Published for SISSA by Springer

Received: October 16, 2023
Accepted: December 31, 2023
Published: January 19, 2024

Hypergraphs in LHC phenomenology — the next frontier of IRC-safe feature extraction

Partha Konar,[a] **Vishal S. Ngairangbam**[a] **and Michael Spannowsky**[b,c]

[a]*Theoretical Physics Division, Physical Research Laboratory, Shree Pannalal Patel Marg, Ahmedabad — 380009, Gujarat, India*
[b]*Institute for Particle Physics Phenomenology, Durham University, Durham DH1 3LE, U.K.*
[c]*Department of Physics, Durham University, Durham DH1 3LE, U.K.*

E-mail: konar@prl.res.in, vishalng@prl.res.in, michael.spannowsky@durham.ac.uk

Abstract: In this study, we critically evaluate the approximation capabilities of existing infra-red and collinear (IRC) safe feature extraction algorithms, namely Energy Flow Networks (EFNs) and Energy-weighted Message Passing Networks (EMPNs). Our analysis reveals that these algorithms fall short in extracting features from any N-point correlation that isn't a power of two, based on the complete basis of IRC safe observables, specifically C-correlators. To address this limitation, we introduce the Hypergraph Energy-weighted Message Passing Networks (H-EMPNs), designed to capture any N-point correlation among particles efficiently. Using the case study of top vs. QCD jets, which holds significant information in its 3-point correlations, we demonstrate that H-EMPNs targeting up to N=3 correlations exhibit superior performance compared to EMPNs focusing on up to N=4 correlations within jet constituents.

Keywords: Jets and Jet Substructure, Top Quark

ArXiv ePrint: 2309.17351

https://doi.org/10.1007/JHEP01(2024)113

Contents

1 Introduction **1**

2 Universal approximation of IRC safe observables **3**
2.1 Energy Flow Networks 4
2.2 Energy-weighted message passing networks 4

3 Hypergraph energy-weighted message passing networks **6**
3.1 IRC safety with heterogenous source and destination embeddings 6
3.2 Building higher point IRC safe feature extractor 7

4 Network architecture and training **10**

5 Results **11**
5.1 Performance 11
5.2 Visualizing the latent graph representation 14

6 Conclusions **17**

1 Introduction

The Large Hadron Collider (LHC) has been a cornerstone in advancing our understanding of particle physics. However, the complexity of the data generated necessitates sophisticated methods for feature extraction and analysis. Traditional approaches often fail to capture intricate relationships among the data points, especially when considering infrared and collinear (IRC) safe observables. In this context, neural networks have shown promise [1–15] but are not without limitations. These include issues regarding interpretibility [16–22], uncertainty quantification [23–30], and a better handle and design of the physical biases [31–40] of the neural networks for better physics generalization capabilities. The intricate nature of the underlying physical description warrants a thorough understanding of these algorithms, particularly as a precise understanding of the Standard Model background within perturbative Quantum Chromodynamics (pQCD) is needed to discover new physics.

With the recorded events naturally represented as sets (of variable sizes) of different reconstructed particles or raw detector hits, point clouds are the natural representation of the recorded data, and architectures to process such data efficiently, particularly Graph Neural Networks [41–48], have been used successfully for LHC phenomenology. However, graphs do not expose higher-order correlations within the data by design, concentrating on two-particle correlations — the natural generalisation being hypergraphs. This generalisation is diagrammatically shown in figure 1 for a three-prong top jet where the graph's edges are defined in terms of two particles, while the order three hyperedges can look into the relevant three-prong structure of the top jet. This paper addresses these challenges by introducing Hypergraph Energy-weighted Message Passing Networks (H-EMPNs) designed to

JHEP01(2024)113

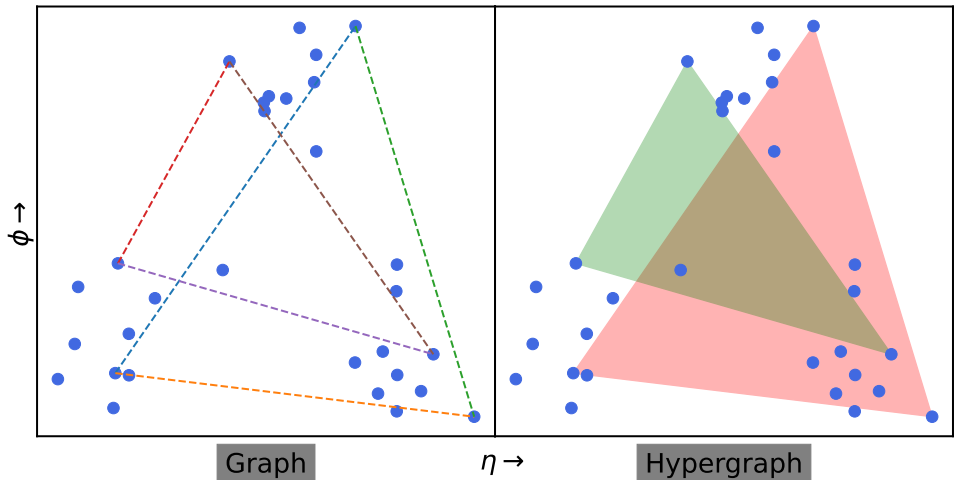

Figure 1. Visualisation of the inter-relations of jet constituents as captured by a graph structure (left) and a hypergraph structure with order-three hyperedges (right). In a graph structure, the edges correlate two constituents at a time and are shown as a line segment connecting two nodes. Instead, the order-three hyperedges simultaneously link properties of three jet constituents at a time and are shown as a triangle with vertices coinciding with three nodes. Thus, hypergraphs are more expressive structures and can access higher-order correlations amongst jet constituents.

extract three-particle correlations better than existing IRC-safe feature extractors. We first examine the universal approximation capabilities of existing infra-red and collinear safe neural network models like Energy Flow Networks (EFNs) [31] and Energy-weighted Message Passing Networks (EMPNs) [36] in approximating any IRC safe observable expressible in terms of C-correlators [49, 50] looking into any general N-body phase space. Finding that EFNs are restricted $N = 1$, and EMPNs have an arguably weak capability for approximating any $N \neq 2^n$ C-correlators, we present H-EMPN as a more robust and versatile model capable of efficiently approximating any general IRC safe observable for any general N. Our method leverages the power of message-passing in graphs and hypergraphs to capture higher-order relationships among the data points, thereby providing a more comprehensive feature extraction mechanism.

Restricting ourselves to $N = 3$ for the top vs QCD jet tagging scenario, where the dominant information lies in the 3-body decay phase space of the top quark, we find that H-EMPNs outperform EMPN, which look up to $N = 4$ interparticle correlations, confirming our initial observation. We demonstrate the efficacy of H-EMPNs through empirical tests to showcase the learned graph representations. Furthermore, we discuss the architectural nuances of H-EMPN, providing insights into its design and training procedures. By doing so, we aim to establish H-EMPN as a powerful tool for LHC phenomenology, opening new avenues for applications in collider phenomenology.

Specifically, in section 2, we discuss the universal approximation of any IRC safe observable by EFNs and EMPNs by taking its correspondence to any generic C-correlator. In section. 3, we devise H-EMPNs that can approximate any general C-correlator. The architecture and training details are presented in section 4, while the results are presented in section. 5. We conclude in section 6.

JHEP01(2024)113

Notation. In the following discussions, we are given the set of four vectors of the jet constituents

$$\mathcal{S} = \{\, p_1, p_2, \ldots, p_{n_{part.}} \},$$

with n_{part} being the number of constituents. These particles will be indexed via small Roman subscripts, while the number of message-passing operations will be indexed as Greek superscripts. Unless otherwise stated, all summations will be over the set $\mathcal{S}$. The four vectors are given in terms of the relative hardness $z_i = p_T^i / \sum p_T^j$ and the rapidity-azimuth variables $\hat{\mathbf{p}}_i = (y_i, \phi_i)$. Bold-faced alphabets like $\mathbf{h}_i$ and $\mathbf{G}$ denote vector quantities with their italicised counterparts h_i and G acting as a placeholder for a component. As we will consider inference on networks after training rather than the training itself, we will not explicitly write the dependence of function approximators on the tunable parameters. For instance, $\mathbf{g}^{(\alpha)}(\mathbf{h}_i^{(\alpha-1)}, \mathbf{h}_j^{(\alpha-1)})$ denotes a MultiLayer Perceptron (MLP) at the α^{th} message passing step, $\mathbf{h}_i^{(\alpha-1)}$ and $\mathbf{h}_j^{(\alpha-1)}$ correspond to the updated node features in the previous operation of particle i and j, respectively, in $\mathcal{S}$.

