Streamlined jet tagging network assisted by jet prong structure

Speaker: Ahmed Hammad with Mihoko Nojiri

Theory center, KEK, Japan

LCWS workshop 2024 @ Tokyo

9th July 2024

MLPhYs 学術変革領域研究(A) 学習物理学の創成
Foundation of "Machine Learning Physics"

Introduction

Jet tagging:

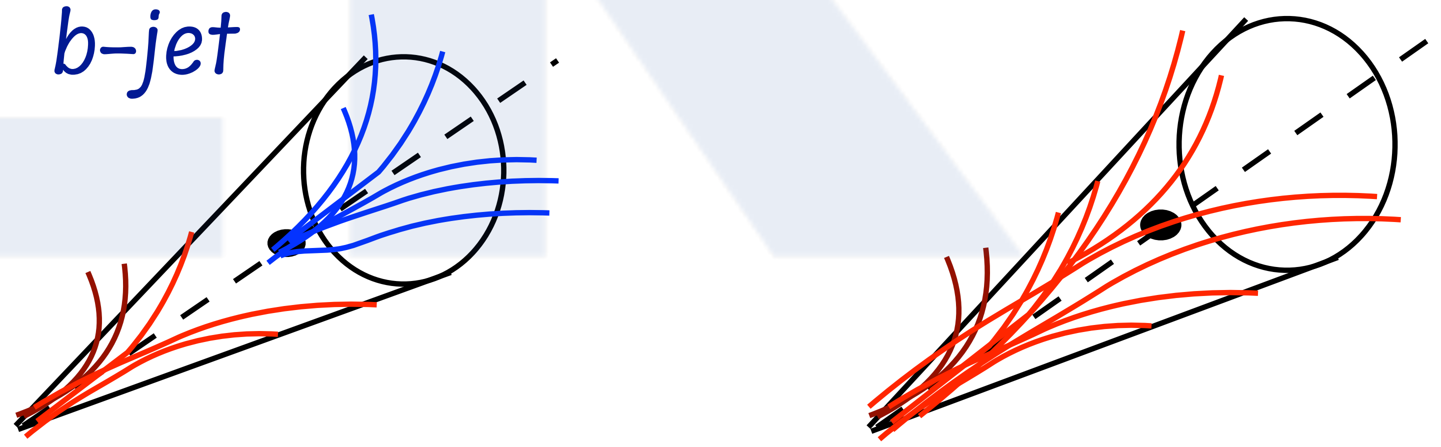
Identify the hard scattering particle that initiates the jet.

Jet identification is important to improve our understanding for the QCD processes.
Moreover, it improves the significance for new physics searches.

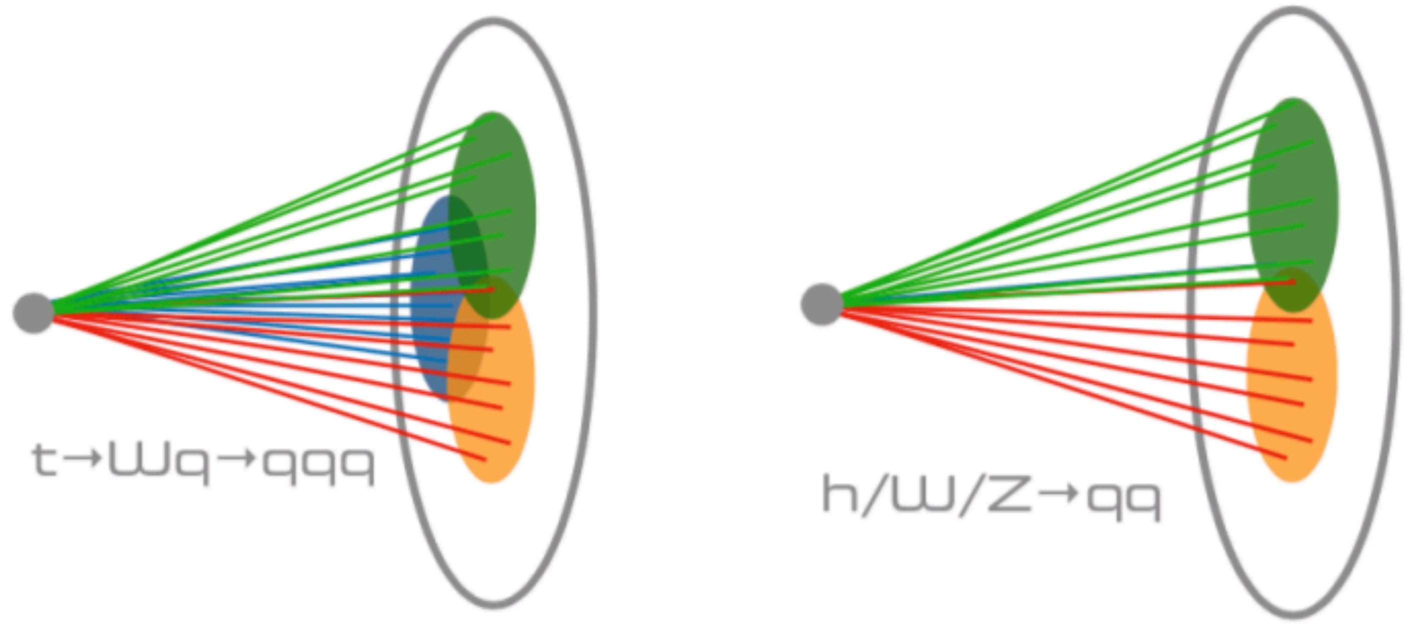
Examples:

