2,\rm soft}(R;\Delta R) = p_{T,\mathbf{J}}^2 \cdot \left(\mathcal{O}\left[f_{\rm trim}\right] + \mathcal{O}\left[f_{\rm trim}^2\right]\right). \tag{2.16}$$

JHEP07(2019)135

The dominant contributions to $S_{2,\text{soft}}(R;\Delta R)$ (i.e., the $\mathcal{O}[f_{\text{trim}}]$ terms) come from the correlations between a constituent in $\mathbf{J}_{\text{trim}}$ and a constituent in $\mathbf{J}-\mathbf{J}_{\text{trim}}$. The subleading $\mathcal{O}\left[f_{\text{trim}}^2\right]$ terms denote the correlations among the constituents in $\mathbf{J}-\mathbf{J}_{\text{trim}}$.

2.2 Derivation of classifiers based on energy flows and jet spectra

We discuss the relation between $S_2(R)$ and neural network classifiers trained on the energy flow $P_T(\vec{R})$. A general softmax classifier that solves K-class jet classification problem can be expressed as a functional $\hat{\Psi}_i$ which maps the energy flow to real numbers h_i, i.e.,

$$h_i = \hat{\Psi}_i[P_T] \tag{2.17}$$

$$\hat{y} = \varphi_{\text{softmax}}(\vec{z}), \quad z_k = w_{ki}^{(\text{out})} h_i + b_k^{(\text{out})}, \quad k \in \{1,\cdots,K\}, \tag{2.18}$$

where $w_{ki}^{(\text{out})}$ and $b_k^{(\text{out})}$ are the weights and biases of the output layer, and $\hat{y}$ is the prediction of the classifier. Here the φ_{softmax} is the softmax function whose k-th component is expressed as follows,

$$\varphi_{\text{softmax},k}(\vec{z}) = \frac{e^{z_k}}{\sum_{k=1}^{K} e^{z_k}}. \tag{2.19}$$

Many jet classifiers can be expressed in the form of eq. (2.17). For example, in the cut-based analysis, $\hat{\Psi}_i$ is a jet substructure variable, such as a ratio of n-subjettiness [12], a ratio of energy correlation functions [16, 17], etc. The deep neural network classifiers, such as artificial neural network tagger [18], convolutional neural network using pixelated jet images [19], energy flow network [31], etc., are also described by eq. (2.17). The neural networks that are introduced in section 3 and section 4 also belong to this category.

The jet spectra S_2 and $S_{2,\text{trim}}$ can be derived from eq. (2.17) using a functional Taylor expansion. The energy flow is decomposed by trimming as follows,

$$P_{T,a}(\vec{R}) = \begin{cases} P_{T,\text{trim}}(\vec{R}) & a = 1, \\ P_T(\vec{R}) - P_{T,\text{trim}}(\vec{R}) & a = 2. \end{cases} \tag{2.20}$$

One can express $\hat{\Psi}_i[P_{T,a}]$ as a functional series at a reference point $P_{T,a}(\vec{R}) = 0$,

$$h_i = w_i^{(0)} + \int d\vec{R}\, P_{T,a}(\vec{R}) w_{i,a}^{(1)}(\vec{R}) + \frac{1}{2!}\int d\vec{R}_1 d\vec{R}_2\, P_{T,a}(\vec{R}_1) P_{T,b}(\vec{R}_2) w_{i,ab}^{(2)}(\vec{R}_1,\vec{R}_2) + \cdots, \tag{2.21}$$

where $w_{i,a_1\cdots a_n}^{(n)}(\vec{R}_1,\cdots,\vec{R}_n)$ is the coefficient of n-th correlation function. The first order coefficient $w_{i,a}^{(1)}$ can be chosen as a constant if we are not interested in features depending on reference vectors, for example, jet axes, beam directions, etc. The linear term in $P_T(\vec{R})$ of eq. (2.21) is related to the jet momentum $p_{T,\mathbf{J}}$ and trimmed jet momentum $p_{T,\mathbf{J},\text{trim}}$ as follows,

$$\int d\vec{R}\, P_{T,1}(\vec{R}) \simeq p_{T,\mathbf{J},\text{trim}}, \quad \int d\vec{R}\, P_{T,2}(\vec{R}) \simeq p_{T,\mathbf{J}} - p_{T,\mathbf{J},\text{trim}}. \tag{2.22}$$

The second order coefficient $w_{i,ab}^{(2)}$, the first non-trivial term of the series expansion, is a function of the relative distance of $\vec{R}_1$ and $\vec{R}_2$. The basis vectors of $w_{i,ab}^{(2)}$ are two-point

JHEP07(2019)135

correlation functions $S_{2,ab}(R)$,

$$h_i = w_i^{(0)} + \int d\vec{R}\, P_{T,a}(\vec{R})\, w_{i,a}^{(1)} + \frac{1}{2!}\int dR\, S_{2,ab}(R)\, w_{i,ab}^{(2)}(R) + \cdots \tag{2.23}$$

$$S_{2,ab}(R) = \int d\vec{R}_1 d\vec{R}_2\, P_{T,a}(\vec{R}_1) P_{T,b}(\vec{R}_2)\, \delta(R - R_{12}). \tag{2.24}$$

The spectra S_2 and $S_{2,\text{trim}}$ are expressed in terms of $S_{2,ab}$ as follows,

$$S_2(R) = \sum_{a,b} S_{2,ab}(R), \quad S_{2,\text{trim}}(R) = S_{2,11}(R). \tag{2.25}$$

Instead of the energy flows, we consider a classifier of $S_{2,A}$ (A = trim, soft),

$$h_i = \Psi_i[S_{2,A}; \vec{x}_{\text{kin}}], \tag{2.26}$$

where $\vec{x}_{\text{kin}}$ is a set of additional inputs to the classifier based on the kinematics of the jet. Similar to eq. (2.23), we expand eq. (2.26) around $S_{2,A}(R) = 0$ as

$$h_i = w_i^{(0)}(\vec{x}_{\text{kin}}) + \int dR\, S_{2,A}(R)\, \frac{w_{i,A}^{(2)}(R; \vec{x}_{\text{kin}})}{2} + \frac{1}{2}\int dR_1 dR_2\, S_{2,A_1}(R_1) S_{2,A_2}(R_2)\, \frac{w_{i,A_1A_2}^{(4)}(R_1, R_2; \vec{x}_{\text{kin}})}{12} + \cdots, \tag{2.27}$$

where $w_{i,A_1\cdots A_{\frac{n}{2}}}^{(n)}$ is the weight function corresponding to $w_i^{(n)}$ in eq. (2.26). One may further truncate the series to get a linear form,

$$h_i = \frac{1}{2}\int dR\, S_{2,A}(R) w_{i,A}^{(2)}(R; \vec{x}_{\text{kin}}). \tag{2.28}$$

The above-mentioned linear setup has an advantage on the interpretability and visualization of the network predictions; we discuss more on this network in section 4.

2.3 Relation between two-point correlation spectra and energy flow polynomials

Both the two-point correlation spectra and the energy flow polynomials [68] with two vertices span the set of bilinear C-correlators; therefore, there is a transformation rule between them. We first extend the definition of the energy flow polynomials to compare them to $S_{2,ab}$. Since $S_{2,ab}$ is a multivariate function of energy flows, we introduce a multivariate energy flow polynomial with two labeled vertices,

$$\text{EFP}_{2,ab}^{(n)} = \int d\vec{R}_1 d\vec{R}_2\, P_{T,a}(\vec{R}_1) P_{T,b}(\vec{R}_1)\, R_{12}^n = \sum_{i\in\mathbf{J}_a}\sum_{j\in\mathbf{J}_b} p_{T,i} p_{T,j} R_{ij}^n. \tag{2.29}$$

This expression suggests that R^n can be considered as an angular weighting function $w_{i,ab}^{(2)}(R)$ in eq. (2.23).

JHEP07(2019)135

The resulting transformation from $S_{2,ab}(R)$ to $\mathrm{EFP}^{(n)}_{2,ab}$ is the Mellin transformation,

$$\mathrm{EFP}^{(n-1)}_{2,ab} = \int_0^\infty dR\, R^{n-1} \cdot S_{2,ab}(R). \tag{2.30}$$

The integral on the right-hand side is finite because $S_{2,ab}(R)$ vanishes on $R \gg 2R_{\mathbf{J}}$. The inverse transform is also well-defined if we allow the exponent n in the angular weighting function of $\mathrm{EFP}^{(n)}_{2,ab}$ to be a complex number.

2.4 Spectra of two-prong jets

The $S_2(R;\Delta R)$ spectrum is useful to identify the substructures of the jet and also to characterize the jet. Typically, $S_2(R;\Delta R)$ of QCD jet has a peak at $R = 0$ with a long tail towards large R. The peak originates from the autocorrelation term $\sum_i p_{T,i}^2$ in eq. (2.3). On the other hand, if a jet originates from a Higgs boson decaying into $b\bar{b}$, the b-partons create two isolated cores inside the jet. The spectrum of the Higgs jet has a peak at the angular scale equal to the angle between the two clusters. In addition, $S_2(R;\Delta R)$ encodes the fragmentation pattern of b-partons.

At the LHC, boosted heavy objects such as top quark, gauge bosons and Higgs boson decaying into quarks can be studied by identifying jet substructures. Usually, these substructures are characterized by parameters such as D_2 defined as,

$$\begin{aligned} D_2^\beta &= e_3^\beta/(e_2^\beta)^3, \\ e_2^\beta &= \frac{1}{p_{T,\mathbf{J}}^2} \sum_{i,j\in\mathbf{J}, i<j} p_{T,i} p_{T,j} R_{ij}^\beta, \\ e_3^\beta &= \frac{1}{p_{T,\mathbf{J}}^3} \sum_{i,j,k\in\mathbf{J}, i<j<k} p_{T,i} p_{T,j} p_{T,k} R_{ij}^\beta R_{jk}^\beta R_{ki}^\beta, \end{aligned} \tag{2.31}$$

where β is the angular exponent. If a jet has a two-prong substructure, D_2 is much less than one. The jet spectrum $S_2(R;\Delta R)$ contains more information than D_2, and therefore, the analysis with $S_2(R;\Delta R)$ goes beyond the one using D_2. It was shown that a neural network trained on $S_2(R;\Delta R)$ distinguishes Higgs jet from QCD jet better than the one trained on D_2 [28].

To study the fragmentation pattern of the b-partons and their color connection to the mother particle, we introduce a color-octet scalar, sgluon (σ). We assume that the Higgs boson (h) and σ decay into $b\bar{b}$ through the interaction,

$$\mathcal{L}_{\mathrm{SM}} \ni y_{hb\bar{b}}\, h\, \bar{b}b + \mathrm{h.c.} \tag{2.32}$$

$$\mathcal{L}_{\mathrm{Sgluon}} \ni y_{\sigma b\bar{b}}\, \sigma^a\, \bar{b} T^a b + \mathrm{h.c.}\,. \tag{2.33}$$

The Higgs boson is a color singlet particle, and the decay $h \to b\bar{b}$ is isolated in color flows. Therefore, $S_2(R;\Delta R)$ beyond the angle between the b-partons, i.e., $R_{b\bar{b}}$, is suppressed due to the color coherence. No such constraint on the angular scale exists for sgluon and QCD jets. Meanwhile, the Higgs jet and sgluon jet have the same two-prong substructure, unlike the QCD jet, as both are originating from a particle decaying into $b\bar{b}$ final states.

JHEP07(2019)135

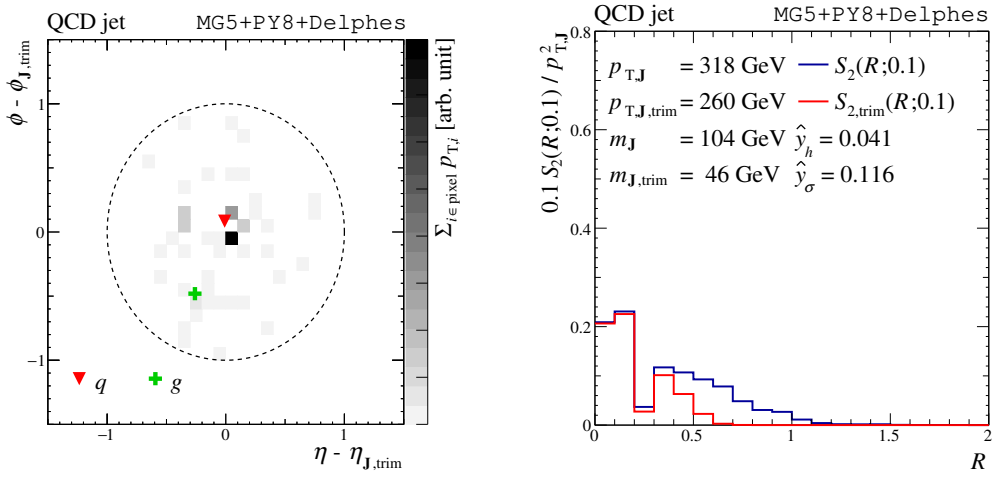

Figure 1. A jet image (left) and the corresponding S_2 (blue) and $S_{2,\mathrm{trim}}$ (red) spectra (right) of the leading jet of a $pp \to Zj$ event. In the jet image, a red triangle is a position of a parton level quark in the jet, and a green "+" shows a leading gluon emitted from the quark.

To study the spectra of two-prong jets, we simulate events as follows. We use `Madgraph5 2.6.1` [69] to generate the events of $pp \to Zh$, $pp \to Z\sigma$, and $pp \to Zj$ processes with a collision energy of 13 TeV and the Z boson decaying to a pair of neutrinos. These events are then passed to `Pythia 8.226` [65] for the parton shower and hadronizations. To study the impact of the parton shower and hadronization schemes, we also pass those parton level events to `Herwig 7.1.3` [66, 67]. A color octet scalar UFO model [70, 71] generated by `Feynrule 2.0` [72, 73] is used to simulate $pp \to Z\sigma$ process. The masses and widths of Higgs boson and sgluon are $m_h = m_\sigma = 125\,\mathrm{GeV}$ and $\Gamma_\sigma = \Gamma_h = 6.4\,\mathrm{MeV}$. The detector response is simulated by `Delphes 3.4.1` [74] with the default ATLAS detector configuration. We use `FastJet 3.3.0` [75, 76] to reconstruct jets from the calorimeter towers using anti-k_T algorithm [77] with the radius parameter $R_{\mathbf{J}} = 1.0$. For jet trimming, we use $R_{\mathrm{trim}} = 0.2$ and $f_{\mathrm{trim}} = 0.05$. We select the events with the leading jet transverse momentum $p_{T,\mathbf{J}} \in [300, 400]\,\mathrm{GeV}$ and its mass $m_{\mathbf{J}} \in [100, 150]\,\mathrm{GeV}$. For Higgs jet and sgluon jet, we additionally require that the two b-partons originating from their decay are located within $R_{\mathbf{J}}$ from the jet axis. More details on our simulations are described in appendix A.

In figure 1, we show the pixelated jet image (left panel) and S_2 and $S_{2,\mathrm{trim}}$ spectra (right panel) of a QCD jet. There are high energy deposits in the jet image near the jet center along with a wide spray of soft activity. It also has a moderate amount of radiation at $(-0.4, 0.0)$. As a result, $S_2(R;\Delta R)$ spectra has a long tail starting from $R = 0.4$. The jet trimming eliminates a significant amount of soft particles and, therefore, the tail does not appear in $S_{2,\mathrm{trim}}(R;\Delta R)$. The remaining cross-correlations contributing to $S_{2,\mathrm{trim}}(R;\Delta R)$ are the ones between high and moderate energy deposits. Most of the energy deposits are concentrated at the center, and the peak intensity at $R = 0.4$ is much lower than the intensity from autocorrelations at $R = 0$.