2 Universal approximation of IRC safe observables

In present scientific literature, it is well-known that MLPs are universal function approximators [51–53]. Without going into mathematical rigour, a parametrized function $f(\mathbf{x}, \Theta)$ of a vector $\mathbf{x}$ and tunable parameters Θ, is a universal approximator if it can approximate any continuous function up to any arbitrary precision in a compact domain and range. On the other hand, physical observables like momenta or position live in an underlying metric space, and notions of completeness have long been the bread-and-butter of physicists to study physical systems. The complete set of IRC safe observables is essential at the LHC and the subject of our present investigation. Any IRC safe observable $\mathcal{O}$ can be expanded in a basis of C-correlators [49] as

$$\mathcal{O} \approx \sum_{N=0}^{N_{\max}} \mathcal{C}_N^{f_N}, \quad \mathcal{C}_N^{f_N} = \sum_{i_1} \sum_{i_2} \ldots \sum_{i_N} E_{i_1} E_{i_2} \ldots E_{i_N} f_N(\hat{p}_{i_1}, \hat{p}_{i_2}, \ldots \hat{p}_{i_N}), \tag{2.1}$$

where f_N is symmetric to any permutation of its arguments. Energy Flow Polynomials (EFPs) [50] expand $\mathcal{O}$ in a basis of polynomials of energy using the Stone-Weierstrass approximation theorem. In this section, we take a look into the approximation capabilities of existing IRC safe neural networks, namely Energy Flow Networks [31], and Energy-weighted Message Passing Network (EMPN) [36], comparing the functional form to any arbitrary N in the basis of C-correlators. As the C-correlators are complete, the network-extracted observables would be expressible as a linear sum of different C-correlators, and we investigate the terms in the sum (as given in eq. (2.1)) that are optimally extracted via these observables.

Although we rely on the statement of universal approximation theorems, it is important to remember that we will strictly talk about the existence of such approximators and not concentrate on the method of finding such a function. However, presently available gradient descent algorithms are powerful enough to efficiently find an approximation given that we have the desired output value on a large enough number of samples. This numerical nature of finding a practical working point in the weight space is one of the significant concerns

JHEP01(2024)113

regarding the interpretability of neural networks in general. Our aim is not to tackle this more difficult problem but to systematically establish the capability of IRC-safe feature extractors based on their ability to approximate different C-correlators. Moreover, we concentrate on the extracted features rather than the final observable approximated by the complete network, i.e. we do not consider the function approximation done by the downstream MLP, which takes the extracted IRC safe features, as this would be akin to a usual multi-variate approach of physics motivated features.

As we will study the general behaviour of the approximated function whose weights are frozen after some training procedure, we will not discuss the explicit dependence of the neural networks on their tuneable parameters in the following discussions.

2.1 Energy Flow Networks

Energy Flow Networks are infra-red and collinear safe deep sets model which learns a per-particle map of each particle's directional coordinates $\hat{\mathbf{p}}_i$ and undergoes an energy-weighted sum to form a fixed length representation of any variable cardinality constituent set. Without loss of generality for a multi-dimensional representation, a single IRC safe observable can be written as

$$C_1 = \sum_i z_i \; g_1(\hat{\mathbf{p}}_i) \,,$$

where $g_1(\hat{p}_i)$ represents a parameterised multilayer perceptron. We have specifically denoted the observable as C_1 to make it self-evident the per-particle map essentially approximates any general $\mathcal{C}_1^{f_1}$. This is because the MLP g_1 is a universal approximator and can approximate any function f_1 suiting a particular objective up to a required precision. In a practical implementation, several related IRC safe observables are approximated, which are fed to a downstream network for classification. The direct implementation of EFNs can, therefore, only extract features expressible in terms of C_1.

2.2 Energy-weighted message passing networks

An energy-weighted message passing operation for any general parametrised function $\bar{\mathbf{g}}^{(\alpha)}$ can be written as

$$\mathbf{h}_i^{(\alpha+1)} = \sum_{j \in \mathcal{N}[i]} \omega_j^{(\mathcal{N}[i])} \; \bar{\mathbf{g}}^{(\alpha+1)}(\mathbf{h}_i^{(\alpha)}, \mathbf{h}_j^{(\alpha)}) \,,$$

where $\mathbf{h}_i^{(\alpha)}$ is the input node features for the α^{th} message passing operation and

$$\omega_j^{(\mathcal{N}[i])} = \frac{p_T^j}{\sum_{k \in \mathcal{N}[i]} \; p_T^k}$$

are the energy weights dependent on the IRC safe neighbourhood set $\mathcal{N}[i]$, with $\omega_j^{(S)} = z_j$, for the whole set $\mathcal{S}$. For notational convenience in the following discussions, we will take the sum over the full set of particles in the jet and replace z_j in place of $\omega_j^{(\mathcal{N}[i])}$ without loss of generality. Therefore, we have

$$\mathbf{h}_i^{(\alpha+1)} = \sum_j \; z_j \; \mathbf{g}^{(\alpha+1)}(\mathbf{h}_i^{(\alpha)}, \mathbf{h}_j^{(\alpha)}) \tag{2.2}$$

JHEP01(2024)113

with the function $\mathbf{g}^{(\alpha+1)}$ expressed as a product of a Heaviside step functions $\Theta(\Delta R_{ij} < R_0)$ and the original message function $\bar{\mathbf{g}}^{(\alpha+1)}$ as

$$\mathbf{g}^{(\alpha+1)}(\mathbf{h}_i^{(\alpha)}, \mathbf{h}_j^{(\alpha)}) = \Theta(\Delta R_{ij} < R_0)\, \bar{\mathbf{g}}^{(\alpha+1)}(\mathbf{h}_i^{(\alpha)}, \mathbf{h}_j^{(\alpha)})\,.$$

Here, ΔR_{ij} is the Euclidean distance in the rapidity-azimuth plane between particle i and j while R_0 is the graph's radius. The requirement of symmetry in the argument of $f_2(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_j)$ for $\mathcal{C}_2^{f_2}$ and its absence in eq. (2.2) is not a contradiction as the node features themselves are defined for each particle and hence are not IRC safe observables. In contrast, the IRC safe graph representation will generally be expressible as some linear combination of $\mathcal{C}_N^{f_N}$.

We have $\mathbf{h}_i^{(0)} = \hat{\mathbf{p}}_i$ which gives $\hat{\mathbf{p}}_i = \hat{\mathbf{p}}_j \implies \mathbf{h}_i^{(\alpha)} = \mathbf{h}_j^{(\alpha)}$ for any $\alpha \geq 0$ and any two collinear particles i and j. The IRC safe graph representation is obtained as

$$\mathbf{G}^{(L)} = \sum_{i=1}^{n_{part}} z_i\, \mathbf{h}_i^{(L)}\,,$$

after L iterations. As we shall see in the following, the complexity of the extracted features via EMPNs will depend on the value of L.

Explicitly for $L = 1$, we have $\mathbf{h}_i^{(1)} = \sum_j z_j\, \mathbf{g}^{(1)}(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_j)$ which gives

$$\mathbf{G}^{(1)} = \sum_{i,j} z_i\, z_j\, \mathbf{g}^{(1)}(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_j)\,.$$

If the symmetry is enforced in $\mathbf{g}^{(1)}$, the approximated observable will contain a $\mathcal{C}_2^{f_2}$ term alone. At the same time, a non-symmetric $\mathbf{g}^{(1)}$ would also have a $\mathcal{C}_1^{f_1}$ component.

For $L = 2$, we have

$$\begin{aligned} \mathbf{G}^{(2)} &= \sum_{i,j} z_i\, z_j\, \mathbf{g}^{(2)}(\mathbf{h}_i^{(1)}, \mathbf{h}_j^{(1)}) \\ \implies \mathbf{G}^{(2)} &= \sum_{i,j} z_i\, z_j\, \mathbf{g}^{(2)}\Big(\sum_k z_k\, \mathbf{g}^{(1)}(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_k), \sum_l z_l\, \mathbf{g}^{(1)}(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_l)\Big)\,. \end{aligned} \tag{2.3}$$

The complicated nature of the arguments makes it difficult to ascertain the exact behaviour of the functional approximation. One expects the universal approximator $\mathbf{g}^{(2)}$ to be expressible as a linear combination of $\mathcal{C}_N^{f_N}$'s up to $N = 4$. However, due to the presence of four angular arguments and four energy weights, it hints against the efficient approximation of any $\mathcal{C}_N^{f_N}$ for any $N < 4$.

The situation is even more futile for $L = 3$ with eight angular arguments and eight energy-weighted sums. For a particular L, we have 2^L angular arguments and the same number of energy-weighted sums. Even if one extracts the graph features at each stage α, and gets a concatenated graph representation for each $\alpha > 0$ up to $\alpha = L$, we have the efficient extraction of $2, 2^2, 2^3, \ldots 2^L$ terms the sum in eq. (2.1) for any general IRC safe observable $\mathcal{O}$. Although, for jet substructure applications, one does not need to go to very high N, we already run into a problem for top-tagging, which has valuable information in the 3-prong structure of the energy deposits.

JHEP01(2024)113

3 Hypergraph energy-weighted message passing networks

As discussed above, although powerful, Graph Neural Networks cannot look into higher-order relational information amongst the nodes efficiently. Therefore, in this section, we develop IRC-safe point cloud architectures capable of efficiently extracting higher-point correlation.