- Heavy flavor tagging, bottom and charm tagging

b-jet Light flavor-jet



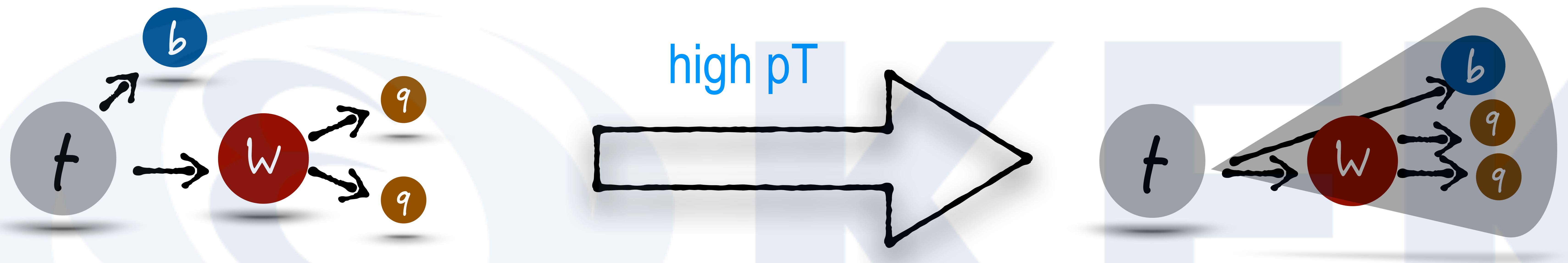
- Heavy resonance tagging, Top-jet, W-jet, Higgs-jet



Introduction

Boosted Jet tagging:

At high pT, the decay products from heavy particles, Higgs,W,Z,top, become collimated and can be contained in a single large-R jet



At the boosted region, heavy resonances exhibit different features from the QCD jets

Machine learning

Existence of many new approaches has been proposed in the past few years, leading to significant improvement in performance and deeper insights into jet physic.

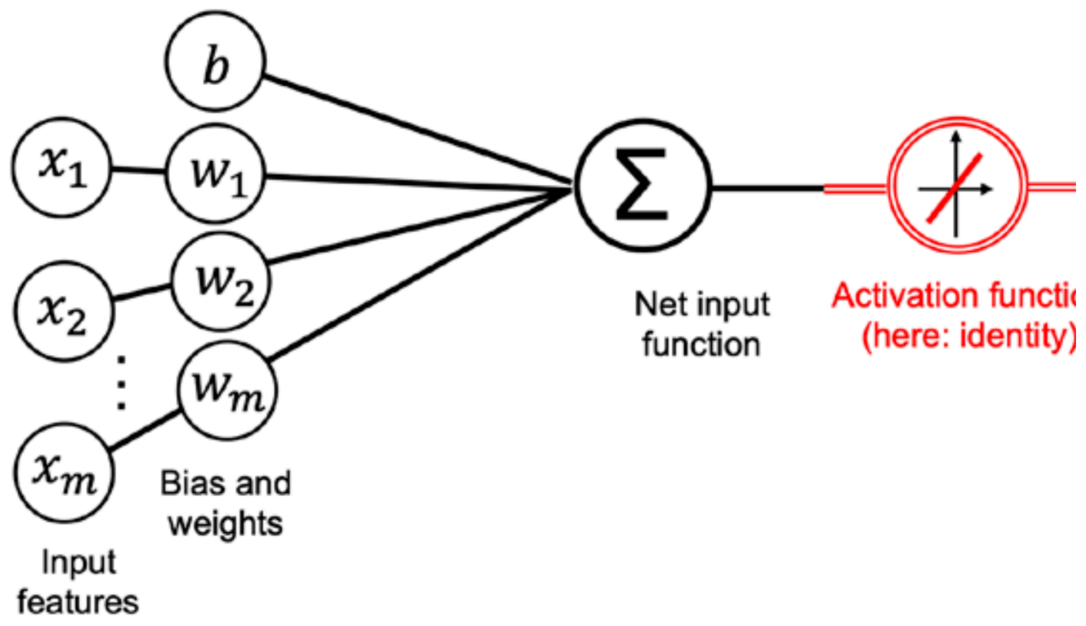
What is the difference between the different ML models? Which is the best one?