In figure 2, we show $S_2(R;\Delta R)$ and $S_{2,\mathrm{trim}}(R;\Delta R)$ distributions of a Higgs jet. For this particular event, $S_2(R;\Delta R)$ distribution is similar to $S_{2,\mathrm{trim}}(R;\Delta R)$ distribution, and

JHEP07(2019)135

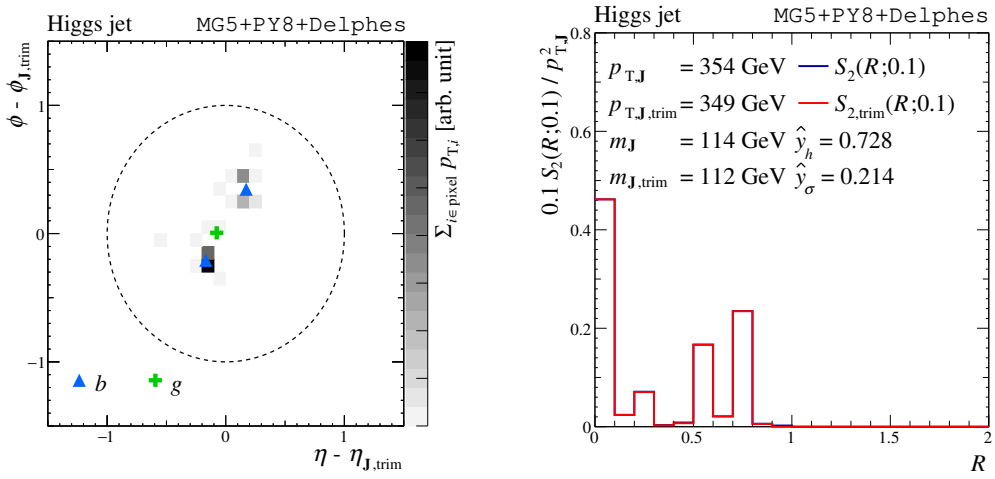

Figure 2. A jet image (left) and the corresponding S_2 (blue) and $S_{2,\mathrm{trim}}$ (red) spectra (right) of the leading jet of a $pp \to Zh$ event. The blue triangles in the jet image are positions of the parton level bottom quarks from the Higgs decay, and a green "+" shows the position of a leading gluon emitted from a bottom quark.

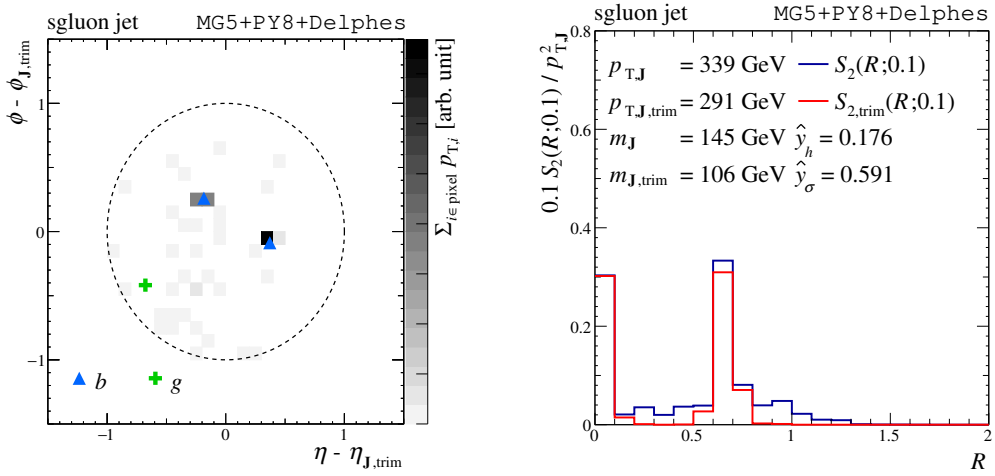

Figure 3. A jet image (left) and the corresponding S_2 (blue) and $S_{2,\mathrm{trim}}$ (red) spectra (right) of the leading jet of a $pp \to Z\sigma$ event. The blue triangles in the jet image are positions of the parton level bottom quarks from the sgluon decay, and green "+" show the position of a leading gluon emitted from a bottom quark.

their difference $S_{2,\mathrm{soft}}(R;\Delta R)$ is hard to be seen. No significant activity has been observed beyond the peak at $R \sim 0.8$, mostly because the Higgs jet is very compact compared to the QCD jet. Correspondingly, there are two prominent subjets in the jet image, while most of the cells have no jets.

Finally, we show the $S_2(R;\Delta R)$ and $S_{2,\mathrm{trim}}(R;\Delta R)$ distributions of a sgluon jet in figure 3. The $S_2(R;\Delta R)$ distribution has a large peak at $R = 0.6$ which is as significant as the one at $R = 0$. This spectrum is qualitatively similar to the Higgs jet in figure 2. However, the $S_2(R;\Delta R)$ spectrum has a long tail beyond $R_{\mathbf{J}}$ as compared with that of a

JHEP07(2019)135

Higgs jet. The tail disappears after jet trimming, like the QCD jet in figure 1, that makes the $S_{2,\text{trim}}(R;\Delta R)$ distribution more compact. From figure 1–3, we observe that $S_{2,\text{trim}}$ and $S_{2,\text{soft}}$ include useful complementary information.

In [28], it was shown that a neural network classifier trained on $S_2(R;\Delta R)$ and $S_{2,\text{trim}}(R;\Delta R)$ spectra performs better than one without $S_{2,\text{trim}}(R;\Delta R)$. The reason is that the hard and soft correlation terms in $S_2(R)$, i.e., $\mathcal{O}[1]$ terms in eq. (2.15) and $\mathcal{O}[f_{\text{trim}}]+\mathcal{O}[f_{\text{trim}}^2]$ in eq. (2.16) respectively, can be resolved by the jet trimming. Therefore, we use the orthogonal combinations, namely $S_{2,\text{trim}}$ and $S_{2,\text{soft}}$, throughout this paper.

The $S_{2,\text{trim}}$ and $S_{2,\text{soft}}$ spectra encode the important features of the parton shower and fragmentation, and, thus, may be regarded as a well-motivated prototype. The hard partons evolve by the parton splittings $i \to i_1 i_2$, which are parameterized by the angle $R_{i_1 i_2}$ and momentum fraction z with $p_{T,i_1} = z p_{T,i}$ and $p_{T,i_2} = (1-z) p_{T,i}$. The splitting generates two-point correlation $z(1-z)p_{T,i}^2$ at $R_{i_1 i_2}$. Therefore, S_2 spectra encode the parton splitting at any angular scale.

3 Classifying Higgs jet, sgluon jet, and QCD jet

In this section, we introduce a neural network trained on the jet spectra for classifying Higgs jet, sgluon jet, and QCD jet. We first discuss the basic kinematic features of these jets and then outline their dependence on the parton shower simulators. Afterward, we show the details of the neural network and then present our results in terms of the receiver operating characteristic (ROC) curves.

3.1 Basic kinematics

In figure 4, we show $p_{T,\mathbf{J}}$ and $m_{\mathbf{J}}$ distributions for the Higgs boson, sgluon, and QCD jets. The solid and dashed lines correspond to `Pythia 8` (`PY8`) and `Herwig 7` (`HW7`) generated jets, respectively. The mild differences in the p_T distribution are due to the difference in their matrix elements. The Higgs jet is produced via s-channel process, while the sgluon and QCD jet are produced via t-channel and u-channel processes; hence, $p_{T,\mathbf{J}}$ scalings are different. Not much difference is observed between the $p_{T,\mathbf{J}}$ distributions of `PY8` and `HW7` samples. This is because $p_{T,\mathbf{J}}$ is mostly determined by the matrix level p_T of the leading parton and the jet algorithm with large radius parameter clusters most of the radiations from this parton into a single jet. However, the difference between $m_{\mathbf{J}}$ distributions is large. The peak at m_σ of sgluon jet is significantly broader than that of Higgs jet because radiations of the b-partons from the Higgs boson decay are mostly confined due to the color coherence, but those of the sgluon are not. As a consequence, `PY8` and `HW7` generate different $m_{\mathbf{J}}$ distributions.

We assume that both Higgs boson and sgluon have narrow-widths although sgluon width can be large. An increase in the width will broaden $R_{b\bar{b}}$ distribution of $\sigma \to b\bar{b}$ that has a peak at the characteristic angular scale,

$$\hat{R}_{b\bar{b}} = \frac{2m_h}{p_{T,\mathbf{J}}} = \frac{2m_\sigma}{p_{T,\mathbf{J}}} \simeq \frac{2m_{\mathbf{J}}}{p_{T,\mathbf{J}}}. \tag{3.1}$$

JHEP07(2019)135

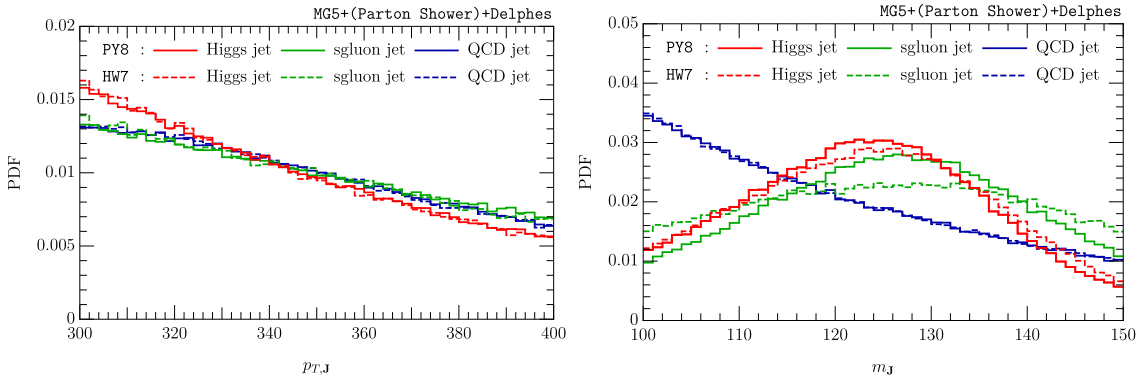

Figure 4. The distribution of $p_{T,\mathbf{J}}$ (left) and $m_{\mathbf{J}}$ (right) of the leading jet. The red, green, and blue solid (dashed) lines correspond to the Higgs jet, sgluon jet, and QCD jet of PY8 (HW7) samples, respectively.

For example, the variation of $R_{b\bar{b}}$ is only about 0.07 for $p_{T,\mathbf{J}} = 300\,\text{GeV}$, $\Gamma_\sigma = 10\,\text{GeV}$ and 0.05 for $p_{T,\mathbf{J}} = 400\,\text{GeV}$, $\Gamma_\sigma = 10\,\text{GeV}$. Those variations are close to the calorimeter angular resolution ~ 0.1 and do not affect the calorimeter level analysis.

We first make a quantitative estimate of the radiation pattern inside the jet. To do so, we define two quantities comparing $S_2(R)$ spectra around $\hat{R}_{b\bar{b}}$,

$$R_{\text{sym}} = \frac{\int_{a\hat{R}_{b\bar{b}}}^{\min[a'\hat{R}_{b\bar{b}},R_{\mathbf{J}}]} dR\, S_2(R)}{\int_0^{a\hat{R}_{b\bar{b}}} dR\, S_2(R) + \int_{\min[a'\hat{R}_{b\bar{b}},R_{\mathbf{J}}]}^{\infty} dR\, S_2(R)}, \tag{3.2}$$

$$R_{\text{rad}} = \frac{C\int_{\min[a'\hat{R}_{b\bar{b}},R_{\mathbf{J}}]}^{\infty} dR\, S_2(R)}{\int_0^{\min[a'\hat{R}_{b\bar{b}},R_{\mathbf{J}}]} dR\, S_2(R) + C\int_{\min[a'\hat{R}_{b\bar{b}},R_{\mathbf{J}}]}^{\infty} dR\, S_2(R)} \tag{3.3}$$

with $a = 0.75$, $a' = 1.25$ and $C = 40$. The ratio R_{sym} compares energy deposits around $\hat{R}_{b\bar{b}}$ and in its surrounding angular scales [28]. The ratio is sensitive to the correlation between the two hard substructures of the Higgs jet. On the other hand, The R_{rad} is sensitive to the color of mother particle as it compares energy deposits in the large angular scales.

We show the R_{sym} distributions in the left panel of figure 5. The distributions of the Higgs jet and sgluon jet are similar because both of the $S_2(R)$ spectra peak at $R_{b\bar{b}}$. Meanwhile, the two-point correlations for the QCD jet are not localized around the $R_{b\bar{b}}$ scale, so the R_{sym} is smaller than that of a Higgs jet and a sgluon jet. In the right panel of figure 5, we show the R_{rad} distributions. The R_{rad} of the sgluon jet and QCD jet are large on average, while R_{rad} is smaller for Higgs jet because large angle radiations are suppressed.

The difference in R_{sym} and R_{rad} distributions between PY8 and HW7 samples is small; however, there is an appreciable difference in the restricted phase space. In figure 6, we plot R_{rad} distributions after the selection, $R_{\text{sym}} > 0.85$, so that the jets always contain two hard subjets with similar transverse momenta. The PY8 (solid line) and HW7 (dashed line) samples have significantly different R_{rad} distributions for the Higgs jet. Such a difference is not observed for the QCD/sgluon jets. The observed deviation for the Higgs jets could be

JHEP07(2019)135

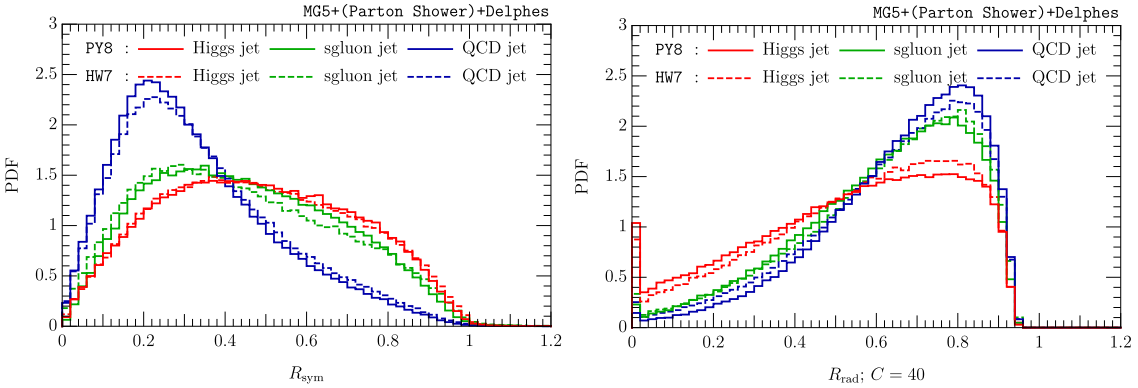

Figure 5. The distribution of $R_{\rm prong,sym}$ (left) and $R_{\rm rad}$ (right) for Higgs jet (red), sgluon jet (green), and QCD jet (blue). The solid (dashed) lines correspond to the jets of PY8 (HW7) samples.