A possible way to extend the capabilities of IRC safe feature extraction to higher-point correlations is to directly implement the form of C-correlators as

$$\mathcal{H}^N = \sum_{i_1} \sum_{i_2} \dots \sum_{i_N} z_{i_1} z_{i_2} \dots z_{i_N} \, \Theta_N(\hat{p}_{i_1}, \hat{p}_{i_2}, \dots, \hat{p}_{i_N}) \, \Phi_N(\hat{p}_{i_1}, \hat{p}_{i_2}, \dots, \hat{p}_{i_N}) \,,$$

where Θ_N are step functions for reducing the sums to localised information, and Φ_N are the neural networks approximating a correlated set (as the output of Φ_N in general, is a vector) of f_N's for the particular training objective. For IRC safety, both Θ_N and Φ_N should be symmetric under the permutation of its arguments. The step function Θ_N for each N essentially endows an N-uniform hypergraph structure onto the constituent set similar to the radius filter $\Theta(\Delta R_{ij} < R_0)$ endowing a graph structure for the case of $N = 2$. Therefore, the concatenated hypergraph representations

$$\mathbf{X} = \oplus_N \mathcal{H}^N \,,$$

up to $N_{\max}$ would extract IRC safe features to be fed to a downstream MLP for some task.

We do not follow this approach for the following reasons. It is well-known [54–57] that automatic feature extraction works best with deeper networks. Depth can only be brought into Φ_N in the above expression, which does nothing to the IRC-safe feature extraction process. The complexity can be increased by increasing N, which increases the width of the network, thereby increasing the model complexity sharply. Although the factorisation of the extracted features in energy and angular components could lead to better all-order behaviour in QCD and is indeed interesting, one needs to have proper control of the behaviour of the parameter optimisation before we can hope to answer such questions as demonstrated in reference [33].

Our approach is based on one-particle and two-particle messages to construct a hybrid message-passing neural network that can extract higher point correlations in a recursive approach. Although it is easily generalisable to higher-point information, we restrict ourselves up to 3-point interactions due to the increasing complexity.

3.1 IRC safety with heterogenous source and destination embeddings

The basic observation which makes it possible to build a higher-point IRC safe feature extractor is that the requirement of IRC safety for EMPN is still valid even when the node embedding for the source $\psi_S(\hat{\mathbf{p}}_i)$, and destination $\psi_D(\hat{\mathbf{p}}_i)$ are different as long as they are separately equal in the collinear limit of two particles. If a particle q has two collinear daughters r and s, then we have

$$\psi_S(\hat{\mathbf{p}}_q) = \psi_S(\hat{\mathbf{p}}_r) = \psi_S(\hat{\mathbf{p}}_s) \,, \text{ and } \psi_D(\hat{\mathbf{p}}_q) = \psi_D(\hat{\mathbf{p}}_r) = \psi_D(\hat{\mathbf{p}}_s)$$

when $\hat{\mathbf{p}}_q = \hat{\mathbf{p}}_r = \hat{\mathbf{p}}_s$, even if $\psi_S(\hat{\mathbf{p}}_i) \neq \psi_D(\hat{\mathbf{p}}_i)$. More importantly, the embeddings ψ_S and ψ_D need not be functions of just a single particle. They can also be the updated node features

JHEP01(2024)113

of the α-hop IRC safe neighbourhood after α energy-weighted message passing operations (as given in eq. (2.2)). For an IRC safe neighbourhood of i, where a particle q splits to two daughters r and s, we have $\mathcal{N}[i] \ni q \implies \mathcal{N}'[i] \ni r \ \wedge \mathcal{N}'[i] \ni s$ when $\hat{\mathbf{p}}_q = \hat{\mathbf{p}}_r = \hat{\mathbf{p}}_s$.

Let us look closer into the statement that we need not have the same embedding in the argument of the message function in an Energy-weighted Message Passing operation even though the statement logically follows from the non-requirement of symmetry of the message function. Since we have heterogeneous source and destination embeddings, we need to fix a uniform direction of messages. We will take the direction of all messages as originating from a neighbourhood node $j \in \mathcal{N}[i]$ to the destination node i. Therefore, we have

$$\mathbf{H}_i^{(\alpha+1,\beta+1)} = \sum_j z_j \, \mathbf{g}^{(\alpha+1,\beta+1)}(\mathbf{h}_{D,i}^{(\alpha)}, \mathbf{h}_{S,j}^{(\beta)}) \,,$$

where $\mathbf{h}_{D,i}^{(\alpha)}$ and $\mathbf{h}_{S,j}^{(\beta)}$ are the destination and source node embeddings, respectively, and $\mathbf{g}^{(\alpha+1,\beta+1)}$ is the corresponding message function. As the destination and source node embeddings differ, the message-passing operations are indexed separately with α and β, respectively. The source embedding satisfying $\mathbf{h}_{S,q}^{(\beta)} = \mathbf{h}_{S,r}^{(\beta)} = \mathbf{h}_{S,s}^{(\beta)}$ in the collinear limit makes the updated node representation $\mathbf{H}_i^{(\alpha+1,\beta+1)}$ equal for $i \notin \{q, r, s\}$, in the splitted and unsplitted case since $z_q = z_r + z_s$. Explicitly, we have

$$z_q \, \mathbf{g}^{(\alpha+1,\beta+1)}(\mathbf{h}_{D,i}^{(\alpha)}, \mathbf{h}_{S,q}^{(\beta)}) = z_r \, \mathbf{g}^{(\alpha+1,\beta+1)}(\mathbf{h}_{D,i}^{(\alpha)}, \mathbf{h}_{S,r}^{(\beta)}) + z_s \, \mathbf{g}^{(\alpha+1,\beta+1)}(\mathbf{h}_{D,i}^{(\alpha)}, \mathbf{h}_{S,s}^{(\beta)}) \,. \tag{3.1}$$

Additionally, we require the equality of the destination embeddings $\mathbf{h}_{D,q}^{(\alpha)} = \mathbf{h}_{D,r}^{(\alpha)} = \mathbf{h}_{D,s}^{(\alpha)}$ when $i \in \{q, r, s\}$. However, we can have $\mathbf{h}_{D,q}^{(\alpha)} \neq \mathbf{h}_{S,q}^{(\beta)}$, as this is not needed to satisfy eq. (3.1). Therefore, $\mathbf{H}_i^{(\alpha+1,\beta+1)}$ satisfies $\mathbf{H}_q^{(\alpha+1,\beta+1)} = \mathbf{H}_r^{(\alpha+1,\beta+1)} = \mathbf{H}_s^{(\alpha+1,\beta+1)}$, in the collinear limit of the two daughters r and s of q.

3.2 Building higher point IRC safe feature extractor

It is now straightforward to build an IRC-safe message-passing operation which looks into three-particle correlations. The structure of the two-particle energy-weighted operation is kept the same as eq. (2.2), and then combined with destination embedding $\psi_D(\hat{\mathbf{p}}_i)$ and source embedding $\psi_S(\hat{\mathbf{p}}_i)$ of the angular coordinates to give an effective three particle message passing of the form

$$\begin{aligned} \mathbf{H}_i^{(1,2)} &= \sum_j z_j \, \mathbf{g}^{(1,2)}(\psi_D(\hat{\mathbf{p}}_i), \mathbf{h}_{S,j}^{(1)}) \,, \\ \mathbf{H}_i^{(2,1)} &= \sum_j z_j \, \mathbf{g}^{(2,1)}(\mathbf{h}_{D,i}^{(1)}, \psi_S(\hat{\mathbf{p}}_j)) \,. \end{aligned} \tag{3.2}$$

As the destination and source embeddings are different, $\mathbf{h}_{D,i}^{(1)}$ and $\mathbf{h}_{S,i}^{(1)}$ denote node features updated after two separate message-passing operations as given in eq. (2.2) with different message functions $\mathbf{g}_D^{(1)}$ and $\mathbf{g}_S^{(1)}$, respectively. The IRC safe feature would be a graph-level representation after an energy-weighted summed graph readout on $\mathbf{H}_i^{(1,2)}$ and $\mathbf{H}_i^{(2,1)}$, as

$$\mathbf{G}_3^{(1,2)} = \sum_i z_i \, \mathbf{H}_i^{(1,2)} \,, \quad \mathbf{G}_3^{(2,1)} = \sum_i z_i \, \mathbf{H}_i^{(2,1)} \,. \tag{3.3}$$

JHEP01(2024)113

We shall see in the following discussions that these two representations look at distinct topological structures in the graph; the IRC safe representation for the order three feature extraction is constructed as a concatenation of these two components

$$\mathbf{G}_3 = \mathbf{G}_3^{(1,2)} \oplus \mathbf{G}_3^{(2,1)} .$$

We can ascertain the behaviour of $\mathbf{G}_3$ by writing down its dependence on the particle's four vectors:

$$\mathbf{G}_3 = \sum_{i,j} z_i z_j \left(\mathbf{g}^{(1,2)}(\psi_D(\hat{\mathbf{p}}_i), \sum_l z_l \, \mathbf{g}_S^{(1)}(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_l)) \right.$$
$$\left. \oplus \, \mathbf{g}^{(2,1)}(\sum_l z_l \, \mathbf{g}_D^{(1)}(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_l), \psi_S(\hat{\mathbf{p}}_j)) \right) .$$

Three energy weights and three angular arguments hint that the learning procedure would directly start looking at the three-particle interrelations. It is important to note that any IRC safe observable looking into n body phase space, by definition, approaches its $n-1$ body phase space limit when one particle approaches the soft or collinear limit. In other words, eq. (2.3) will also look into the three-body limit of any four-particle combination when one is soft or collinear to any other particle. However, we expect the above form to extract better the three-particle correlations required for tagging three-prong jets like top quarks.