Introduction

Low computation cost

Multi-Layers perceptron:

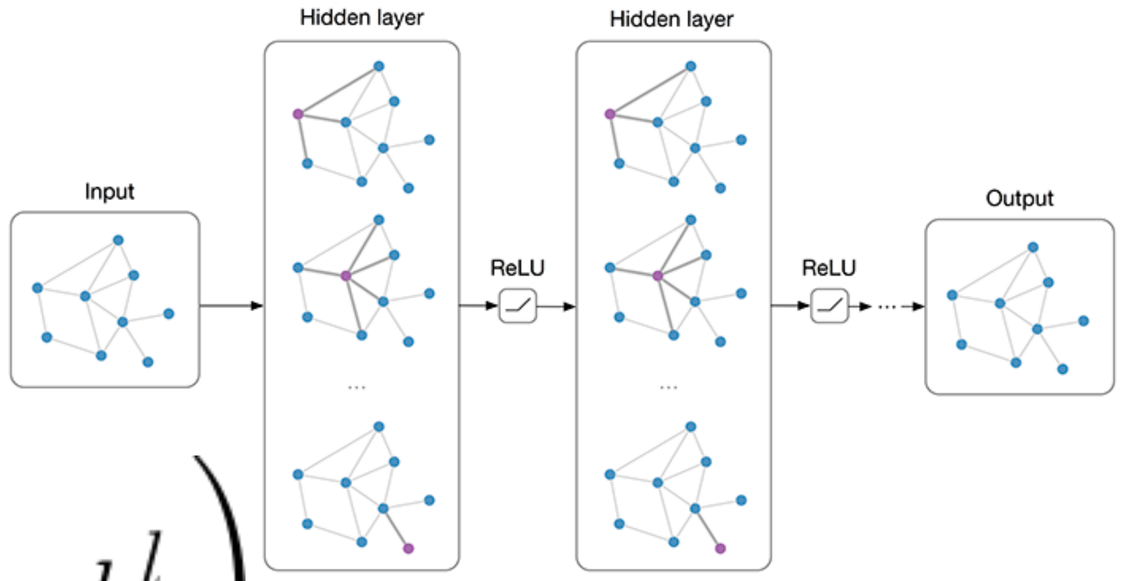
$$Z_i^l = \sigma\left(\sum_{j=1}^{l} W_{i,j}^l \cdot x_j^{l-1} + b_i^l\right)$$



The sum runs over each neuron in the layer fixed pattern of the latent neurons

MLP is not invariant under translation, rotation and permutation.
Suffer from the input sparsity.

Graph NeuralNetwork



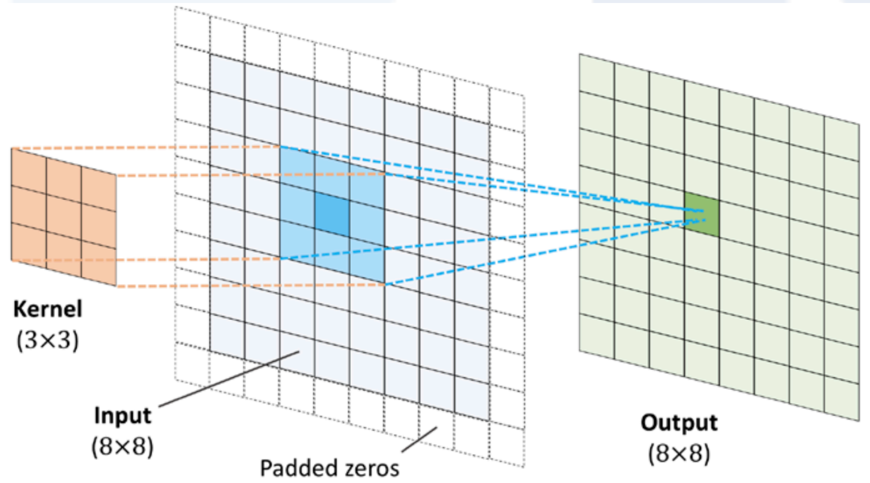
High computation cost

$$h_\nu^{l+1} = \sigma\left(W^{(l)} h^l + \sum^{n} W_n^l \cdot h_n^l\right)$$

For a given graph the nodes for the next layer are updated by aggregating the information from the near by nodes from the previous layer