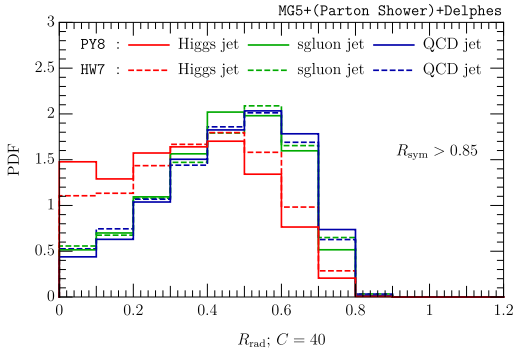

Figure 6. The distribution of $R_{\rm rad}$ for Higgs jet (red), sgluon jet (green), and QCD jet (blue) with an additional selection of $R_{\rm sym} > 0.85$. The solid (dashed) lines correspond to the jets of PY8 (HW7) samples.

originating from the difference of the parton shower scheme. The angular-ordered shower is adopted in HW7. On the other hand, the p_T-ordered shower is the default shower algorithm for PY8 where angular ordering is enforced by hand. An artificial veto in p_T-ordered shower introduces the mismatch to the angular-ordered shower at double-leading log level [78, 79].

3.2 Multilayer perceptron of spectra

We introduce a neural network trained on the kinematic and spectrum ($S_{2,\rm trim}$ and $S_{2,\rm soft}$) variables to classify the jets. A schematic diagram of the architecture of the classifier is shown in figure 7. The following set of inputs is used,

$$\vec{x} = \{p_{T,\mathbf{J}}, m_{\mathbf{J}}, p_{T,\mathbf{J},\rm trim}, m_{\mathbf{J},\rm trim}\} \cup \left\{ S^k_{2,\rm trim}, S^k_{2,\rm soft} \,|\, k \in \{0, \cdots, 19\} \right\}, \tag{3.4}$$

where $p_{T,\mathbf{J},\rm trim}$ and $m_{\mathbf{J},\rm trim}$ are the transverse momentum and mass of the trimmed jet, respectively. The discretized spectra $S^k_{2,\rm trim}$ and $S^k_{2,\rm soft}$ are used to analyze the radiation pattern of the jet,

$$S^k_{2,\rm trim} = S_{2,\rm trim}(0.1\,k; 0.1), \tag{3.5}$$

$$S^k_{2,\rm soft} = S_2(0.1\,k; 0.1) - S_{2,\rm trim}(0.1\,k; 0.1). \tag{3.6}$$

JHEP07(2019)135

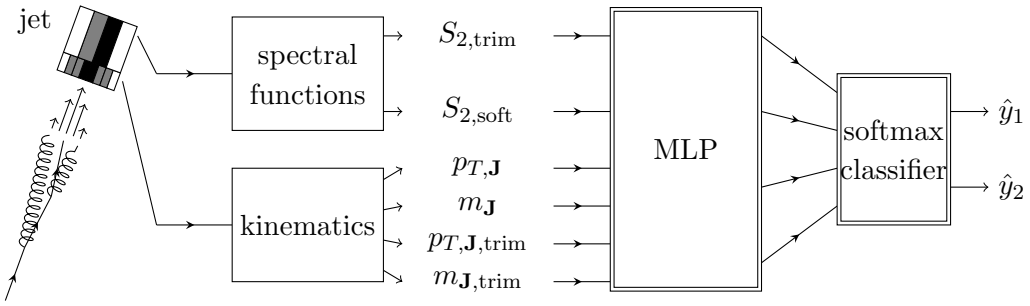

Figure 7. Schematic diagram of the classifier, including the multilayer perceptron (MLP). The double bordered boxes represent trainable modules.

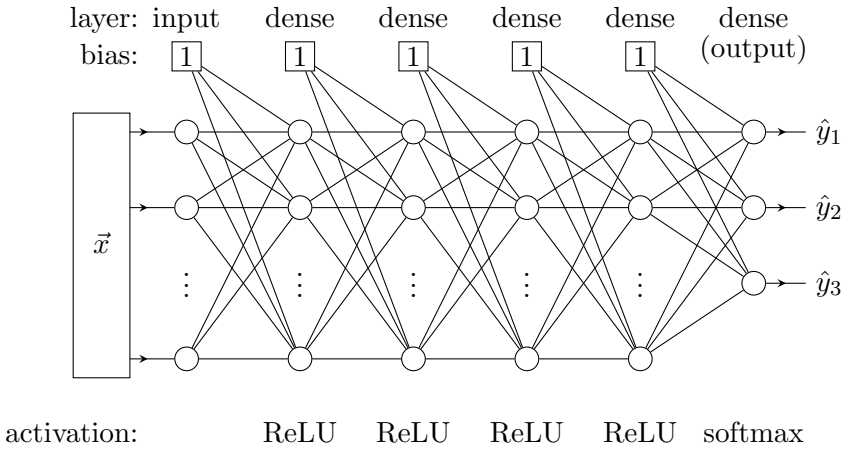

Figure 8. Schematic diagram of the multilayer perceptron.

Here we take the bin width $\Delta R = 0.1$, which is approximately the angular resolution of hadronic calorimeter of the ATLAS detector. Note that the maximum separation between any two constituents of the jet is $2R_{\mathbf{J}}$.

A multilayer perceptron (MLP) with L layers is used to map the inputs to the class prediction. The following first-order recurrence relation between the layers describes an MLP,

$$h_i^{(\ell)} = \varphi^{(\ell)}\left(w_{ij}^{(\ell)} h_j^{(\ell-1)} + b_i^{(\ell)}\right), \quad \vec{h}^{(0)} = \vec{x}, \tag{3.7}$$

where $w_{ij}^{(\ell)}$ and $b_i^{(\ell)}$ are the weight and bias of the ℓ-th layer. The activation function of the ℓ-th layer, $\varphi^{(\ell)} : \mathbb{R} \to \mathbb{R}$, is a monotonic and nonlinear function. We use four hidden layers with 1000, 800, 400, and 200 nodes, respectively, with a rectified linear unit (ReLU), $\varphi_{\text{ReLU}}(x) = \max(0, x)$, as the activation function. This MLP will identify important features of inputs for the classification after training. To make a class prediction, we provide the outputs of the MLP to a softmax classifier in eq. (2.18). The whole network architecture is illustrated in figure 8.

The MLP is trained by minimizing a loss function including categorical cross-entropy and L_2 weight regularization [80],

$$\mathcal{L} = \frac{1}{N_{\text{events}}} \sum_{\text{events}}^{N_{\text{events}}} \sum_i y_i \log \hat{y}_i + \lambda \sum_{\ell=1}^{L} \sum_{i,j} |w_{ij}^{(\ell)}|^2, \tag{3.8}$$

JHEP07(2019)135

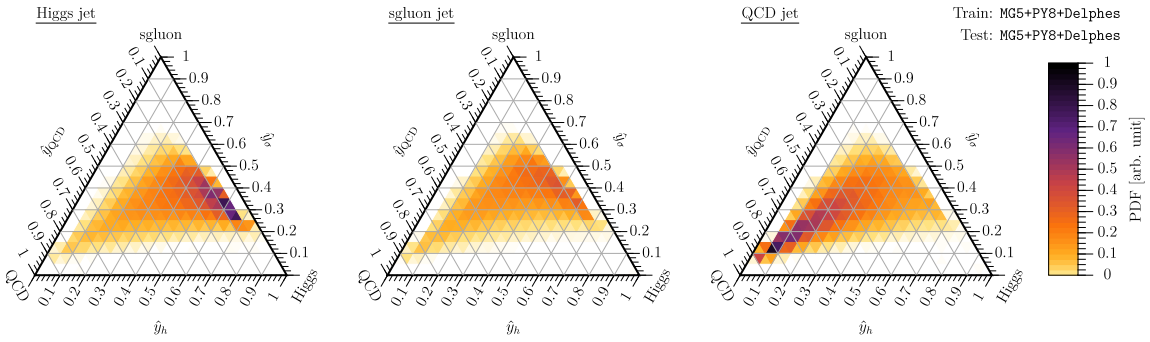

Figure 9. Ternary plots of the predicted label vector $\hat{y}$ of the MLP for the Higgs jet (left), sgluon jet (center), and QCD jet (right).

where N_{events} is the total number of events in the training data set, λ is a weight decay constant associated to the L_2 regularization. We choose $\lambda = 0.01$. The y_i ($\hat{y}_i$) denotes the components of the truth (predicted) label vector $\vec{y}$ ($\hat{y}$). The L_2 weight regularization reduces the over-fitting on the training data and also allows smooth extrapolation to the phase space that is not covered by the training sample. The minimization is done with ADAM optimizer [81]. We stop training when the validation loss has stopped improving for 50 epochs. After the minimization of the loss function, the softmax layer provides scores of the classes of a given event. The truth label vectors are defined as follows,

$$\vec{y} = \begin{cases} (1,0,0) & \text{Higgs jet}, \\ (0,1,0) & \text{sgluon jet}, \\ (0,0,1) & \text{QCD jet}. \end{cases} \tag{3.9}$$

The unnecessary symmetries in the neural network are broken by using the Glorot uniform initialization method [82]. The weights in the hidden layers are initialized by assigning random numbers between $[-\sqrt{6/(N_{\text{in}}+N_{\text{out}})}, \sqrt{6/(N_{\text{in}}+N_{\text{out}})}]$, where N_{in} and N_{out} are numbers of inputs and outputs of a layer, respectively. The biases are initialized to zero. All the inputs are standardized before training. The architecture is implemented in `Keras` [83] with backend `TensorFlow` [84].

In figure 9, we show ternary plots of the predicted label vector $\hat{y}$. The three sides of the triangle (starting from the base of the triangle and then counterclockwise) are $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$ axis; we denote them as $\hat{y}_h$, $\hat{y}_\sigma$ and $\hat{y}_{\text{QCD}}$, respectively. The $\hat{y}$ distributions of the Higgs jet and QCD jet have high-density spots that do not overlap with each other. It means that the network has found the exclusive features of those two kinds of jets. The two-prong substructure of a Higgs jet and the one-prong structure of a QCD jet are the exclusive features. However, the two-prong substructure of a sgluon jet is more radiative and less exclusive, and therefore, there are no high-density spots in the $\hat{y}$ distribution of the sgluon jet.

Next, we show ROC curves of binary classifications in figure 10 with the red dotted lines. The following signal-background classifications are considered: Higgs-QCD, sgluon-QCD, and Higgs-sgluon. We assign the truth label vectors $\vec{y} = (1,0)$ for the signal and

JHEP07(2019)135

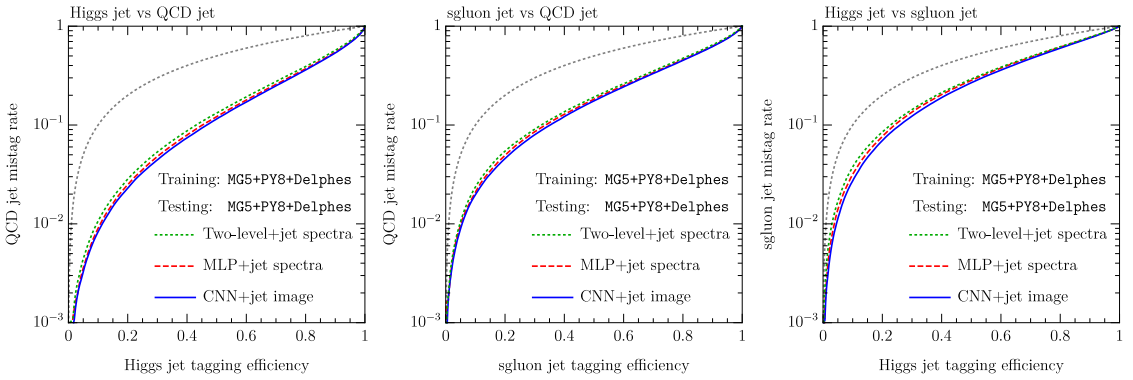

Figure 10. The ROC curves of the binary classifiers: the MLP trained on $S_{2,\text{trim}}$ and $S_{2,\text{soft}}$ (red dashed), the CNN trained on jet images (blue solid), and the two-level architecture (see section 4) trained on $S_{2,\text{trim}}$ and $S_{2,\text{soft}}$ (green dotted) with PY8 samples. The dashed gray lines represent the ROC curves of the random guess. We show the results of Higgs jet vs. QCD jet (left), sgluon jet vs. QCD jet (center), and Higgs jet vs. sgluon jet (right) classifications.

$\vec{y} = (0, 1)$ for the background. The QCD jet mistag rates are comparable for both Higgs-QCD and sgluon-QCD classifications; however, the separation between the Higgs jet and sgluon jet is weaker.

We now compare these ROC curves with that of CNN trained on jet images.[3] The CNN classifier takes 20×20 inputs of the jet images, while 2×20 inputs of $S_{2,\text{trim}}$ and $S_{2,\text{soft}}$ spectra are used for the MLP. The solid blue lines in figure 10 denote the ROC curves of the CNN. Some improvement in the background mistag rates is observed compared with the MLP classifier. Quantitatively, it is only 0.2% (= 2.5%–2.3%) at the signal acceptance of 20% for Higgs-QCD classification.

3.3 Event generator dependence

The classifier introduced in the previous subsection uses not only the information of hard subjets encoded in $S_{2,\text{trim}}$ but also the soft activities captured in $S_{2,\text{soft}}$ as well. This leads to concerns about the accuracy of the models of soft physics. Specifically, the performance of the classifier could be sensitive to the soft activities in the jet while the simulated soft activities may be significantly different from the truth.

In figure 11, we compare the ROC curves of the MLP trained with PY8 and HW7 samples. As these two event generators are based on different modeling of parton shower and hadronization scheme, the comparison would give us a reasonable estimate of the systematic uncertainty originating from the generator choice.

In the left panel of figure 11, we compare the ROC curves of the Higgs jet vs. QCD jet classification for different generator choices. By doing this exercise, we estimate a systematic uncertainty in the predictions of the classifier by comparing ROC(PY8, PY8) and ROC(HW7, HW7) curves, where the first and second entries in the parenthesis correspond to the generators used to simulate the training and test samples, respectively. On the other

[3]The CNN setup is explained in detail in appendix C.

JHEP07(2019)135

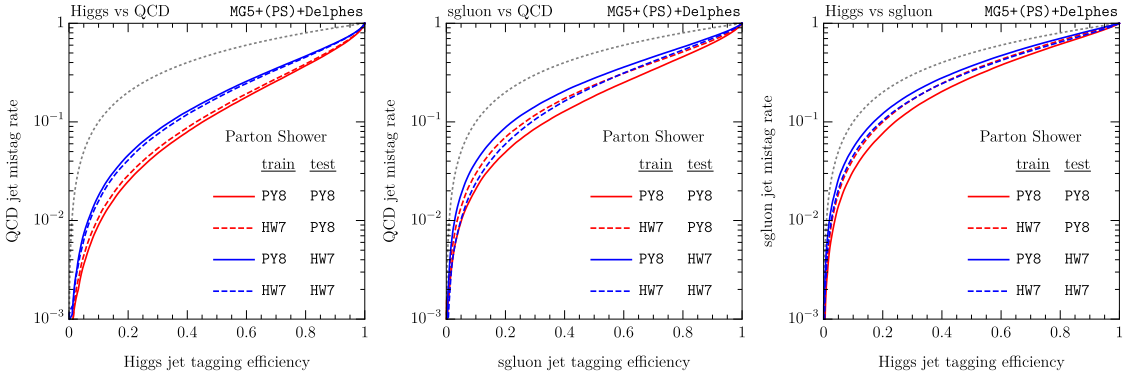

Figure 11. The ROC curves of the MLP trained on $S_{2,\text{trim}}$ and $S_{2,\text{soft}}$. The solid and dashed lines correspond to the classifier trained with `PY8` and `HW7`, respectively. The red and blue lines correspond to the classifier tested with `PY8` and `HW7`, respectively. The dashed gray lines represent the ROC curves of the random guess. We show the results of discriminating Higgs jet vs. QCD jet (left), sgluon jet vs. QCD jet (center), and Higgs jet vs. sgluon jet (right).

hand, ROC(`HW7`, `PY8`) and ROC(`PY8`, `HW7`) show the degradation of the performance of classifier trained on the "wrong sample" to analyze "real events".