A schematic representation of the feature extraction procedure using different source and destination embeddings of order one and order two operations is shown in figure 2. We focus on the red node whose neighbours are the coloured. On the top left, the per-particle embeddings for the source and destination can only look into the individual particle information. On the right, however, the energy-weighted message-passing operation gathers information from each node's neighbourhood, which are shown with the identically coloured arrows for the coloured nodes. The order three feature extractors are built by combining the per-particle destination embedding with the order-two source embedding (on the left) and the order-two destination embedding with the per-particle source embedding (on the right).

From a feature extraction perspective, there are two essential differences in comparison to the $L = 2$ case given in eq. (2.3):

- One argument in both $\mathbf{g}^{(1,2)}$ and $\mathbf{g}^{(2,1)}$ is an embedding of the angular coordinates of a single particle and hence contain single-particle information. In contrast, both arguments already contain the aggregated neighbourhood information in $\mathbf{g}^{(2)}$.
- The embedding of the two arguments in $\mathbf{g}^{(1,2)}$ and $\mathbf{g}^{(2,1)}$ have independently trainable weights while they are shared for $\mathbf{g}^{(2)}$.

The first difference makes it possible for the function $\mathbf{g}^{(1,2)}$ to effectively extract the relation of node i with the updated neighbourhood information of its neighbours (2-hop neighbourhood of i), while the function $\mathbf{g}^{(2,1)}$ looks at the aggregated node feature of i's immediate neighbourhood with individual nodes in the same neighbourhood. The difference is also seen in figure 2, where on the left $\mathbf{H}_i^{(1,2)}$ looks into the features of the nodes within each coloured circle with the red node, while on the right, $\mathbf{H}_i^{(2,1)}$ looks into the feature of the

JHEP01(2024)113

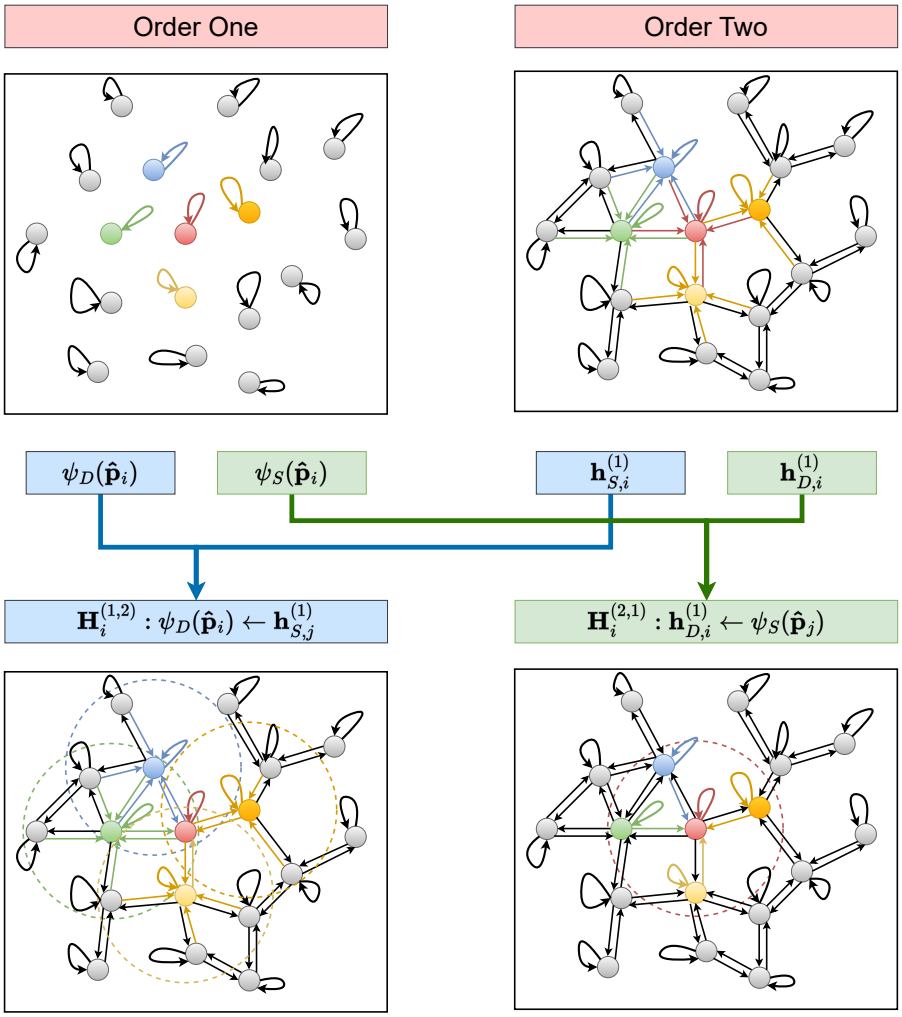

Figure 2. The figure shows a schematic representation of the message passing operation to build hybrid order three node representations for Hypergraph Energy-weighted Message Passing Networks by combining order one and two node representations.

aggregated neighbourhood information of the red node with the individual nodes within its neighbourhood. This essential difference in the feature extraction procedure makes it imperative to devise the two separate message-passing operations as they need to extract topologically different features within the graph.

It is straightforward to generalize this procedure to any arbitrary N, with substantial flexibility to choose the extractor guided by the requirement to divide N into two parts in any possible way. Any feature extractor looking into less than N correlations can be used to extract features from topologically distinct paths of length N within the graph. Due to the different combinatorial factors involved, the complexity rises relatively fast with increasing N, and we restrict our discussion to $N = 3$.

JHEP01(2024)113

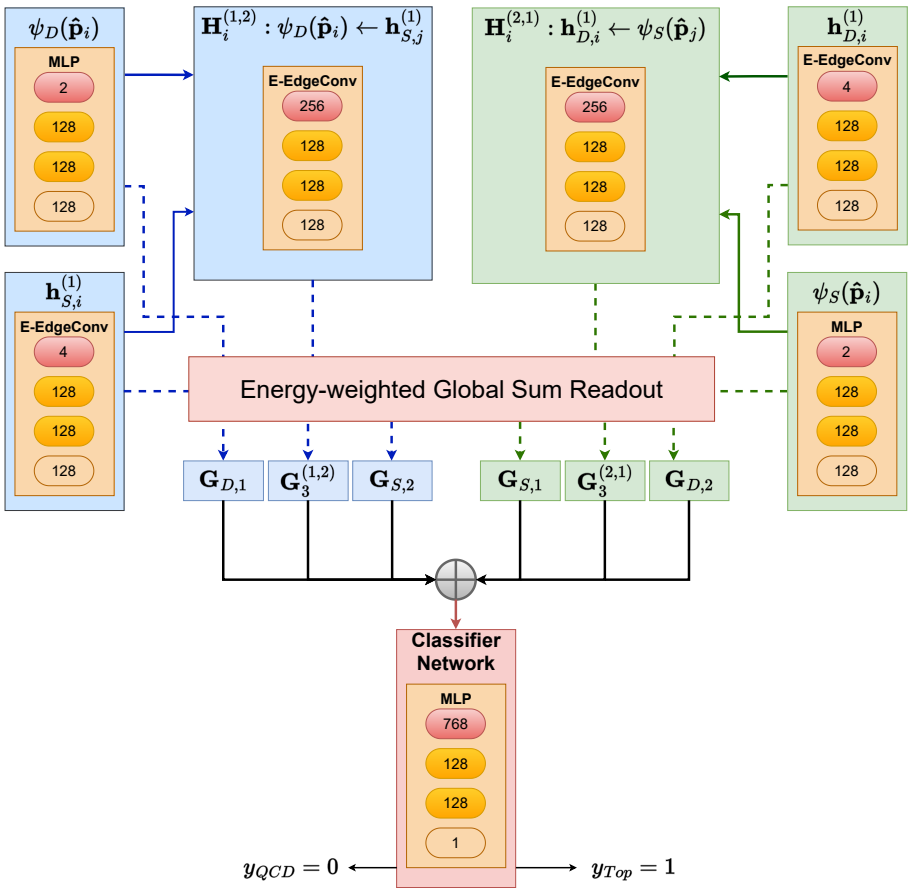

Figure 3. The architecture of the H-EMPN network utilized in this study is shown as a flowchart.