The node update runs sequentially over the graph nodes and thus GNN is not invariant under permutation but no sparsity issue.

Convolution NeuralNetwork



$$Z_{i,j}^l = \sigma\left(\sum_m \sum_n X(i+m, j+n) \times W(m,n)\right)$$

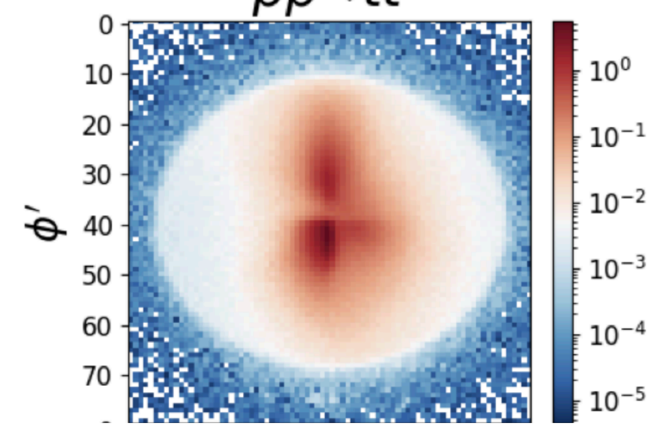
I,j run over spatial dimensions of the image. Local weights captured by each kernel is shared to allow for Translation invariance

Low computation cost

Only the important information are kept via pooling

$$P_{i,j}^l = \mathrm{Max}_{(m,n)}\left[Z^l_{(i\times M+m, j\times M+n)}\right]$$

$pp \to \bar{t}t$

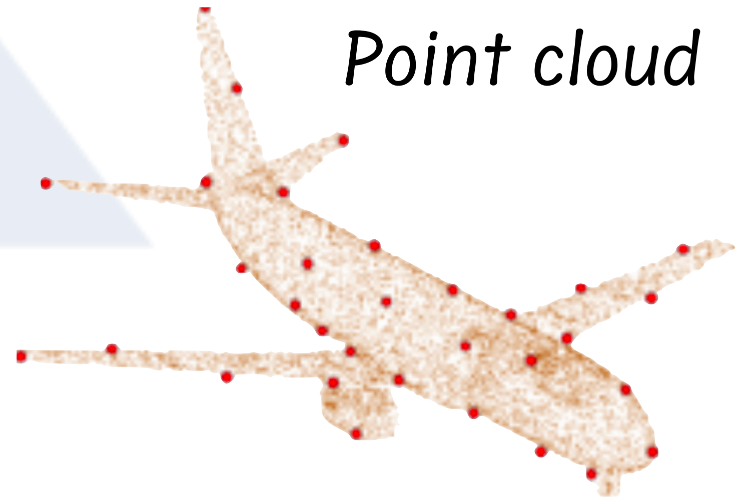


CNN is invariant only under translation. Suffer from the input sparsity.

Transformer Network

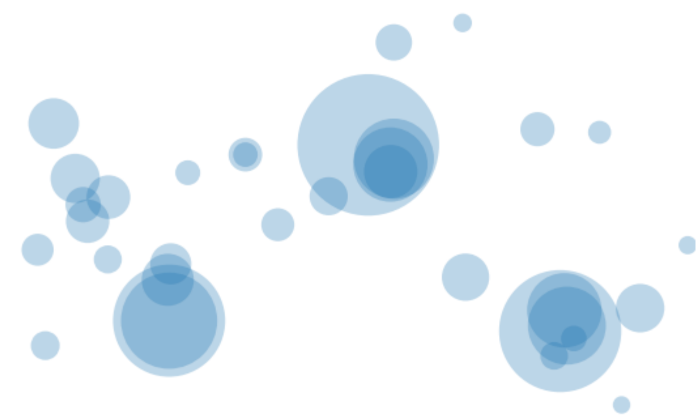
High computation cost

Point cloud



$$\mathrm{Attention}_{(i,j)} = \mathrm{softmax}\left(\frac{Q.K^T}{\sqrt{d}} \cdot V\right)$$