The performance of the classifier improves as we vary generator combinations in the following order: ROC(`PY8`, `HW7`), ROC(`HW7`, `HW7`), ROC(`HW7`, `PY8`), and ROC(`PY8`, `PY8`). We find that the classification performance is significantly better for `PY8` test samples than that of `HW7` samples. On the other hand, the classification performance for the same test samples hardly depends on the classifiers, namely ROC(`PY8`, `HW7`) $\sim$ ROC(`HW7`, `HW7`) and ROC(`HW7`, `PY8`) $\sim$ ROC(`PY8`, `PY8`). For the Higgs jet vs. QCD jet classification, the classifier mostly concentrates on the core substructures within the jet, and here both `PY8` and `HW7` provide similar kinematics and radiation spectra. Therefore, we do not observe any significant change in the ROC curves by varying training samples while keeping the test samples the same.

In the middle panel of figure 11, we compare the classifier performance for the sgluon jet vs. QCD jet classification. It improves in the following order: ROC(`PY8`, `HW7`), ROC(`HW7`, `PY8`), ROC(`HW7`, `HW7`), and ROC(`PY8`, `PY8`). The classifiers are indeed sensitive to the choice of generators. The network trained with `PY8` (`HW7`) samples has failed to capture the features of `HW7` (`PY8`) test samples. The networks have focused on different portions of the distribution of the fragmentation functions. In the right panel of figure 11, the ROC curves for the Higgs jet vs. sgluon jet classification show similar behavior.

4 Interpretable two-level architecture

A quantitative understanding of a neural network is not straightforward because the parameters and intermediate outputs of the neural network are less readable. In this section, we propose an architecture constructed from the truncated series in eq. (2.28) and try to explain quantitatively how this network classifies events. In the case of binary classifications,

JHEP07(2019)135

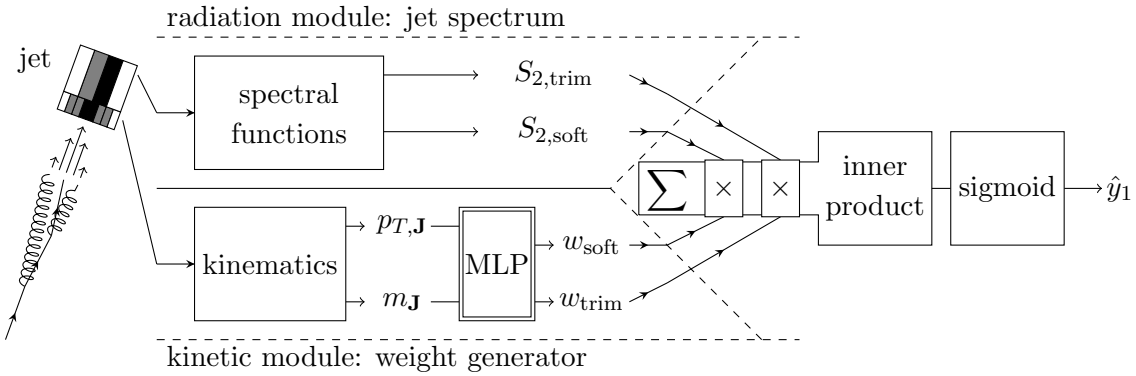

Figure 12. A schematic diagram of a two-level architecture for binary classification. An MLP trained on $p_{T,\mathbf{J}}$ and $m_{\mathbf{J}}$ generates weights w_{trim} and w_{soft} for analyzing radiation patterns encoded in $S_{2,\mathrm{trim}}$ and $S_{2,\mathrm{soft}}$ spectra. Double bordered boxes represent trainable modules. The $\hat{y}_2$ is given by the normalization, i.e., $\hat{y}_2 = 1 - \hat{y}_1$.

the discretized architecture is defined as follows,

$$h = \sum_k S^k_{2,\mathrm{trim}}\, w^k_{\mathrm{trim}} + \sum_k S^k_{2,\mathrm{soft}}\, w^k_{\mathrm{soft}}, \quad w^k_A = \frac{1}{2}\int_{R_k}^{R_k+\Delta R_k} dR\, w^{(2)}_A(R) \;\; (A = \mathrm{trim},\, \mathrm{soft}), \tag{4.1}$$

$$\hat{y}_1 = \frac{e^h}{e^h + 1}, \quad \hat{y}_2 = 1 - \hat{y}_1 \tag{4.2}$$

where w^k_A is a trainable weight. We change the activation of the output layer to a sigmoid activation since the softmax function for binary classification is essentially a sigmoid with a scale factor on its argument. The loss function is the categorical cross-entropy as defined in eq. (3.8). This setup is effectively a logistic classifier on $S^k_{2,\mathrm{trim}}$ and $S^k_{2,\mathrm{soft}}$. After the training, the magnitude of $S^k_{2,\mathrm{trim}} w^k_{\mathrm{trim}}$ or $S^k_{2,\mathrm{soft}} w^k_{\mathrm{soft}}$ is high when the corresponding $S^k_{2,\mathrm{trim}}$ or $S^k_{2,\mathrm{soft}}$ is useful for the classification.

The logistic classifier does not take into account the $p_{T,\mathbf{J}}$ dependence of $\hat{R}_{b\bar{b}}$; therefore, we introduce a two-level architecture, which is a variant of the logistic classifier. The weights w^k_A are calculated by a kinetic module $\Phi^k_A(\vec{x}_{\mathrm{kin}})$ of an MLP trained on $\vec{x}_{\mathrm{kin}} = (p_{T,\mathbf{J}}, m_{\mathbf{J}})$,

$$w^k_A = \Phi^k_A(\vec{x}_{\mathrm{kin}}). \tag{4.3}$$

A schematic diagram of this setup is shown in figure 12. The inputs $\vec{x}_{\mathrm{kin}}$ are standardized before training, and $S^k_{2,\mathrm{trim}}$ and $S^k_{2,\mathrm{soft}}$ are divided by their maximum value of the training sample because standardizing the spectra reintroduce the zeroth order term of eq. (2.27). This architecture is similar to the self-explaining neural network [85]. The Φ^k_A is modeled with an MLP of two hidden layers with exponential linear unit (ELU) activations [86],

$$\varphi_{\mathrm{ELU}}(x) = \begin{cases} x & x > 0, \\ e^x - 1 & x < 0. \end{cases} \tag{4.4}$$

JHEP07(2019)135

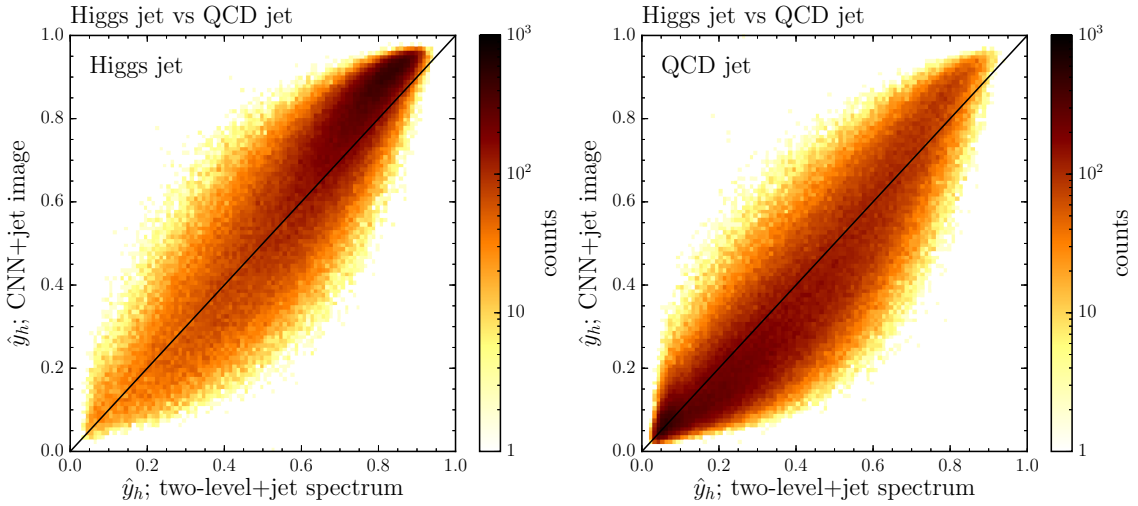

Figure 13. Two-dimensional histograms of two $\hat{y}_h$'s from the two-level architecture and the jet image CNN in Higgs jet vs. QCD jet classification. The left panel is the histogram of Higgs jets, and the right panel is that of QCD jets.

The nodes of the successive layers are configured as 400 ELU, 200 ELU, 2×20 linear respectively. We do not use ReLU in modeling Φ_A^k because dead ReLU nodes with a zero gradient kill $\vec{x}_{\rm kin}$ dependency of the weights. This is known as the dying ReLU problem. In this case, the architecture is reduced to the logistic classifier.

The vanishing gradient problem arises when the momentum range of the training sample is too small, which generates constant weights. The characteristic scale of $\hat{R}_{b\bar{b}}$ is $[0.625, 0.833]$ for the $p_{T,\mathbf{J}}$ range $[300, 400]$ GeV. The variation in $\hat{R}_{b\bar{b}}$ is about 0.2, which is not significantly large compared with the calorimeter resolution of 0.1. Therefore, we extend the $p_{T,\mathbf{J}}$ range of all the samples to $[300, 600]$ GeV. In addition, we avoid vanishing gradient problem by using He uniform initializer [87]. The weights and biases are initialized by uniform random numbers in $[-\sqrt{6/N_{\rm in}}, \sqrt{6/N_{\rm in}}]$ where $N_{\rm in}$ is the number of inputs to the layer. The advantage of using He initializer over Glorot initializer is that it generates random numbers in a wider range so that the neural network can start up from wider initial weights and gradient. The weight decay parameter λ of the L_2 weight regularizer is set to 0.001 so that the weights do not vanish too early.

After the successful training, the performance of the two-level architecture is close to that of the MLP in section 3. The green dotted lines in figure 10 are the ROC curves of the classifier. The difference is smaller than the systematic uncertainty shown in figure 11. This makes a good reason to believe that the weights in eq. (4.3) capture the essential features of the MLP and CNN in section 3. The correlation between the output of the two-level architecture and the CNN model is shown in figure 13. We can see a positive correlation between them, but the correlation is slightly tilted towards the lower triangle (upper triangle) for small (large) $\hat{y}_h$ values because the CNN performs better than the two-level architecture.

In figure 14, we show the weight functions $w_{\rm trim}$ (left) and $w_{\rm soft}$ (right) of Higgs jet vs. QCD jet classifier trained with `MG5+PY8+Delphes` samples. Note that the weights are

JHEP07(2019)135

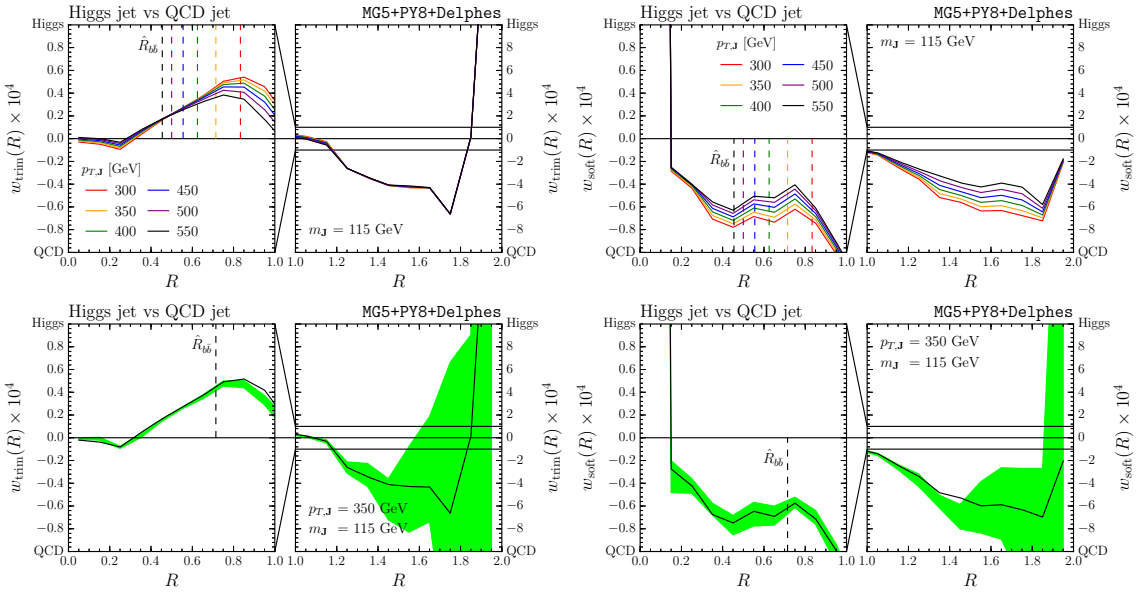

Figure 14. The weights $w_{\rm trim}$ (left) and $w_{\rm soft}$ (right) of Higgs jet vs. QCD jet classification. We show the weights for $p_{T,\mathbf{J}}$ values in the range $[300, 550]$ GeV while fixing the $m_{\mathbf{J}}$ to 115 GeV. The weights at an angular scale smaller than 1.0 are magnified by 10. The value of $w_{\rm soft}$ in the bin $[0, 0.1)$, i.e., $w^0_{\rm soft}$, is approximately 32 for all the values of $p_{T,\mathbf{J}}$. In the lower panels, we show the statistical uncertainty of the weights from the training dataset at $p_{T,\mathbf{J}} = 350$ GeV.

outputs of the neural network for the given $\vec{x}_{\rm kin}$ and not the output with the sample at the indicated $p_{T,\mathbf{J}}$ and $m_{\mathbf{J}}$. The weights in $R > R_{\mathbf{J}}$ and $w_{\rm soft}$ in $R < R_{\rm trim}$ are large compared to the weights in other angular scales. The $S_{2,\rm trim}$ and $S_{2,\rm soft}$ on these angular scales are typically smaller than that in other scales. Therefore, their weights become large to compensate for the energy difference when the corresponding value of the spectrum is useful for jet classification. The dotted lines in figure 14 denote the values of $\hat{R}_{b\bar{b}}$, defined in eq. (3.1), for different values of $p_{T,\mathbf{J}}$.

The $w_{\rm trim}$ around $\hat{R}_{b\bar{b}}$ is positive because the correlation at the scale is a characteristic feature of Higgs jet. If a Higgs boson is decaying to a pair of bottom quarks perpendicular to the boosted direction in its rest frame, the relative angular separation of the decay products is $\hat{R}_{b\bar{b}}$ in the lab frame. Due to the phase space of the decay, most of the events are distributed near $\theta = \pi/2$, where θ is Higgs decay angle relative to the boost direction. As $p_{T,\mathbf{J}}$ increases, $\hat{R}_{b\bar{b}}$ decreases, and the lower edge of $w_{\rm trim} > 0$ moves toward smaller values. The region with $w_{\rm trim} > 0$ also shifts towards smaller values of R. The weight $w_{\rm trim}$ on $R > \hat{R}_{b\bar{b}}$ is positive for capturing a Higgs jet whose θ is smaller than $\pi/2$. These events have p_T asymmetric subjets, and the cross-correlation terms in $S_{2,\rm trim}(R)$ are smaller than that of p_T symmetric case. These correlations are still useful for the classification because S_2 spectrum of QCD jet reduces much faster than that of Higgs jet. As a result, the $w_{\rm trim}$ is an increasing function in this region to compensate for the $S_{2,\rm trim}$ reduction.