To look into the learnt features of the order one and two feature extractors, we define the graph representation as a concatenation of the source and destination embeddings as

$$\begin{aligned} \mathbf{G}_1 &= \mathbf{G}_{D,1} \oplus \mathbf{G}_{S,1} = \sum_i z_i \left(\psi_D(\hat{\mathbf{p}}_i) \oplus \psi_S(\hat{\mathbf{p}}_i)\right), \\ \mathbf{G}_2 &= \mathbf{G}_{D,2} \oplus \mathbf{G}_{S,2} = \sum_i z_i \left(\mathbf{h}_{D,i}^{(1)} \oplus \mathbf{h}_{S,i}^{(1)}\right). \end{aligned} \tag{3.4}$$

This gives the concatenated graph readout to be fed to the classifier network as

$$\mathbf{G} = \mathbf{G}_1 \oplus \mathbf{G}_2 \oplus \mathbf{G}_3 \,. \tag{3.5}$$

4 Network architecture and training

To gauge the properties of the proposed network, we utilise the public top-tagging dataset [58] for a supervised classifier. These events were generated with `Pythia 8.2.15` [59] and were showered and hadronised without MPI effects. The showered events additionally underwent a parametrised detector response via `Delphes3` [60] with the default ATLAS detector card. The particle-flow objects of the Delphes output were used as inputs to construct anti-k_T [61] jets with $R = 0.8$ via `FastJet` [62]. with additional requirements of p_T within the range

JHEP01(2024)113

$[550, 650]$ GeV, and pseudorapidity $|\eta| < 2$. Further, for the signal events, the top quark and its decay products' parton level information were used to reject falsely reconstructed jets with the partons falling outside the jet's area. The training data comprises 1.2 million samples, while the test and validation datasets contain 400k samples. The network analysis uses PyTorch-Geometric [63].

We compare order three Hypergraph Energy-weighted Message Passing Networks (H-EMPNs) with $L = 2$ EMPNs. For a reasonable comparison with the H-EMPN, we will extract the graph features for $\alpha = 1$ and $\alpha = 2$ stages separately for the EMPN and feed the concatenated graph representation into the classifier network. As shown in figure 3, the IRC-safe feature extractor module for the H-EMPN, in total, contains two per-particle maps for ψ_D and ψ_S, and four energy-weighted edge convolution (E-EdgeConv) operations to give the updated node embeddings $\mathbf{h}^{(1)}_{D,i}$, $\mathbf{h}^{(1)}_{S,i}$, $\mathbf{H}^{(1,2)}_{i}$, and $\mathbf{H}^{(2,1)}_{i}$. Including the classifier MLP, which takes in the concatenated graph readout, we have seven MLPs. We have one for each per-particle map and a message function for each E-EdgeConv operation from the feature extractor module. All these seven MLPs contain two hidden layers with 128 nodes and a rectified linear unit activation function. Except for the classifier network, which has a one-dimensional output with sigmoid activation, all other MLPs have a 128-dimensional output layer with a linear activation function. The per-particle maps take the rapidity-azimuth coordinates $\hat{\mathbf{p}}_i = (\Delta y_{iJ}, \Delta\phi_{iJ})$ of each constituent i as inputs with the differences taken from the jet axis defined by the four-vector $p^{\mu}_{J} = \sum_{k=1}^{n_{part}} p^{k}_{\mu}$. For a destination node embedding $\mathbf{h}_{S,i}$ and source node embedding $\mathbf{h}_{D,i}$, the message function takes in the concatenated vector $\mathbf{h}_{S,i} \oplus \mathbf{h}_{S,i} - \mathbf{h}_{S,j}$ as the input. The EMPN network sequentially applies the E-EdgeConv operation twice to the input graph's node features. The first and the second E-EdgeConv operations have the same MLP architecture corresponding to the ones that give $\mathbf{h}^{(1)}_{D,i}$ (or $\mathbf{h}^{(1)}_{S,i}$) and $\mathbf{H}^{(1,2)}_{i}$ (or $\mathbf{H}^{(2,1)}_{i}$), respectively. The classifier MLP for the EMPN and H-EMPN takes in 256 and 768-dimensional concatenated graph representations, respectively. The whole network is trained using the binary-cross entropy loss function.

We construct graphs with $R_0 \in \{0.4, 0.5, 0.6\}$ and $R_0 \to \infty$ corresponding to complete graphs.[1] For all these four instances of input graphs, we train each network five times from random initialization for 100 epochs with the `Adam` optimizer [64] and a learning rate of 0.001. A decay-on-plateau condition is applied to the learning rate with a decay factor of 0.5 if the validation loss does not decrease for three epochs. The epoch with minimum validation loss is used for inference for each training instance.

5 Results

5.1 Performance

The receiver operator characteristics (ROC) curve for the network with highest area under the ROC (AUC) curve from all training instances between the signal acceptance ϵ_S and the

[1] Strictly speaking, the maximum value that R_0 can take is determined by the jet's diameter as we are always confined to particles contained in the jet. We use the $R_0 \to \infty$ limit for defining the complete graph, as we are using sequential recombination algorithms and the maximum area of the jet is not compactly defined even for the anti-k_t algorithm which gives almost conical jets.

JHEP01(2024)113

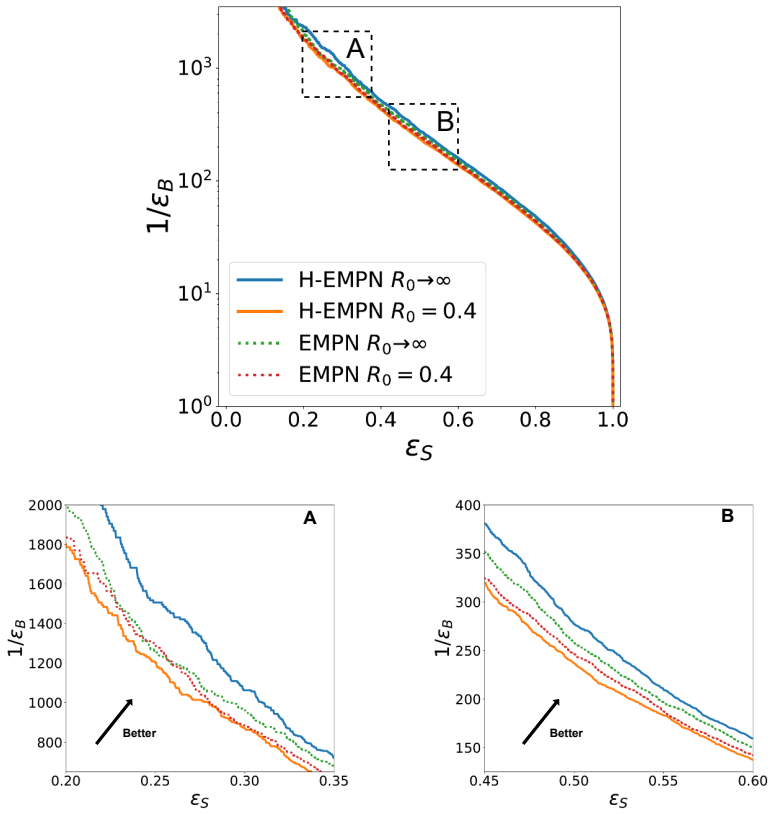

Figure 4. The receiver operator characteristics curve for the best performing network (in terms of AUC) over the five training instances for $R_0 = 0.4$ and $R_0 \to \infty$ for the EMPN and H-EMPN for different ranges of signal acceptance ϵ_S. The figure on the top shows $1/\epsilon_b$ in log scale over the full range of ϵ_S, while on the center and right, it is shown in linear scale over different regions of ϵ_S to highlight the differences.

inverse of background acceptance $1/\epsilon_B$ for the two models for $R_0 = 0.4$ and $R_0 \to \infty$ is shown in figure 4. We see that the EMPN has almost an overlapping ROC curve for these two radii, while for the H-EMPN, there is a noticeable improvement. The area under the receiver operator curve for the EMPN and H-EMPN for different graph construction radii are tabulated in table 1. The values correspond to the mean over the five training instances, while the errors correspond to the standard deviation. For $R_0 = 0.4$, the EMPN and H-EMPN have almost identical discrimination power with an AUC of 0.9823 and 0.9821, respectively. As the radius increases, there is a steady increase for the H-EMPN, while for the EMPN, it increases for $R_0 = 0.5$ and stays at a similar value for $R_0 = 0.6$ and there is a noticeable dip in performance when going to complete graphs with $R_0 \to \infty$. This trend can be understood from the structural difference between the EMPN and H-EMPN and the three-prong nature of the top jet. The EMPN's feature extraction is sequential, with the second E-EdgeConv being fed by the first E-EdgeConv's updated node features. With increasing radius, the feature-extraction, which looks at aggregated two-particle correlations, suffers from a redundancy of the information as the first E-EdgeConv already looks at a much larger neighbourhood in the rapidity-azimuth plane. On the other hand, the H-EMPN has a much larger width, with four

JHEP01(2024)113

Model	Area Under the ROC Curve			
	$R_0 = 0.4$	$R_0 = 0.5$	$R_0 = 0.6$	$R_0 \to \infty$
EMPN	0.9823 ± 0.00015	0.9827 ± 0.00009	0.9826 ± 0.00024	0.9825 ± 0.00015
H-EMPN	0.9821 ± 0.00012	0.9826 ± 0.00010	0.9828 ± 0.00029	0.9834 ± 0.00012

Table 1. The table shows the mean AUC for five training instances evaluated on the test dataset of the public top-tagging dataset for different architectures. The errors shown are the standard deviation of the five training instances.

Model	$1/\epsilon_B$ at $\epsilon_S = 0.5$			
	$R_0 = 0.4$	$R_0 = 0.5$	$R_0 = 0.6$	$R_0 \to \infty$
EMPN	235 ± 7	250 ± 2	246 ± 4	255 ± 6
H-EMPN	236 ± 2	258 ± 6	258 ± 11	276 ± 6

Table 2. The table shows the background rejection at a signal acceptance of 50% for different models. The values correspond to the mean from the evaluation of the test dataset for five different training instances from random initialization, while the standard deviations are shown as errors.

modules taking the input jet constituents parallelly, which are then combined non-trivially to feed the order-three feature extractors. Even though the order-three extractors take in the updated order-two node features from the full jet in the $R_0 \to \infty$ limit, the combination with the per-particle maps drives the extraction process to look at any relevant three-prong structure in the whole jet. From a purely QCD perspective, the radius R_0 puts in an additional scale beyond the jet radius, and going to the $R_0 \to \infty$ limit takes away this dependence in the feature extraction procedure. Although it is possible to define R_0 as a function of the IRC safe kinematic information of the jet which could possibly improve the feature extraction, we do not consider this as our aim is to move towards theoretically transparent ways of improving feature extraction. Therefore, the H-EMPN can extract features from the full jet more efficiently without being restrained by an arbitrary angular scale R_0.

The AUC paints a global picture of the discrimination power of a binary classifier; however, a classifier is almost always used at a specific working point, depending on the analysis. This practical aspect demands a local figure of merit, which we show with the inverse of the background acceptance ϵ_B, the background rejection $1/\epsilon_B$, at fixed values of signal acceptance ϵ_S. The background rejection for the EMPN and H-EMPN for the different graph construction radii are shown for $\epsilon_S = 0.5$ and $\epsilon_S = 0.3$ in tables 2 and 3, respectively. The values are averaged over the five training instances, with the standard deviations shown as errors. Although the trend for separate models is similar to that of the AUCs, the H-EMPN already starts having a noticeably better background rejection for $R_0 = 0.5$ even though the EMPN has a nominally higher AUC. As a matter of fact, except for $R_0 = 0.4$ at $\epsilon_S = 0.3$, the H-EMPN has a numerically higher mean background rejection for all other instances.

JHEP01(2024)113

Model	$1/\epsilon_B$ at $\epsilon_S = 0.3$			
	$R_0 = 0.4$	$R_0 = 0.5$	$R_0 = 0.6$	$R_0 \to \infty$
EMPN	819 ± 39	882 ± 11	839 ± 31	895 ± 36
H-EMPN	817 ± 33	917 ± 25	911 ± 34	995 ± 48

Table 3. The table shows the background rejection ($1/\epsilon_B$) at a signal acceptance (ϵ_S) of 30% for different models. The values correspond to the mean from the evaluation of the test dataset for five different training instances from random initialization, while the standard deviations are shown as errors.

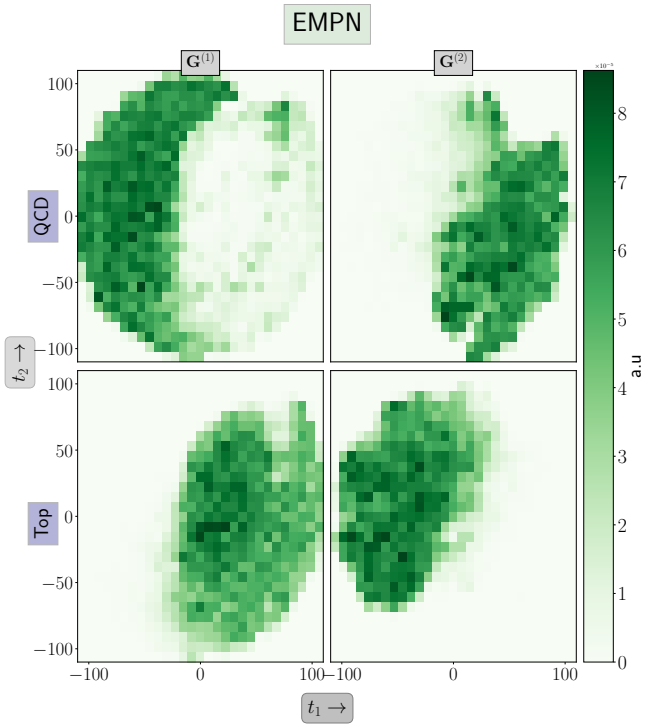

Figure 5. The two-dimensional histogram of the QCD (above) and top (below) test datasets in the two-dimensional latent space obtained after a t-SNE embedding of the 128-dimensional graph representation $\mathbf{G}^{(1)}$ (left) and $\mathbf{G}^{(2)}$ (right) of the best performing EMPN trained with complete graphs.

5.2 Visualizing the latent graph representation

In this section, we investigate whether all the graph representations that the H-EMPN learns can contribute to separating the signal and the background for the final classifier output. We choose the best-performing complete graph, which has the possibility of the highest information redundancy besides being the strongest classifier. Although a relatively high linear correlation with the network output does point to the classification using that particular information, it is defined for each component of the graph representation, which dilutes the importance of the underlying vector representations. Moreover, the absence of linear correlation does not imply the lack of discriminatory information, as neural networks can be highly non-linear functions of their inputs.

JHEP01(2024)113

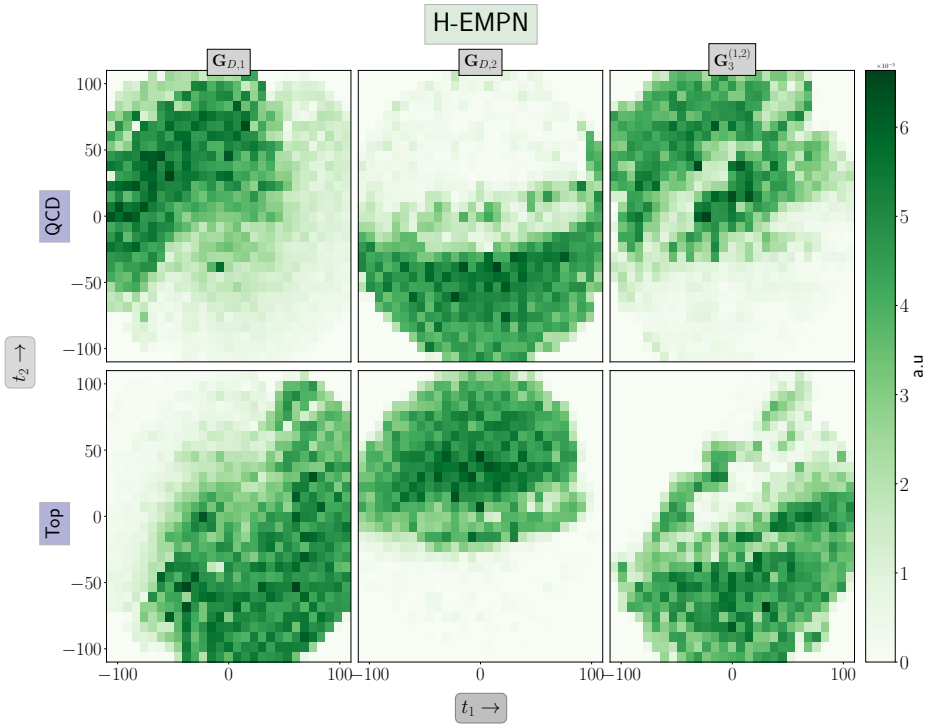

Figure 6. The two-dimensional histogram of the QCD (above) and top (below) test datasets in the two-dimensional latent space obtained after a t-SNE embedding of the 128-dimensional graph representation $\mathbf{G}_{D,1}$ (left), $\mathbf{G}_{D,2}$ (center) and $\mathbf{G}_3^{(1,2)}$ (right) of the best performing H-EMPN trained with complete graphs.

We look into the separating power of the different graph representations by visualizing them in a two-dimensional latent space using the t-distributed Stochastic Neighbourhood Embedding (t-SNE) [65] — an unsupervised data representation technique, where high dimensional data is embedded non-linearly in a lower dimensional space by maximally conserving the neighbourhood information endowed by a Euclidean metric in both spaces. In other words, nearby points in the high-dimensional representation get mapped to a local neighbourhood in the low-dimensional space. As it is an unsupervised technique, no explicit class information (QCD and top for our case) is fed when learning the map, and the clusters that arise in the low-dimensional space are a consequence of their proximity in the high-dimensional space. Therefore, a well-separated cluster in the lower-dimensional space implies that the higher-dimensional space also has well-separated regions.