For $R \gtrsim R_{\mathbf{J}}$, weight $w_{\rm trim}$ is negative. The score $\hat{y}_h$ decreases whenever there are any energy deposits at $R \gtrsim R_{\mathbf{J}}$. The crossover point from $w_{\rm trim} > 0$ to $w_{\rm trim} < 0$ shifts towards smaller values of R with an increase of $p_{T,\mathbf{J}}$ because the Higgs decay products become

JHEP07(2019)135

more asymmetric with respect to the boost direction. In such a case, one of the subjets tends to be soft so that the two-point correlations are included in $S_{2,\rm soft}$, instead of $S_{2,\rm trim}$. These contributions to $S_{2,\rm soft}$ do not affect $w_{\rm soft}$, because $S_{2,\rm soft}$ spectra of QCD jets are overwhelming at large R.

The $S_{2,\rm soft}$ on $R > R_{\rm trim}$ always reduces $\hat{y}_h$, and there is no prominent structure around $\hat{R}_{b\bar{b}}$. Moreover, $|w_{\rm soft}|$ decreases as $p_{T,\mathbf{J}}$ increases. The reduction of $w_{\rm soft}$ compensates the increase of $S_{2,\rm soft}$, and the prediction is more or less $p_{T,\mathbf{J}}$ independent. On $R > R_{\mathbf{J}}$, $|w_{\rm soft}|$ increases with R because activity in this region is a sign of QCD jet even though corresponding $S_{2,\rm soft}$ decreases due to suppressed large angle radiations.

The $w_{\rm soft}$ on $R \lesssim R_{\rm trim}$ is positive and $w_{\rm soft}$ has a break at $R \sim R_{\rm trim}$. Correlations between the constituents in a soft subjet contributes to the $S_{2,\rm soft}$ on $R < R_{\rm trim}$, i.e., $S_{2,\rm soft}(0; R_{\rm trim}) \propto f_{\rm trim}^2$. Let us assume $\mathbf{J}_a$ is a single soft subjet, then $S_{2,\rm soft}(0; R_{\rm trim}) \sim p_{T,\mathbf{J}_a}^2 \sim (p_{T,\mathbf{J}} - p_{T,\mathbf{J},\rm trim})^2$. If there are multiple soft subjets, then $S_{2,\rm soft}(0; R_{\rm trim}) \sim \sum_a p_{T,\mathbf{J}_a}^2 < (\sum_a p_{T,\mathbf{J}_a})^2 \sim (p_{T,\mathbf{J}} - p_{T,\mathbf{J},\rm trim})^2$. This triangular inequality suggests that the magnitude of $S_{2,\rm soft}(0; R_{\rm trim})$ is small for a jet with a given $p_{T,\mathbf{J}} - p_{T,\mathbf{J},\rm trim}$ when there are multiple soft jets. The positive w_2 on $R < R_{\rm trim}$ means that Higgs jet has less soft subjets than QCD jet. The $S_{2,\rm soft}^0$ consists of the autocorrelation of soft subjets, which has different energy scaling behavior compared with the other $S_{2,\rm soft}^k$. The $S_{2,\rm soft}$ on $R < R_{\rm trim}$ is $S_{2,22} \sim \mathcal{O}[f_{\rm trim}^2]$ in eq. (2.24). On the other hand, $S_{2,\rm soft}^k$ $(k \geq 1)$ is dominated by $S_{2,12}$. The $S_{2,12}$ on $R < R_{\rm trim}$ does not contribute to $S_{2,\rm soft}^0$ because it vanishes. Therefore, we may rewrite h as follows,

$$h = \int dR\, S_{2,\rm trim}(R) w_{\rm trim}(R) + \int_0^{R_{\rm trim}} dR\, S_{2,\rm soft}(R) w'_{\rm soft}(R) + \int dR\, S_{2,\rm soft}(R) w''_{\rm soft}(R) \tag{4.5}$$

where $w'_{\rm soft}$ and $w''_{\rm soft}$ are continuous functions with $w'_{\rm soft}(R_{\rm trim}) = 0$ and $w_{\rm soft}(R) = w'_{\rm soft}(R) + w''_{\rm soft}(R)$. The second term is essentially the same as $\int dR\, S_{2,22}(R) w_{2,22}^{(2)}(R)$ in eq. (2.23), and the last term is $\int dR\, S_{2,12}(R) w_{2,12}^{(2)}(R) + \int dR\, S_{2,21}(R) w_{2,21}^{(2)}(R)$.

The sudden changes of $w_{\rm trim}$ and $w_{\rm soft}$ at $R \simeq 1.8$ are due to the statistical fluctuations of the training sample. The $S_{2,\rm trim}$ and $S_{2,\rm soft}$ may have a non-zero value at large R if the jet has multiple large angle radiations with the opposite direction from the jet axis; however, the probability of such a radiation pattern is small. As a result, the number of events used for training the weights at large R is not sufficient. The large weights also do not contribute much to the classification (see figure 16).

To estimate this statistical uncertainty, we use both training and test datasets. The merged dataset is divided into ten subsets and we train a network for each of them. This will decrease the number of events in each subset by a factor 5. We estimate the uncertainty of the fit by calculating the mean and variance of the $w_{\rm trim}^k$ and $w_{\rm soft}^k$ from the ten subsets. The green band in the bottom panels of figure 14 represents the estimated uncertainty. Note that $w_{\rm trim}^k$ and $w_{\rm soft}^k$ are not sensitive to the network initialization because the two-level network is effectively a logistic regression for a fixed $p_{T,\mathbf{J}}$ and $m_{\mathbf{J}}$ and the loss function is a convex function of them.

JHEP07(2019)135

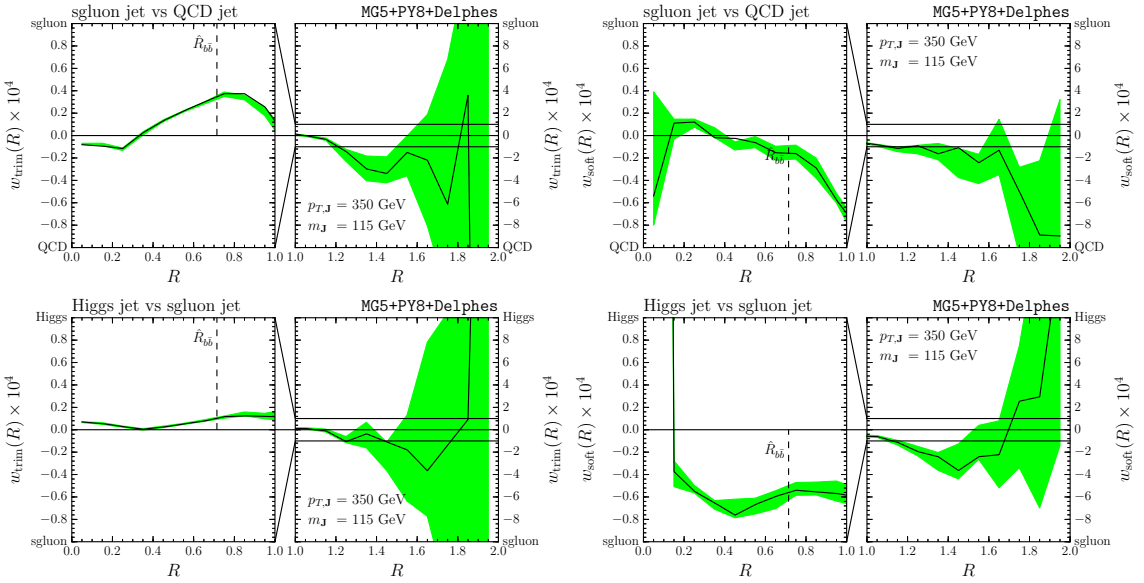

Figure 15. The weights $w_{\rm trim}$ (left) and $w_{\rm soft}$ (right) at $p_{T,\mathbf{J}} = 350\,\text{GeV}$ and $m_{\mathbf{J}} = 115\,\text{GeV}$ for the sgluon jet vs. QCD jet (top) and the Higgs jet vs. sgluon jet (bottom) classification. We show the weights with statistical uncertainty from the training dataset.

In the top panels of figure 15, we show the weights $w_{\rm trim}$ and $w_{\rm soft}$ for the sgluon jet vs. QCD jet classification with $p_{T,\mathbf{J}} = 350\,\text{GeV}$ and $m_{\mathbf{J}} = 115\,\text{GeV}$. The $w_{\rm trim}$ distribution is similar to that of the Higgs jet vs. QCD jet classification. However, the $|w_{\rm soft}|$ is much smaller. This comes from the fact that $S_{2,\rm soft}$ of sgluon jet is similar to that of QCD jet and it is less important in the classification. No peak of $S_{2,\rm soft}$ around $R \lesssim R_{\rm trim}$ also indicates that the soft substructures of sgluon jet are as radiative as QCD jet. Additionally, there is no color coherence restriction of soft radiations for the sgluon jet. This leads to small $|w_{\rm soft}|$ for $R > R_{\mathbf{J}}$. In the bottom panels of figure 15, we show the weights for the Higgs jet vs. sgluon jet classification. The peak of $w_{\rm trim}$ around $R = \hat{R}_{b\bar{b}}$ is small as the hard substructures of Higgs jet and sgluon jet are (almost) the same. However, a sgluon is more radiative than a Higgs boson, and $w_{\rm soft}$ is negative in the entire region of $R < 1.5$.

As described above, weights $w_{\rm trim}$ and $w_{\rm soft}$ may take large values, but it does not necessarily mean that the corresponding $S_{2,\rm trim}$ and $S_{2,\rm soft}$ contribute dominantly in the jet classification. The energy scaling factors on the $S_{2,\rm trim}$ ($S_{2,\rm soft}$) and its weight $w_{\rm trim}$ ($w_{\rm soft}$) cancel out in the quantity of our interest $h = \sum_k (S^k_{2,\rm trim} w^k_{\rm trim} + S^k_{2,\rm soft} w^k_{\rm soft})$. For example, $\mathcal{O}[1]$ terms in $S_{2,\rm trim}$ and $\mathcal{O}[f_{\rm trim}]$ terms in $S_{2,\rm soft}$ contribute equally to the classifier if $w_{\rm soft}$ is around $f_{\rm trim}\, w_{\rm trim}$. In the left panel of figure 16, we draw the mean values $\langle S^k_{2,\rm trim} w^k_{\rm trim} \rangle$ and $\langle S^k_{2,\rm soft} w^k_{\rm soft} \rangle$ of Higgs jet vs. QCD jet classification, which are more directly related to the jet classification. The solid and dashed red lines correspond to the distributions of the Higgs jet, while the solid and dashed blue lines are for the QCD jet. The regions where Higgs jet and QCD jet distributions differ significantly are important for the network predictions. We find $\langle S^k_{2,\rm trim} w^k_{\rm trim} \rangle$ around $R \sim \hat{R}_{b\bar{b}}$ and $\langle S^k_{2,\rm soft} w^k_{\rm soft} \rangle$ in the region $R < 1.2$ mostly contribute to the jet classification.

The average distribution may not illustrate all the features of the classifier performance. The energy deposits in each bin fluctuate, and the bins with hits higher than the average

JHEP07(2019)135

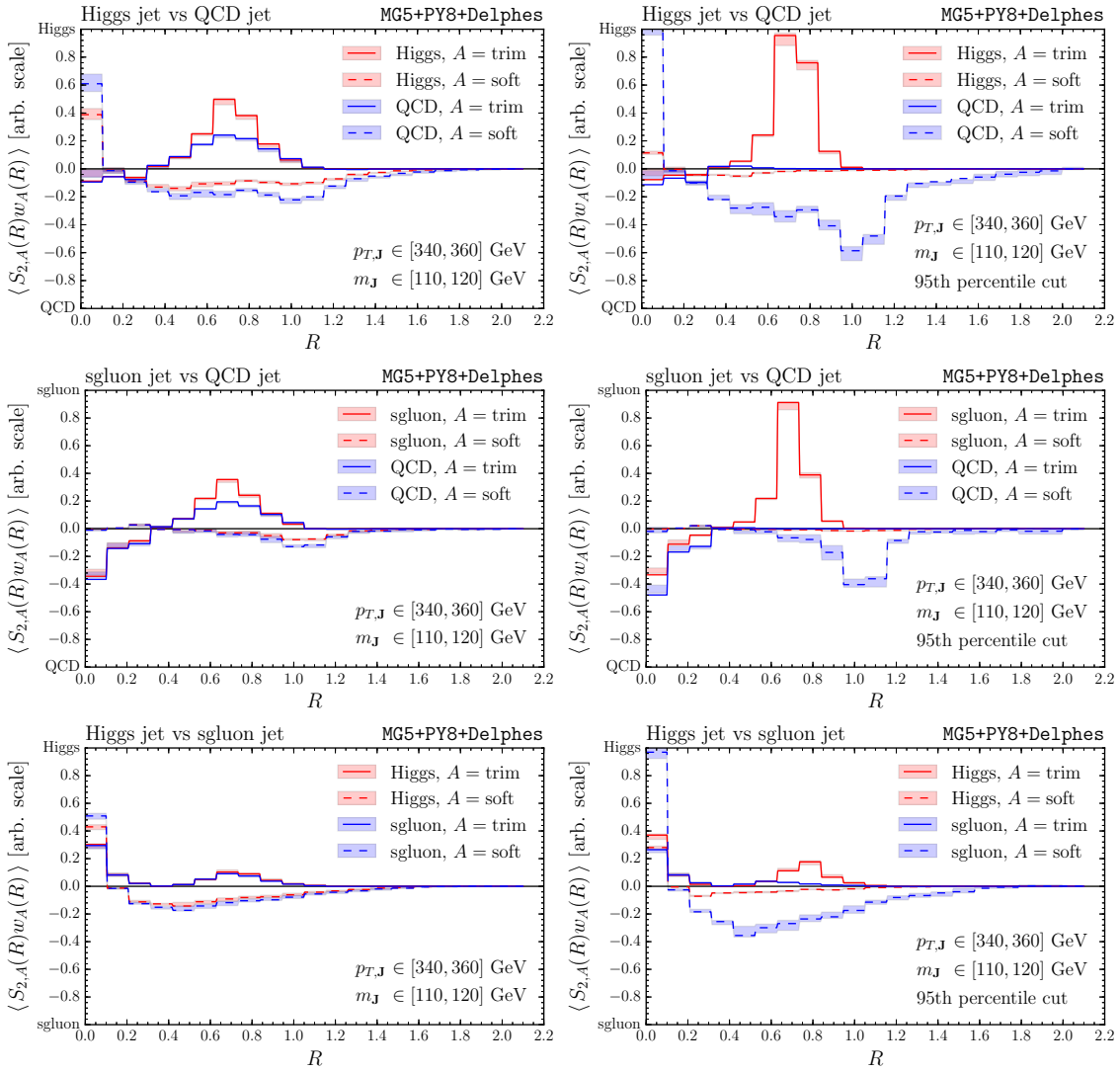

Figure 16. The distribution of the mean value $\langle S_{2,\mathrm{trim}}(R)\, w_{\mathrm{trim}}(R)\rangle$ (solid) and $\langle S_{2,\mathrm{soft}}(R) \cdot w_{\mathrm{soft}}(R)\rangle$ (dashed). We show the $\langle S_{2,A}(R)\, w_A(R)\rangle$ for Higgs jet vs. QCD jet (top), sgluon jet vs. QCD jet (center), and Higgs jet vs. sgluon jet (bottom) classifications. In the right figure, we additionally demand that $\hat{y}_h$ of the Higgs jet and $\hat{y}_{\mathrm{QCD}}$ of the QCD jet are larger than their 95th percentile respectively. We show their statistical uncertainty from the training samples as colored bands.

value contribute more to the network decisions. For example, soft emissions outside the angle between the two hardest subjets are rare in the Higgs jet. Once there is large angle radiation outside the cone of hard subjets, the network is likely to identify the jet as a QCD jet. In the right panel of figure 16, we plot the $\langle S^k_{2,\mathrm{trim}} w^k_{\mathrm{trim}}\rangle$ and $\langle S^k_{2,\mathrm{soft}} w^k_{\mathrm{soft}}\rangle$ distributions of the Higgs jet (QCD jet) with $\hat{y}_h$ ($\hat{y}_{\mathrm{QCD}}$) higher than the 95th percentile. The distributions indicate that the selected Higgs jets are mostly classified because of the $S_{2,\mathrm{trim}}$ excess at $\hat{R} \sim R_{b\bar{b}}$, while the QCD jets are classified using $S_{2,\mathrm{soft}}$ excess above $R > 0.2$.