We use the implementation of t-SNE in `Scikit-learn` [66] package to embed the various 128-dimensional graph representations of the test dataset evaluated on the best performing EMPN and H-EMPN for the complete graph in a two-dimensional space separately for each representation. The class-wise two-dimensional histogram in the embedding space (t_1, t_2) for $\mathbf{G}^{(1)}$ and $\mathbf{G}^{(2)}$ for the EMPN are shown in figure 5. We can see that both the graph representations have relatively distinct regions in (t_1, t_2) for the QCD samples (shown above) and top samples (shown below). Similarly, the two-dimensional histograms for the graph representations constructed out of the destination and source node-embeddings for the H-

JHEP01(2024)113

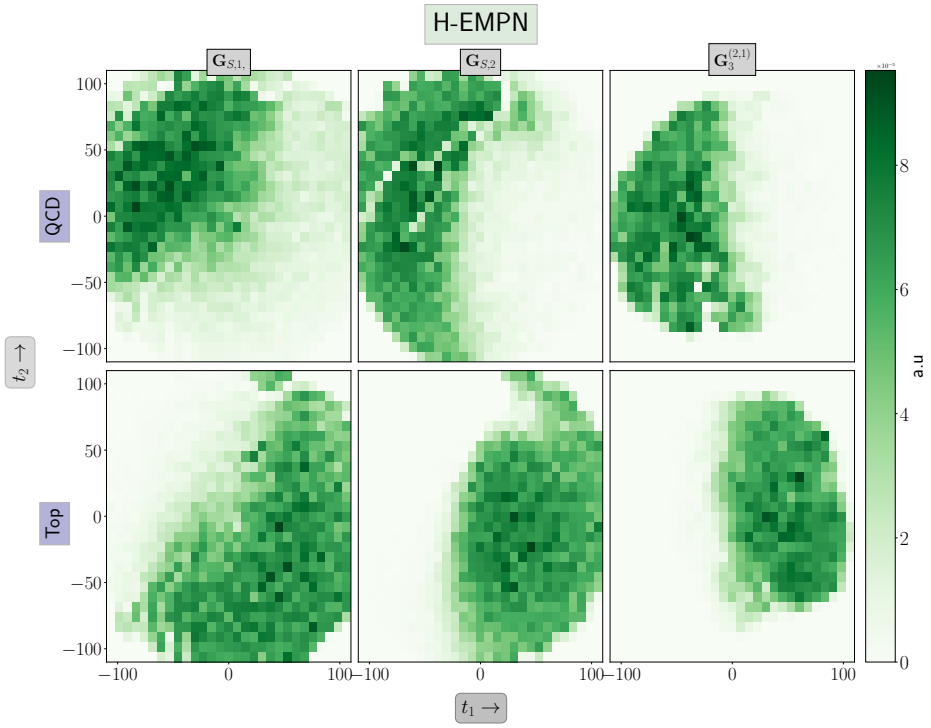

Figure 7. The two-dimensional histogram of the QCD (above) and top (below) test datasets in the two-dimensional latent space obtained after a t-SNE embedding of the 128-dimensional graph representation $\mathbf{G}_{S,1}$ (left), $\mathbf{G}_{S,2}$ (center) and $\mathbf{G}_3^{(2,1)}$ (right) of the best performing H-EMPN trained with complete graphs.

EMPN are shown in figures 6 and 7, respectively. All these embedded graph representations exhibit clear clustering of the QCD and top samples in different regions, confirming that the H-EMPN has extracted discriminating features from all of its component modules.

Although the EMPN and H-EMPN can utilize their constituent graph representation to separate the QCD jets from top jets as seen from these two-dimensional histograms, we reiterate the qualitative differences between these two networks from the QCD perspective. The $L = 2$ EMPN looks up to order four relations. In contrast, the H-EMPN in its present guise only looks up to order three — the sequential application of E-EdgeConv (to give $\mathbf{H}_i^{(1,2)}$ and $\mathbf{H}_i^{(2,1)}$) takes in the per-particle map with single particle information rather than an updated node feature with the local neighbourhood information in one of its arguments. However, we can see the better ability of the H-EMPN network from its performance studies and potentially better behaviour in QCD with its greater efficacy in the absence of an arbitrary angular scale R_0. Since we took the top vs QCD jets classification example, we already knew that there is beneficial information in the three-prong structure within the jet, which prompted our design of the specific H-EMPN.[2] The first observation from the finite R_0 cases is that the H-EMPN architecture is more critical in extracting the order three relational information from the jets than the $L = 2$ EMPN. On the other hand, our

[2]The situation may be different, for instance, in the quark vs gluon case where the separating information is not in the hard prong structure but the soft radiation pattern surrounding the one prong core within the jet.

JHEP01(2024)113

a priori knowledge of QCD, prompting the design of the H-EMPN, validates that physical inductive biases, or more specifically, QCD, have an important role in the design of performant feature extractors. Therefore, rather than throwing a currently "fashionable network" under the hood, designing architectures based on the underlying physical intuition can help push the performance boundaries of deep learning algorithms and gain (at least) a qualitative understanding of their inner workings.

6 Conclusions

This study delved deep into the intricacies of generalised automatic infrared and collinear safe feature extraction for LHC phenomenology, focusing on the potential of Graphs and Hypergraphs. Hypergraphs are a generalisation of traditional graphs. While a standard graph consists of vertices connected by edges, each connecting exactly two vertices, a hypergraph allows edges to connect any number of vertices, offering a more flexible way to represent relationships between entities.

First, we explored the behaviour of energy-weighted message passing and its capability to approximate general infrared and collinear safe observables. We highlighted the significance of IRC-safe observables, especially in the context of data interpretation at LHC experiments. The study further explored the capabilities of Energy Flow Networks and Energy-weighted message-passing networks, shedding light on their potential and constraints utilising the usage of multilayer perceptrons as universal function approximators within the architecture with the IRC-safe observables expressible in terms of C-correlators.

To enhance the capabilities of IRC safe feature extraction, especially for higher-point correlations, a novel method was introduced by leveraging the form of C-correlators and heterogenous source and destination node embeddings. This approach presents a renewed outlook on feature extraction.

Qualitatively assessing the two models, while the EMPN model provides a robust foundation for feature extraction, the H-EMPN model, designed to look at order-three interparticle relations, demonstrates an edge in performance metrics even though the EMPN model via the application of two-message passing operations could theoretically look up to order-four. This suggests that incorporating hypergraph structures in the H-EMPN model offers enhanced capabilities in extracting higher-point correlations, making it a promising tool for more intricate analyses in LHC phenomenology.

Our findings underscore the potential of hypergraph-based methods in enhancing the extraction of IRC-safe features. The research paves the way for further exploration into LHC phenomenology, focusing on optimising feature extraction techniques.

Acknowledgments

M.S. is supported by the STFC under grant ST/P001246/1. Computational work were performed on the Param Vikram-1000 High Performance Computing Cluster and TDP resources at the Physical Research Laboratory (PRL).

JHEP01(2024)113

References

[1] A. Andreassen et al., *OmniFold: a method to simultaneously unfold all observables*, *Phys. Rev. Lett.* **124** (2020) 182001 [arXiv:1911.09107] [INSPIRE].

[2] P.T. Komiske, E.M. Metodiev and J. Thaler, *Metric space of collider events*, *Phys. Rev. Lett.* **123** (2019) 041801 [arXiv:1902.02346] [INSPIRE].

[3] S. Bieringer et al., *Measuring QCD splittings with invertible networks*, *SciPost Phys.* **10** (2021) 126 [arXiv:2012.09873] [INSPIRE].

[4] D. Kim et al., *Deep-learned event variables for collider phenomenology*, *Phys. Rev. D* **107** (2023) L031904 [arXiv:2105.10126] [INSPIRE].

[5] Y.S. Lai, J. Mulligan, M. Płoskoń and F. Ringer, *The information content of jet quenching and machine learning assisted observable design*, *JHEP* **10** (2022) 011 [arXiv:2111.14589] [INSPIRE].

[6] A. Romero et al., *Safety of quark/gluon jet classification*, arXiv:2103.09103 [INSPIRE].

[7] J. Batson, C.G. Haaf, Y. Kahn and D.A. Roberts, *Topological obstructions to autoencoding*, *JHEP* **04** (2021) 280 [arXiv:2102.08380] [INSPIRE].

[8] I. Chahrour and J.D. Wells, *Comparing machine learning and interpolation methods for loop-level calculations*, *SciPost Phys.* **12** (2022) 187 [arXiv:2111.14788] [INSPIRE].

[9] A. Butter, B.M. Dillon, T. Plehn and L. Vogel, *Performance versus resilience in modern quark-gluon tagging*, *SciPost Phys. Core* **6** (2023) 085 [arXiv:2212.10493] [INSPIRE].

[10] F.A. Dreyer, R. Grabarczyk and P.F. Monni, *Leveraging universality of jet taggers through transfer learning*, *Eur. Phys. J. C* **82** (2022) 564 [arXiv:2203.06210] [INSPIRE].

[11] P. Onyisi, D. Shen and J. Thaler, *Comparing point cloud strategies for collider event classification*, *Phys. Rev. D* **108** (2023) 012001 [arXiv:2212.10659] [INSPIRE].