We now use the two-level architecture to compare `PY8` and `HW7`. As we have already shown in section 3, the performance of the classifier depends on the event generators

JHEP07(2019)135

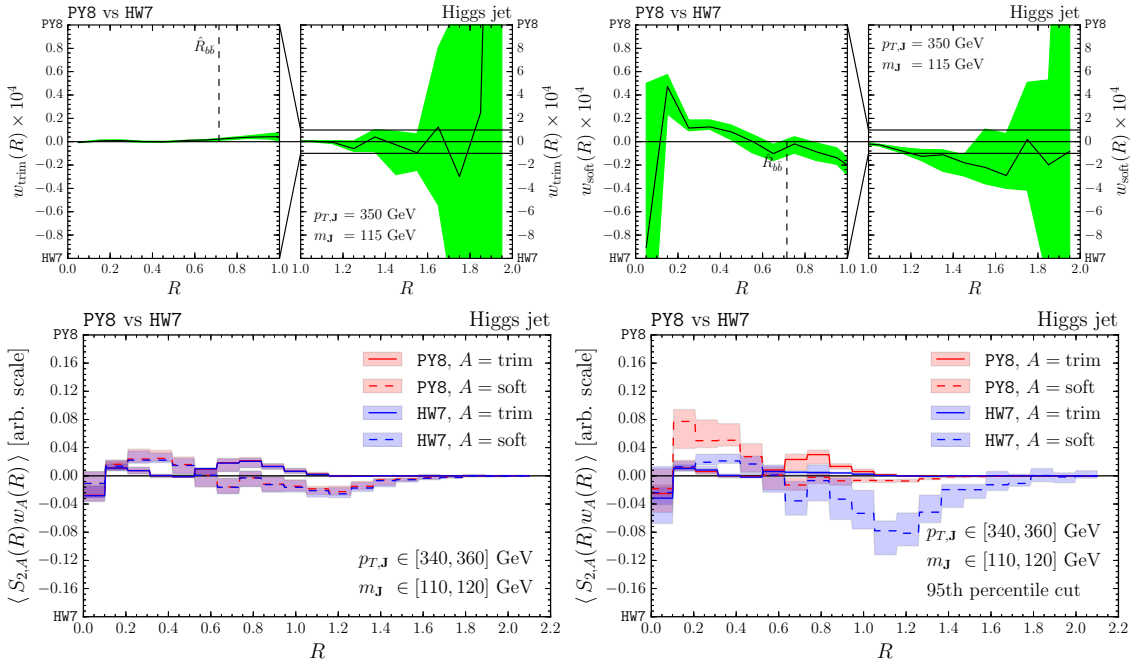

Figure 17. Top figures are the weights $w_{\rm trim}$ (left) and $w_{\rm soft}$ (right) at $p_{T,\mathbf{J}} = 350\,\text{GeV}$ and $m_{\mathbf{J}} = 115\,\text{GeV}$ for classifying Higgs jet of PY8 and HW7 events. Bottom figures are $\langle S_{2,A}(R)\, w_A(R) \rangle$. In the right bottom figure, we additionally demand that $\hat{y}_1$ of the PY8 generated jets and $\hat{y}_2$ of the HW7 generated jets are larger than their 95th percentile respectively. We show their statistical uncertainty from the training samples as colored bands.

significantly. We show the weights of the classifiers for $p_{T,\mathbf{J}} = 350\,\text{GeV}$ and $m_{\mathbf{J}} = 115\,\text{GeV}$ trained with Higgs jets in figure 17, sgluon jets in figure 18, and QCD jets in figure 19. For each plot, the signals are PY8 events, and the backgrounds are HW7 events. The $w_{\rm trim}$ in $R \lesssim R_{\mathbf{J}}$ is close to zero everywhere, representing $S_{2,\rm trim}$ spectra of PY8 and HW7 events are similar. It is not surprising because both PY8 and HW7 events from identical hard partons and these partons create the trimmed subjets inside the jet. On the other hand, the correlation involving constituents of the soft activities, $S_{2,\rm soft}$, is manifestly different and so $w_{\rm soft}$ is nonzero. For Higgs jet and sgluon jet, the $w_{\rm soft}$ distribution of PY8 events is significantly large (and positive) for $R \sim R_{\rm trim}$ and it decreases as R increases. The weight $w_{\rm soft}$ is negative for $R > R_{\mathbf{J}}$, which means that HW7 events have more soft activity in the region $R \gg \hat{R}_{b\bar{b}}$.

For the case of QCD jet, the distribution of $w_{\rm soft}$ is always positive and flat for $R < R_{\mathbf{J}}$ and negative for $R > 1.5$. It would be interesting to evaluate the weights for the classifiers trained with the experimental data and compare with the simulated results to tune the parameters of the event generators further.

5 Summary and outlook

The classification of jets with deep learning has gained significant attention in recent times. Majority of these analyses take advantage of the significant development in computing

JHEP07(2019)135

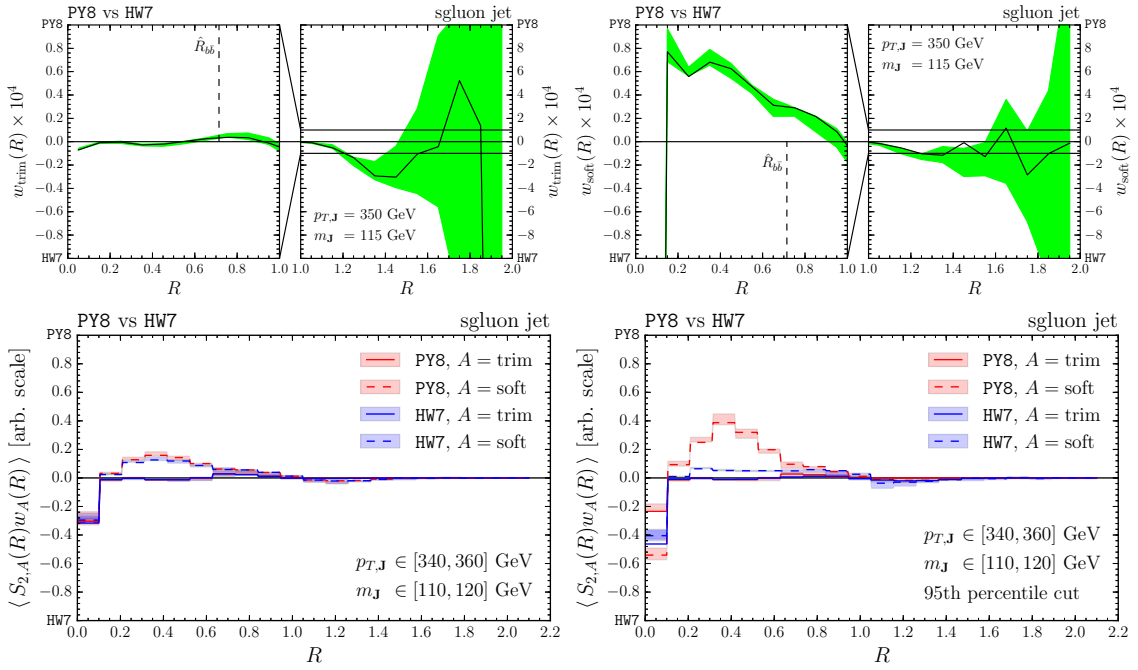

Figure 18. Top figures are the weights $w_{\rm trim}$ (left) and $w_{\rm soft}$ (right) at $p_{T,\mathbf{J}} = 350\,\text{GeV}$ and $m_{\mathbf{J}} = 115\,\text{GeV}$ for classifying sgluon jet of PY8 and HW7 events. Bottom figures are $\langle S_{2,A}(R)\, w_A(R)\rangle$. In the right bottom figure, we additionally demand that $\hat{y}_1$ of the PY8 generated jets and $\hat{y}_2$ of the HW7 generated jets are larger than their 95th percentile respectively. We show their statistical uncertainty from the training samples as colored bands.

power. These deep learning architectures utilize the complete event information in terms of low-level observables. These deep learning based strategies can be compared with the previous approaches for tagging jets, for example, mass drop tagger, n-subjettiness, energy correlation function where each of them has solid physics motivation.

In this paper, we introduce neural networks trained on "jet spectrum" $S_2(R)$, which are essentially two-point correlation functions of jet constituents. We also introduce $S_{2,\rm trim}(R)$, which is $S_2(R)$ calculated from the trimmed jet, to encode hard substructure of the jet. The difference $S_{2,\rm soft}(R) = S_2(R) - S_{2,\rm trim}(R)$ encodes the remaining correlations with soft radiations and is less affected by the correlations among the hard constituents. Our neural networks are trained on $S_{2,\rm trim}$ and $S_{2,\rm soft}$ integrated over certain bins. If the $S_{2,\rm trim}$ and $S_{2,\rm soft}$ spectra are multiplied by smooth functions and integrated over R, it forms an IRC safe C-correlator. This feature assures that the classifiers trained on binned $S_{2,\rm trim}$ and $S_{2,\rm soft}$ are approximately IRC safe.

The performance of MLP trained on $S_{2,\rm trim}$ and $S_{2,\rm soft}$ is compared to that of CNN trained on jet images. The CNN shows better performance than the MLP, but the difference is small. The key reason is efficient preprocessing of parton shower effects with less number of free parameters. Parton shower is the multiple splittings of the partons where each splitting is parametrized by the angular scale and momentum fraction of the partons. The binned $S_2(R)$ spectra collect the information of the parton splitting successfully. The spectra provide comparable jet classification performance with a fewer number of inputs.

JHEP07(2019)135

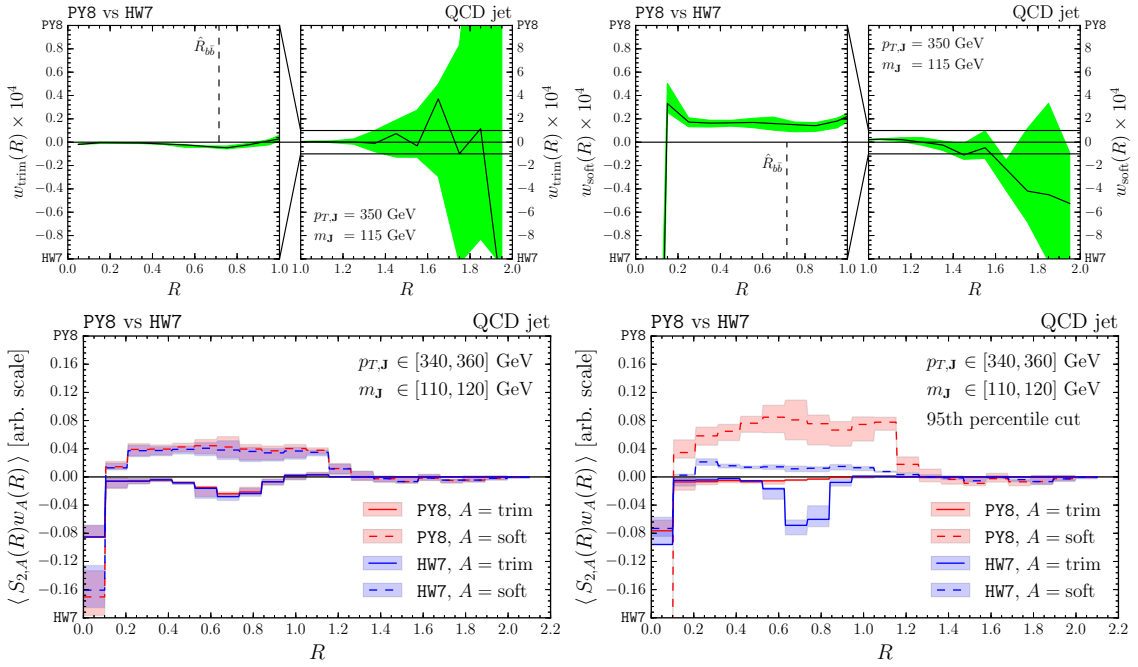

Figure 19. Top figures are the weights $w_{\rm trim}$ (left) and $w_{\rm soft}$ (right) at $p_{T,\mathbf{J}} = 350\,\text{GeV}$ and $m_{\mathbf{J}} = 115\,\text{GeV}$ for classifying QCD jet of PY8 and HW7 events. Bottom figures are $\langle S_{2,A}(R)\, w_A(R)\rangle$. In the right bottom figure, we additionally demand that $\hat{y}_1$ of the PY8 generated jets and $\hat{y}_2$ of the HW7 generated jets are larger than their 95th percentile respectively. We show their statistical uncertainty from the training samples as colored bands.

Furthermore, the MLP is computationally economical than the CNN because the MLP has smaller complexity than the CNN and takes only $\mathcal{O}(40)$ inputs.

The $S_{2,\rm trim}(R)$ and $S_{2,\rm soft}(R)$ spectra can be obtained from a functional Taylor series of an arbitrary classifier in energy flows. The spectra are basis vectors of the second order term in the expansion. In this context, the MLP trained on $S_{2,\rm trim}(R)$ and $S_{2,\rm soft}(R)$ can be considered as a sum of $2n$-linear C-correlators which can be reduced to products of the bilinear C-correlators in $S_{2,\rm trim}$ and $S_{2,\rm soft}$. The (mild) difference in the performance of the CNN and MLP comes from the remaining irreducible n-linear C-correlators.

The terms linear in $S_{2,\rm trim}(R)$ and $S_{2,\rm soft}(R)$ provide an opportunity to visualize and interpret the network predictions; therefore, we study a novel two-level architecture that involves an interpretable layer of a single node in the form of a C-correlator. The output is the sum of the product of the trained weights ($w_{\rm trim}$ and $w_{\rm soft}$) and jet spectra ($S_{2,\rm trim}$ and $S_{2,\rm soft}$). The absolute values of the weights signify the impact of the corresponding $S_{2,\rm trim}$ and $S_{2,\rm soft}$ bin values on the jet classification. In the context of classification between Higgs jet and QCD jet, the distribution of $w_{\rm trim}$ shows that $S_{2,\rm trim}$ spectrum around $\hat{R}_{b\bar{b}} = 2m_h/p_{T,\mathbf{J}}$ increases the output, and the classifier regards the jet as a Higgs jet. We have also shown that the dependence of $w_{\rm trim}$ on jet p_T can be qualitatively understood (at the parton level) from the decay of a boosted Higgs boson. In short, the network is using $S_{2,\rm trim}$ inputs to obtain the core substructure information inside the jet.

The soft activity is also useful for Higgs jet vs. QCD jet classification. The probability for assigning the given jet as a QCD jet increases with increase in $S_{2,\rm soft}$ on $R > R_{\rm trim}$. To

JHEP07(2019)135

study the impact of soft physics in jet classification, we also introduce sgluon, a hypothetical color octet scalar, and compare the classifier performance among Higgs jet, sgluon jet, and QCD jet. The network predictions for sgluon jet vs. QCD jet classification are primarily determined by the core substructure information as expected. However, the network uses the difference in the $S_{2,\mathrm{soft}}$ spectra arising from the different color structure of the decaying particle for Higgs jet vs. sgluon jet classification.

The non-trivial role of soft radiations in the predictions of the classifiers implies the results are highly sensitive to the choice of event generators. The weights associated with the $S_{2,\mathrm{trim}}$ are almost insensitive to the choice; however, the weights of the $S_{2,\mathrm{soft}}$ are strongly affected. This behavior is expected as modeling of soft physics is quite different in `Pythia 8` and `Herwig 7`.

The two-point correlation spectra and the architectures introduced in this paper can be applied for solving other interesting problems, thanks to flexibility on designing neural network. For jets with more complex substructures, e.g., top jet, the higher order terms in the energy flow series expansion may be included. It would be worthwhile to study the classifier performances when the network is trained with the experimental data and compare with the predictions of event generators to tune their parameters to reduce the uncertainty in modeling the soft physics. It is also interesting to use this interpretable architecture as a model-agnostic interpreter for black box architectures [88]. We leave these possibilities for future works.

Acknowledgments

We thank the organizers of *Beyond the BSM* and *Machine Learning for Jet Physics 2018* workshops. This work is supported by the Grant-in-Aid for Scientific Research on Scientific Research B (No. 16H03991, 17H02878) and Innovative Areas (16H06492), and by World Premier International Research Center Initiative (WPI Initiative), MEXT, Japan.

A Event generation and reconstruction

The parton level event samples, namely $pp \to Zj$, $pp \to Zh$ and $pp \to Z\sigma$ events, are generated at the leading order in QCD using `MadGraph5_aMC@NLO 2.6.1` [69]. We force the Higgs boson (h) and the sgluon (σ) to decay to a pair of bottom quarks, while Z boson to decay invisibly. For sgluon, we use a UFO model in [70, 71] with the following interaction term for the decay,

$$\mathcal{L}_{\mathrm{sgluon}} \ni y_{\sigma b\bar{b}}\, \sigma^a\, \bar{b} T^a b + \mathrm{h.c.} \tag{A.1}$$

The parton distribution function (PDF) set NNPDF 2.3 LO at $\alpha_S(m_Z) = 0.130$ [89] is used. To generate Higgs jets and Sgluon jets, we impose a parton level selection criterion on the Z boson transverse momentum, $p_{T,Z} > 250\,\mathrm{GeV}$. We simulate approximately 3 million events of $pp \to Zh$ and $pp \to Z\sigma$ processes and 18 million events of $pp \to Zj$.

We use two parton shower and hadronization simulators to compare the results. Namely, we use `Pythia 8.226` [65] with Monash tune [90] and `Herwig 7.1.3` [66, 67] with default

JHEP07(2019)135

tune [91, 92]. The shower starting scale is $H_T/2$ for $pp \to Zh$ and $pp \to Z\sigma$ processes and $p_{T,j}$ for $pp \to Zj$ process, where H_T is the transverse energy sum of the produced partons. The effects of underlying events and multi-parton interactions are taken into account, but we neglect the contaminations coming from the pile-ups. The PDF set for simulating all these effects are the same as that in the parton level simulation.

We use `Delphes 3.4.1` [74] with its default ATLAS configuration for fast detector simulations. Jets are reconstructed from the calorimeter towers using `FastJet 3.3.0` [75, 76] with anti-k_T algorithm [77] and jet radius parameter $R_{\mathbf{J}} = 1$. The leading jet of each event with $p_{T,\mathbf{J}} \in [300, 400]\,\text{GeV}$ and $m_{\mathbf{J}} \in [100, 150]\,\text{GeV}$ is selected. Since the scale $H_T/2$ for $pp \to Zh$ and $pp \to Z\sigma$ is higher than $p_{T,h}$ and $p_{T,\sigma}$, respectively, there is a chance that the leading jet is from the initial state radiation rather than from the decay of Higgs boson or sgluon. To filter out such jets from the Higgs jet and sgluon jet samples, we require that b-partons produced from the decay are within $R_{\mathbf{J}}$ from the leading jet axis.

B Oversampling and $p_{T,\mathbf{J}}$-bias removal

The neural networks may learn the inherent $p_{T,\mathbf{J}}$ difference among Higgs jet, sgluon jet, and QCD jet in figure 4 to classify them instead of learning the difference in their substructures. To penalize the learning from the $p_{T,\mathbf{J}}$ distribution, we augment the training and validation samples by oversampling as follows,

1. The samples (of each class) are binned in $p_{T,\mathbf{J}}$ with bin-width 1 GeV with b_i entries in the i-th bin and $b_{\max} = \max\{b_i\}$.

2. For each bin, oversample the events so that the number of the bin contents becomes a certain value $n_{\max} = c_d\, b_{\max}$. This oversampling is identical to repeating events in bin i sequentially for
$$M_i = \text{ceil}\left(\frac{c_d \cdot b_{\max}}{b_i}\right) \tag{B.1}$$
times and stop the oversampling when the number of events in the bin reaches $n_{\max}$. It may introduce a small bias due to unequal oversampling among different bins. We choose $c_d = 2$ and ignore the small residual bias.

In the upper (lower) panel of figure 20, we show the conditional probability density of the predicted class vector $\hat{y}_h$ ($\hat{y}_\sigma$) for a given p_T. The probability density has a mild dependence on $p_{T,\mathbf{J}}$ which is originating from the interplay of the phase-space selection and the jet radius parameter.

C Jet image and convolution neural network

We obtain and pre-process the jet image as follows,[4]

1. Recluster the jet constituents by k_T algorithm with a jet radius parameter $R_{\mathbf{J}} = 0.2$.

2. Set the center of (η, ϕ) coordinate to the leading (in p_T) subjet.

[4]This setup is similar to [21].

JHEP07(2019)135

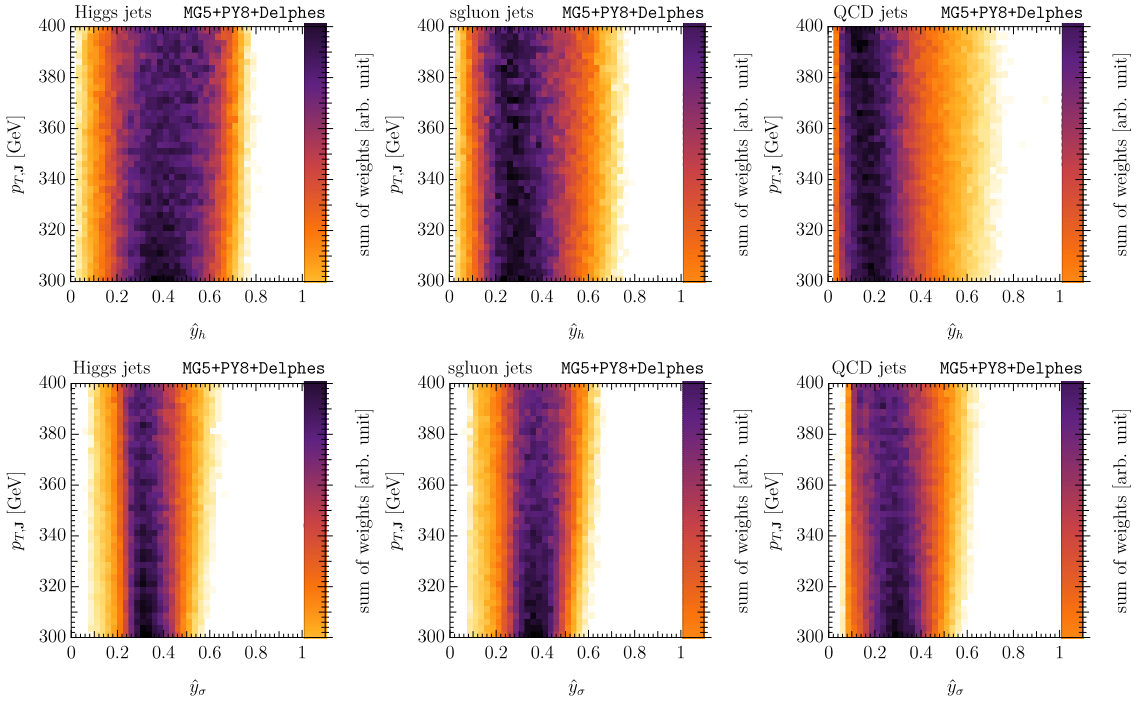

Figure 20. Validation of the p_T-bias removal on the training samples. Each row is a conditional probability density on a given $p_{T,\mathbf{J}}$ range, i.e., the sum of each row is 1. Upper (lower) panel displays the conditional probability density histograms of the predicted class vector $\hat{y}_h$ ($\hat{y}_\sigma$) for the Higgs jet (left), sgluon jet (center), and QCD jet (right).

3. If a second leading subjet is found, rotate the jet constituents on (η, ϕ) plane about the jet center so that the sub-leading jet is on the positive y-axis.

4. If a third leading subjet is found, flip the image about y-axis when x coordinate of the subjet is negative.

5. Select jet constituents within $[-1.5, 1.5] \otimes [-1.5, 1.5]$.

6. Finally, pixelate the jet constituents with pixel size 0.1×0.1. The (k, l)-th pixel intensity $P_T^{k,l}$ is determined by the total transverse energy of the jet constituents present in a given pixel, i.e.,

$$P_T^{k,l} = \sum_{i \in \mathbf{J}} p_{T,i} I_{\mathrm{bin}_{k,l}}(\vec{R}_i) = \int_{\mathrm{bin}_{k,l}} d\vec{R}\, P_T(\vec{R}) \tag{C.1}$$

   where $\mathrm{bin}_{k,l}$ is the region of (k, l)-th bin.

7. Standardize all the $P_T^{k,l}$.

This jet image is analyzed by a CNN which consists of two-dimensional convolutional layers (CONV) and max-pooling layers (figure 21). In particular, we use the following CNN setup,

JHEP07(2019)135

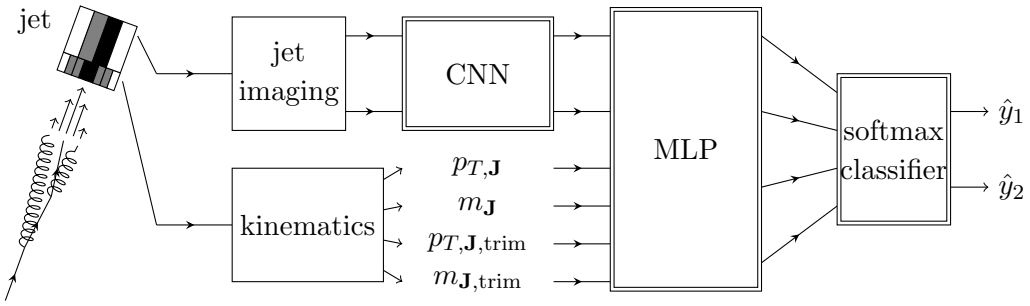

Figure 21. A schematic diagram of a convolutional neural network trained on jet image. The double bordered boxes represent trainable modules.

- Layer 1: convolutional layer with 64 filters with kernel size 3×3, and ReLU activation,
- Layer 2: max-pooling layer with pool size 2×2,
- Layer 3: convolutional layer with 32 filters with kernel size 4×4, and ReLU activation,
- Layer 4: max-pooling with pool size 2×2.

The first convolutional layer deals with angular scale up to 0.3 to treat collinear radiations while the second convolutional layer operates up to 0.8. The outputs of Layer 4 are flattened into a one-dimensional array and concatenated with a set of kinematic inputs, $\{p_{T,\mathbf{J}}, m_{\mathbf{J}}, p_{T,\mathbf{J},\text{trim}}, m_{\mathbf{J},\text{trim}}\}$. The flattened output array is fed into an MLP with two hidden layers with 300, 100 filters respectively, and ReLU activation function. The outputs of the MLP are fed into a softmax layer to make a prediction. The training setup is the same as in section 3.

References

[1] J.M. Butterworth, A.R. Davison, M. Rubin and G.P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, *Phys. Rev. Lett.* **100** (2008) 242001 [arXiv:0802.2470] [INSPIRE].

[2] J. Thaler and L.-T. Wang, *Strategies to Identify Boosted Tops*, *JHEP* **07** (2008) 092 [arXiv:0806.0023] [INSPIRE].

[3] D.E. Kaplan, K. Rehermann, M.D. Schwartz and B. Tweedie, *Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks*, *Phys. Rev. Lett.* **101** (2008) 142001 [arXiv:0806.0848] [INSPIRE].

[4] T. Plehn, G.P. Salam and M. Spannowsky, *Fat Jets for a Light Higgs*, *Phys. Rev. Lett.* **104** (2010) 111801 [arXiv:0910.5472] [INSPIRE].

[5] T. Plehn, M. Spannowsky, M. Takeuchi and D. Zerwas, *Stop Reconstruction with Tagged Tops*, *JHEP* **10** (2010) 078 [arXiv:1006.2833] [INSPIRE].

JHEP07(2019)135

[6] D.E. Soper and M. Spannowsky, *Finding physics signals with shower deconstruction*, *Phys. Rev.* **D 84** (2011) 074002 [arXiv:1102.3480] [INSPIRE].

[7] D.E. Soper and M. Spannowsky, *Finding top quarks with shower deconstruction*, *Phys. Rev.* **D 87** (2013) 054012 [arXiv:1211.3140] [INSPIRE].

[8] M. Dasgupta, A. Fregoso, S. Marzani and G.P. Salam, *Towards an understanding of jet substructure*, *JHEP* **09** (2013) 029 [arXiv:1307.0007] [INSPIRE].

[9] D.E. Soper and M. Spannowsky, *Finding physics signals with event deconstruction*, *Phys. Rev.* **D 89** (2014) 094005 [arXiv:1402.1189] [INSPIRE].

[10] A.J. Larkoski, S. Marzani, G. Soyez and J. Thaler, *Soft Drop*, *JHEP* **05** (2014) 146 [arXiv:1402.2657] [INSPIRE].

[11] J. Gallicchio and M.D. Schwartz, *Seeing in Color: Jet Superstructure*, *Phys. Rev. Lett.* **105** (2010) 022001 [arXiv:1001.5027] [INSPIRE].

[12] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, *JHEP* **03** (2011) 015 [arXiv:1011.2268] [INSPIRE].

[13] J. Gallicchio and M.D. Schwartz, *Quark and Gluon Tagging at the LHC*, *Phys. Rev. Lett.* **107** (2011) 172001 [arXiv:1106.3076] [INSPIRE].

[14] Y.-T. Chien, *Telescoping jets: Probing hadronic event structure with multiple R's*, *Phys. Rev.* **D 90** (2014) 054008 [arXiv:1304.5240] [INSPIRE].

[15] A.J. Larkoski, G.P. Salam and J. Thaler, *Energy Correlation Functions for Jet Substructure*, *JHEP* **06** (2013) 108 [arXiv:1305.0007] [INSPIRE].