[12] Z. Kassabov et al., *The top quark legacy of the LHC run II for PDF and SMEFT analyses*, *JHEP* **05** (2023) 205 [arXiv:2303.06159] [INSPIRE].

[13] A. Butter et al., *Jet diffusion versus JetGPT — modern networks for the LHC*, arXiv:2305.10475 [INSPIRE].

[14] W. Shen, D. Wang and J.M. Yang, *Hierarchical high-point energy flow network for jet tagging*, *JHEP* **09** (2023) 135 [arXiv:2308.08300] [INSPIRE].

[15] A. Rousselot and M. Spannowsky, *Generative invertible quantum neural networks*, arXiv:2302.12906 [INSPIRE].

[16] 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].

[17] 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].

[18] F.A. Dreyer and H. Qu, *Jet tagging in the Lund plane with graph networks*, *JHEP* **03** (2021) 052 [arXiv:2012.08526] [INSPIRE].

JHEP01(2024)113

[19] T. Faucett, J. Thaler and D. Whiteson, *Mapping machine-learned physics into a human-readable space*, *Phys. Rev. D* **103** (2021) 036020 [arXiv:2010.11998] [INSPIRE].

[20] Y.S. Lai, D. Neill, M. Płoskoń and F. Ringer, *Explainable machine learning of the underlying physics of high-energy particle collisions*, *Phys. Lett. B* **829** (2022) 137055 [arXiv:2012.06582] [INSPIRE].

[21] A. Bogatskiy et al., *Explainable equivariant neural networks for particle physics: PELICAN*, arXiv:2307.16506 [INSPIRE].

[22] D. Athanasakos et al., *Is infrared-collinear safe information all you need for jet classification?*, arXiv:2305.08979 [INSPIRE].

[23] S. Bollweg et al., *Deep-learning jets with uncertainties and more*, *SciPost Phys.* **8** (2020) 006 [arXiv:1904.10004] [INSPIRE].

[24] A. Ghosh, B. Nachman and D. Whiteson, *Uncertainty-aware machine learning for high energy physics*, *Phys. Rev. D* **104** (2021) 056026 [arXiv:2105.08742] [INSPIRE].

[25] R. Gambhir, B. Nachman and J. Thaler, *Learning uncertainties the frequentist way: calibration and correlation in high energy physics*, *Phys. Rev. Lett.* **129** (2022) 082001 [arXiv:2205.03413] [INSPIRE].

[26] A. Butter et al., *Generative networks for precision enthusiasts*, *SciPost Phys.* **14** (2023) 078 [arXiv:2110.13632] [INSPIRE].

[27] R.T. d'Agnolo et al., *Learning new physics from an imperfect machine*, *Eur. Phys. J. C* **82** (2022) 275 [arXiv:2111.13633] [INSPIRE].

[28] M. Bellagente, M. Haussmann, M. Luchmann and T. Plehn, *Understanding event-generation networks via uncertainties*, *SciPost Phys.* **13** (2022) 003 [arXiv:2104.04543] [INSPIRE].

[29] A. Ghosh and B. Nachman, *A cautionary tale of decorrelating theory uncertainties*, *Eur. Phys. J. C* **82** (2022) 46 [arXiv:2109.08159] [INSPIRE].

[30] A. Ghosh et al., *Statistical patterns of theory uncertainties*, *SciPost Phys. Core* **6** (2023) 045 [arXiv:2210.15167] [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] A. Bogatskiy et al., *Lorentz group equivariant neural network for particle physics*, arXiv:2006.04780 [INSPIRE].

[33] G. Kasieczka, S. Marzani, G. Soyez and G. Stagnitto, *Towards machine learning analytics for jet substructure*, *JHEP* **09** (2020) 195 [arXiv:2007.04319] [INSPIRE].

[34] S. Badger and J. Bullock, *Using neural networks for efficient evaluation of high multiplicity scattering amplitudes*, *JHEP* **06** (2020) 114 [arXiv:2002.07516] [INSPIRE].

[35] D. Maître and H. Truong, *A factorisation-aware matrix element emulator*, *JHEP* **11** (2021) 066 [arXiv:2107.06625] [INSPIRE].

[36] P. Konar, V.S. Ngairangbam and M. Spannowsky, *Energy-weighted message passing: an infra-red and collinear safe graph neural network algorithm*, *JHEP* **02** (2022) 060 [arXiv:2109.14636] [INSPIRE].

[37] S. Gong et al., *An efficient Lorentz equivariant graph neural network for jet tagging*, *JHEP* **07** (2022) 030 [arXiv:2201.08187] [INSPIRE].

JHEP01(2024)113

[38] Z. Hao, R. Kansal, J. Duarte and N. Chernyavskaya, *Lorentz group equivariant autoencoders*, *Eur. Phys. J. C* **83** (2023) 485 [arXiv:2212.07347] [INSPIRE].

[39] S.E. Park, P. Harris and B. Ostdiek, *Neural embedding: learning the embedding of the manifold of physics data*, *JHEP* **07** (2023) 108 [arXiv:2208.05484] [INSPIRE].

[40] O. Atkinson et al., *IRC-safe graph autoencoder for unsupervised anomaly detection*, *Front. Artif. Intell.* **5** (2022) 943135 [arXiv:2204.12231] [INSPIRE].

[41] H. Qu and L. Gouskos, *ParticleNet: jet tagging via particle clouds*, *Phys. Rev. D* **101** (2020) 056019 [arXiv:1902.08570] [INSPIRE].

[42] V. Mikuni and F. Canelli, *ABCNet: an attention-based method for particle tagging*, *Eur. Phys. J. Plus* **135** (2020) 463 [arXiv:2001.05311] [INSPIRE].

[43] E. Bernreuther et al., *Casting a graph net to catch dark showers*, *SciPost Phys.* **10** (2021) 046 [arXiv:2006.08639] [INSPIRE].

[44] A. Blance and M. Spannowsky, *Unsupervised event classification with graphs on classical and photonic quantum computers*, *JHEP* **21** (2020) 170 [arXiv:2103.03897] [INSPIRE].

[45] E.A. Moreno et al., *JEDI-net: a jet identification algorithm based on interaction networks*, *Eur. Phys. J. C* **80** (2020) 58 [arXiv:1908.05318] [INSPIRE].

[46] O. Atkinson et al., *Anomaly detection with convolutional graph neural networks*, *JHEP* **08** (2021) 080 [arXiv:2105.07988] [INSPIRE].

[47] O. Atkinson et al., *Improved constraints on effective top quark interactions using edge convolution networks*, *JHEP* **04** (2022) 137 [arXiv:2111.01838] [INSPIRE].

[48] S. Tsan et al., *Particle graph autoencoders and differentiable, learned energy mover's distance*, in the proceedings of the 35th *conference on neural information processing systems*, (2021) [arXiv:2111.12849] [INSPIRE].

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

[50] 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].

[51] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, *Math. Control Signals Syst.* **2** (1989) 303 [INSPIRE].

[52] M. Leshno, V.Y. Lin, A. Pinkus and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, *Neural Networks* **6** (1993) 861.

[53] R. Arora, A. Basu, P. Mianjy and A. Mukherjee, *Understanding deep neural networks with rectified linear units*, in *International conference on learning representations*, ICLR 2018 (2018).

[54] Y. Bengio and Y. LeCun, *Scaling learning algorithms toward AI*, in *Large-scale kernel machines*, The MIT Press (2007), p. 321 [DOI:10.7551/mitpress/7496.003.0016].

[55] P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol, *Extracting and composing robust features with denoising autoencoders*, in *Proceedings of the* 25th *International Conference on Machine Learning — ICML '*08, ACM Press (2008) [DOI:10.1145/1390156.1390294].

[56] Y. Bengio, *Learning deep architectures for AI*, *Found. Trends Machine Learn.* **2** (2009) 1.

[57] K. He, X. Zhang, S. Ren and J. Sun, *Deep residual learning for image recognition*, in 2016 *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), p. 770 [DOI:10.1109/CVPR.2016.90] [arXiv:1512.03385] [INSPIRE].

JHEP01(2024)113

[58] G. Kasieczka, T. Plehn, J. Thompson and M. Russel, *Top quark tagging reference dataset*, Zenodo, March 2019 [DOI:10.5281/ZENODO.2603256].

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

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

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

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

[63] M. Fey and J.E. Lenssen, *Fast graph representation learning with PyTorch geometric*, arXiv:1903.02428 [INSPIRE].

[64] D.P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, in 3rd *International Conference on Learning Representations, ICLR* 2015, Y. Bengio and Y. LeCun eds., San Diego, CA, U.S.A., 7–9 May 2015 [arXiv:1412.6980] [INSPIRE].

[65] L. van der Maaten and G. Hinton, *Visualizing data using t-SNE*, *J. Machine Learning Res.* **9** (2008) 2579.

[66] F. Pedregosa et al., *Scikit-learn: machine learning in python*, *J. Machine Learning Res.* **12** (2011) 2825 [arXiv:1201.0490] [INSPIRE].