[16] A.J. Larkoski, I. Moult and D. Neill, *Power Counting to Better Jet Observables*, *JHEP* **12** (2014) 009 [arXiv:1409.6298] [INSPIRE].

[17] I. Moult, L. Necib and J. Thaler, *New Angles on Energy Correlation Functions*, *JHEP* **12** (2016) 153 [arXiv:1609.07483] [INSPIRE].

[18] L.G. Almeida, M. Backović, M. Cliche, S.J. Lee and M. Perelstein, *Playing Tag with ANN: Boosted Top Identification with Pattern Recognition*, *JHEP* **07** (2015) 086 [arXiv:1501.05968] [INSPIRE].

[19] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman and A. Schwartzman, *Jet-images — deep learning edition*, *JHEP* **07** (2016) 069 [arXiv:1511.05190] [INSPIRE].

[20] P.T. Komiske, E.M. Metodiev and M.D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, *JHEP* **01** (2017) 110 [arXiv:1612.01551] [INSPIRE].

[21] G. Kasieczka, T. Plehn, M. Russell and T. Schell, *Deep-learning Top Taggers or The End of QCD?*, *JHEP* **05** (2017) 006 [arXiv:1701.08784] [INSPIRE].

[22] G. Louppe, K. Cho, C. Becot and K. Cranmer, *QCD-Aware Recursive Neural Networks for Jet Physics*, *JHEP* **01** (2019) 057 [arXiv:1702.00748] [INSPIRE].

[23] P.T. Komiske, E.M. Metodiev, B. Nachman and M.D. Schwartz, *Pileup Mitigation with Machine Learning (PUMML)*, *JHEP* **12** (2017) 051 [arXiv:1707.08600] [INSPIRE].

[24] A. Butter, G. Kasieczka, T. Plehn and M. Russell, *Deep-learned Top Tagging with a Lorentz Layer*, *SciPost Phys.* **5** (2018) 028 [arXiv:1707.08966] [INSPIRE].

[25] T. Cheng, *Recursive Neural Networks in Quark/Gluon Tagging*, *Comput. Softw. Big Sci.* **2** (2018) 3 [arXiv:1711.02633] [INSPIRE].

JHEP07(2019)135

[26] A. Andreassen, I. Feige, C. Frye and M.D. Schwartz, *JUNIPR: a Framework for Unsupervised Machine Learning in Particle Physics*, *Eur. Phys. J.* **C 79** (2019) 102 [arXiv:1804.09720] [INSPIRE].

[27] S. Choi, S.J. Lee and M. Perelstein, *Infrared Safety of a Neural-Net Top Tagging Algorithm*, *JHEP* **02** (2019) 132 [arXiv:1806.01263] [INSPIRE].

[28] S.H. Lim and M.M. Nojiri, *Spectral Analysis of Jet Substructure with Neural Networks: Boosted Higgs Case*, *JHEP* **10** (2018) 181 [arXiv:1807.03312] [INSPIRE].

[29] F.A. Dreyer, G.P. Salam and G. Soyez, *The Lund Jet Plane*, *JHEP* **12** (2018) 064 [arXiv:1807.04758] [INSPIRE].

[30] J. Lin, M. Freytsis, I. Moult and B. Nachman, *Boosting $H \to b\bar{b}$ with Machine Learning*, *JHEP* **10** (2018) 101 [arXiv:1807.10768] [INSPIRE].

[31] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy Flow Networks: Deep Sets for Particle Jets*, *JHEP* **01** (2019) 121 [arXiv:1810.05165] [INSPIRE].

[32] J. Arjona Martínez, O. Cerri, M. Pierini, M. Spiropulu and J.-R. Vlimant, *Pileup mitigation at the Large Hadron Collider with Graph Neural Networks*, *Eur. Phys. J. Plus* **134** (2019) 333 [arXiv:1810.07988] [INSPIRE].

[33] G. Kasieczka, N. Kiefer, T. Plehn and J.M. Thompson, *Quark-gluon Tagging: Machine Learning vs. Detector*, *SciPost Phys.* **6** (2019) 069 [arXiv:1812.09223] [INSPIRE].

[34] H. Qu and L. Gouskos, *ParticleNet: Jet Tagging via Particle Clouds*, arXiv:1902.08570 [INSPIRE].

[35] A.J. Larkoski, I. Moult and B. Nachman, *Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning*, arXiv:1709.04464 [INSPIRE].

[36] L. Asquith et al., *Jet Substructure at the Large Hadron Collider: Experimental Review*, arXiv:1803.06991 [INSPIRE].

[37] L.M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, *Weakly Supervised Classification in High Energy Physics*, *JHEP* **05** (2017) 145 [arXiv:1702.00414] [INSPIRE].

[38] T. Cohen, M. Freytsis and B. Ostdiek, *(Machine) Learning to Do More with Less*, *JHEP* **02** (2018) 034 [arXiv:1706.09451] [INSPIRE].

[39] E.M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, *JHEP* **10** (2017) 174 [arXiv:1708.02949] [INSPIRE].

[40] G. Louppe, M. Kagan and K. Cranmer, *Learning to Pivot with Adversarial Networks*, arXiv:1611.01046 [INSPIRE].

[41] C. Shimmin et al., *Decorrelated Jet Substructure Tagging using Adversarial Neural Networks*, *Phys. Rev.* **D 96** (2017) 074034 [arXiv:1703.03507] [INSPIRE].

[42] A. Chakraborty, A.M. Iyer and T.S. Roy, *A Framework for Finding Anomalous Objects at the LHC*, *Nucl. Phys.* **B 932** (2018) 439 [arXiv:1707.07084] [INSPIRE].

[43] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty Detection Meets Collider Physics*, arXiv:1807.10261 [INSPIRE].

[44] T. Heimel, G. Kasieczka, T. Plehn and J.M. Thompson, *QCD or What?*, *SciPost Phys.* **6** (2019) 030 [arXiv:1808.08979] [INSPIRE].

JHEP07(2019)135

[45] M. Farina, Y. Nakai and D. Shih, *Searching for New Physics with Deep Autoencoders*, arXiv:1808.08992 [INSPIRE].

[46] O. Cerri, T.Q. Nguyen, M. Pierini, M. Spiropulu and J.-R. Vlimant, *Variational Autoencoders for New Physics Mining at the Large Hadron Collider*, *JHEP* **05** (2019) 036 [arXiv:1811.10276] [INSPIRE].

[47] T.S. Roy and A.H. Vijay, *A robust anomaly finder based on autoencoder*, arXiv:1903.02032 [INSPIRE].

[48] T. Roxlo and M. Reece, *Opening the black box of neural nets: case studies in stop/top discrimination*, arXiv:1804.09278 [INSPIRE].

[49] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, *Constraining Effective Field Theories with Machine Learning*, *Phys. Rev. Lett.* **121** (2018) 111801 [arXiv:1805.00013] [INSPIRE].

[50] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, *A Guide to Constraining Effective Field Theories with Machine Learning*, *Phys. Rev.* **D 98** (2018) 052004 [arXiv:1805.00020] [INSPIRE].

[51] J. Guo, J. Li, T. Li, F. Xu and W. Zhang, *Deep learning for R-parity violating supersymmetry searches at the LHC*, *Phys. Rev.* **D 98** (2018) 076017 [arXiv:1805.10730] [INSPIRE].

[52] J.H. Collins, K. Howe and B. Nachman, *Anomaly Detection for Resonant New Physics with Machine Learning*, *Phys. Rev. Lett.* **121** (2018) 241803 [arXiv:1805.02664] [INSPIRE].

[53] R.T. D'Agnolo and A. Wulzer, *Learning New Physics from a Machine*, *Phys. Rev.* **D 99** (2019) 015014 [arXiv:1806.02350] [INSPIRE].

[54] A. De Simone and T. Jacques, *Guiding New Physics Searches with Unsupervised Learning*, *Eur. Phys. J.* **C 79** (2019) 289 [arXiv:1807.06038] [INSPIRE].

[55] C. Englert, P. Galler, A. Pilkington and M. Spannowsky, *Approaching robust EFT limits for CP-violation in the Higgs sector*, *Phys. Rev.* **D 99** (2019) 095007 [arXiv:1901.05982] [INSPIRE].

[56] J.H. Collins, K. Howe and B. Nachman, *Extending the search for new resonances with machine learning*, *Phys. Rev.* **D 99** (2019) 014038 [arXiv:1902.02634] [INSPIRE].

[57] Z.C. Lipton, *The mythos of model interpretability*, in proceedings of the *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, U.S.A., 23 June 2016, arXiv:1606.03490.

[58] K. Datta, A. Larkoski and B. Nachman, *Automating the Construction of Jet Observables with Machine Learning*, arXiv:1902.07180 [INSPIRE].

[59] D. Krohn, J. Thaler and L.-T. Wang, *Jet Trimming*, *JHEP* **02** (2010) 084 [arXiv:0912.1342] [INSPIRE].

[60] F.V. Tkachov, *Measuring multijet structure of hadronic energy flow or, what is a jet?*, *Int. J. Mod. Phys.* **A 12** (1997) 5411 [hep-ph/9601308] [INSPIRE].

[61] M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam and P. Vandergheynst, *Geometric Deep Learning: Going beyond Euclidean data*, *IEEE Sig. Proc. Mag.* **34** (2017) 18 [arXiv:1611.08097] [INSPIRE].

[62] J. Cogan, M. Kagan, E. Strauss and A. Schwartzman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, *JHEP* **02** (2015) 118 [arXiv:1407.5675] [INSPIRE].

JHEP07(2019)135

[63] S.D. Ellis, C.K. Vermilion and J.R. Walsh, *Techniques for improved heavy particle searches with jet substructure*, *Phys. Rev.* **D 80** (2009) 051501 [arXiv:0903.5081] [INSPIRE].

[64] S.D. Ellis, C.K. Vermilion and J.R. Walsh, *Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches*, *Phys. Rev.* **D 81** (2010) 094023 [arXiv:0912.0033] [INSPIRE].

[65] T. Sjöstrand et al., *An Introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012] [INSPIRE].

[66] J. Bellm et al., *HERWIG 7.0/HERWIG++ 3.0 release note*, *Eur. Phys. J.* **C 76** (2016) 196 [arXiv:1512.01178] [INSPIRE].

[67] M. Bahr et al., *HERWIG++ Physics and Manual*, *Eur. Phys. J.* **C 58** (2008) 639 [arXiv:0803.0883] [INSPIRE].

[68] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy flow polynomials: A complete linear basis for jet substructure*, *JHEP* **04** (2018) 013 [arXiv:1712.07124] [INSPIRE].

[69] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections and their matching to parton shower simulations*, *JHEP* **07** (2014) 079 [arXiv:1405.0301] [INSPIRE].

[70] C. Degrande, B. Fuks, V. Hirschi, J. Proudom and H.-S. Shao, *Automated next-to-leading order predictions for new physics at the LHC: the case of colored scalar pair production*, *Phys. Rev.* **D 91** (2015) 094005 [arXiv:1412.5589] [INSPIRE].

[71] *FeynRules models to be used for NLO calculations with aMC@NLO*, (2019) https://feynrules.irmp.ucl.ac.be/wiki/NLOModels.

[72] A. Alloul, N.D. Christensen, C. Degrande, C. Duhr and B. Fuks, *FeynRules 2.0 — A complete toolbox for tree-level phenomenology*, *Comput. Phys. Commun.* **185** (2014) 2250 [arXiv:1310.1921] [INSPIRE].

[73] C. Degrande, C. Duhr, B. Fuks, D. Grellscheid, O. Mattelaer and T. Reiter, *UFO — The Universal FeynRules Output*, *Comput. Phys. Commun.* **183** (2012) 1201 [arXiv:1108.2040] [INSPIRE].

[74] DELPHES 3 collaboration, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [arXiv:1307.6346] [INSPIRE].

[75] M. Cacciari, G.P. Salam and G. Soyez, *FastJet User Manual*, *Eur. Phys. J.* **C 72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

[76] M. Cacciari and G.P. Salam, *Dispelling the N^3 myth for the k_t jet-finder*, *Phys. Lett.* **B 641** (2006) 57 [hep-ph/0512210] [INSPIRE].

[77] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063 [arXiv:0802.1189] [INSPIRE].

[78] B.R. Webber, *QCD Jets and Parton Showers*, in proceedings of the *Gribov-80 Memorial Workshop on Quantum Chromodynamics and Beyond*, Trieste, Italy, 26–28 May 2010, pp. 82–92 [https://doi.org/10.1142/9789814350198_0010] [arXiv:1009.5871] [INSPIRE].

[79] B. Bhattacherjee, S. Mukhopadhyay, M.M. Nojiri, Y. Sakaki and B.R. Webber, *Associated jet and subjet rates in light-quark and gluon jet discrimination*, *JHEP* **04** (2015) 131 [arXiv:1501.04794] [INSPIRE].

JHEP07(2019)135

[80] A. Krogh and J.A. Hertz, *A simple weight decay can improve generalization*, in proceedings of the *Advances in Neural Information Processing Systems 4 (NIPS 1991)*, Denver, Colorado, U.S.A., 2–5 December 1991, J.E. Moody, S.J. Hanson and R.P. Lippmann eds., Morgan-Kaufmann (1992), pp. 950–957.

[81] D.P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, arXiv:1412.6980 [INSPIRE].

[82] X. Glorot and Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*, in proceedings of the *Thirteenth International Conference on Artificial Intelligence and Statistics*, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010, Y.W. Teh and M. Titterington eds., pp. 249–256 [*Proc. Mach. Learn. Res.* **9** (2010) 249].

[83] F. Chollet et al., *Keras*, (2015) https://keras.io.

[84] M. Abadi et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, (2015) https://www.tensorflow.org/.

[85] D. Alvarez Melis and T. Jaakkola, *Towards robust interpretability with self-explaining neural networks*, in proceedings of the *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, Montreal, Canada, 3–8 December 2018, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett eds., Curran Associates, Inc. (2018), pp. 7786–7795 [arXiv:1806.07538].

[86] D. Clevert, T. Unterthiner and S. Hochreiter, *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*, in proceedings of the *International Conference on Learning Representations (ICLR)*, Caribe Hilton, San Juan, Puerto Rico, 2–4 May 2016, arXiv:1511.07289.

[87] K. He, X. Zhang, S. Ren and J. Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, in proceedings of the *IEEE International Conference on Computer Vision (ICCV)*, Washington, D.C., U.S.A., 7–13 December 2015, arXiv:1502.01852 [INSPIRE].

[88] M.T. Ribeiro, S. Singh and C. Guestrin, *Model-Agnostic Interpretability of Machine Learning*, in proceedings of the *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, U.S.A., 23 June 2016, arXiv:1606.05386.

[89] R.D. Ball et al., *Parton distributions with LHC data*, *Nucl. Phys.* **B 867** (2013) 244 [arXiv:1207.1303] [INSPIRE].

[90] P. Skands, S. Carrazza and J. Rojo, *Tuning PYTHIA 8.1: the Monash 2013 Tune*, *Eur. Phys. J.* **C 74** (2014) 3024 [arXiv:1404.5630] [INSPIRE].

[91] HERWIG collaboration, *Minimum-bias and underlying-event tunes*, (2015) https://herwig.hepforge.org/tutorials/mpi/tunes.html.

[92] S. Gieseke, C. Rohr and A. Siodmok, *Colour reconnections in HERWIG++*, *Eur. Phys. J.* **C 72** (2012) 2225 [arXiv:1206.0041] [INSPIRE].

JHEP07(2019)135