Particle cloud



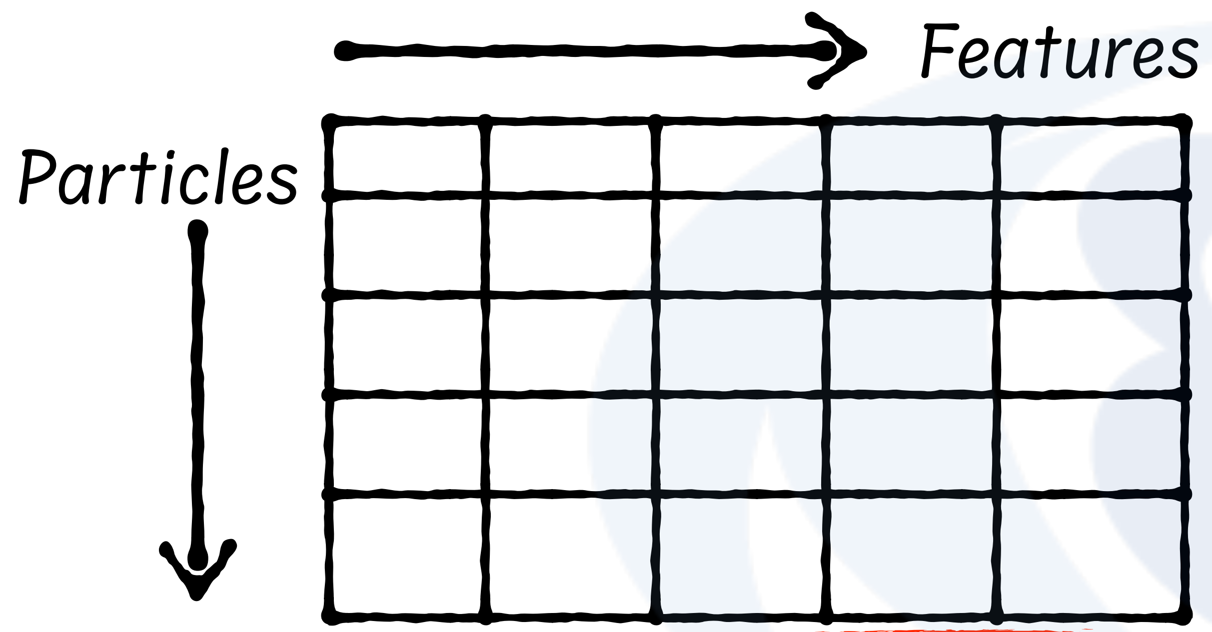
Query, key and value matrices mix up all the particle and feature tokens

Transformer networks are permutation invariant with no sparsity issue.

Transformer network

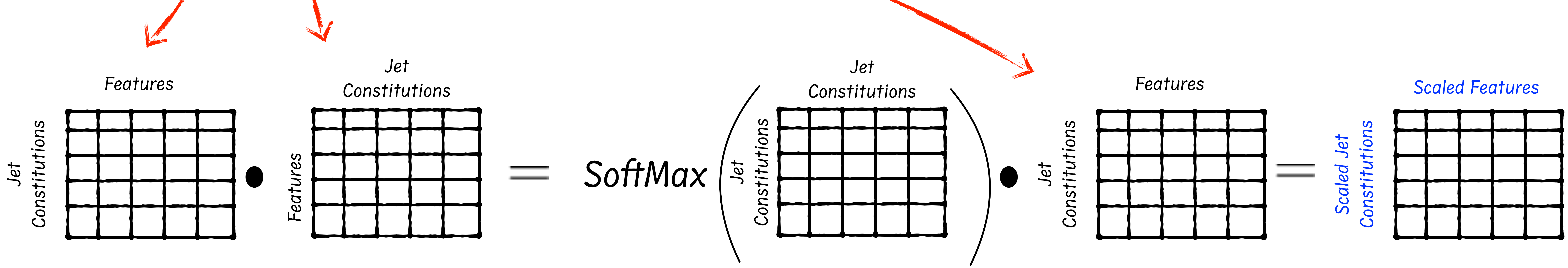
Transformer network mixes particle and feature tokens to highlight the most important tokens for the model decision making.
It allows for global and local information extraction.

Input data structured as a fixed size unordered grid



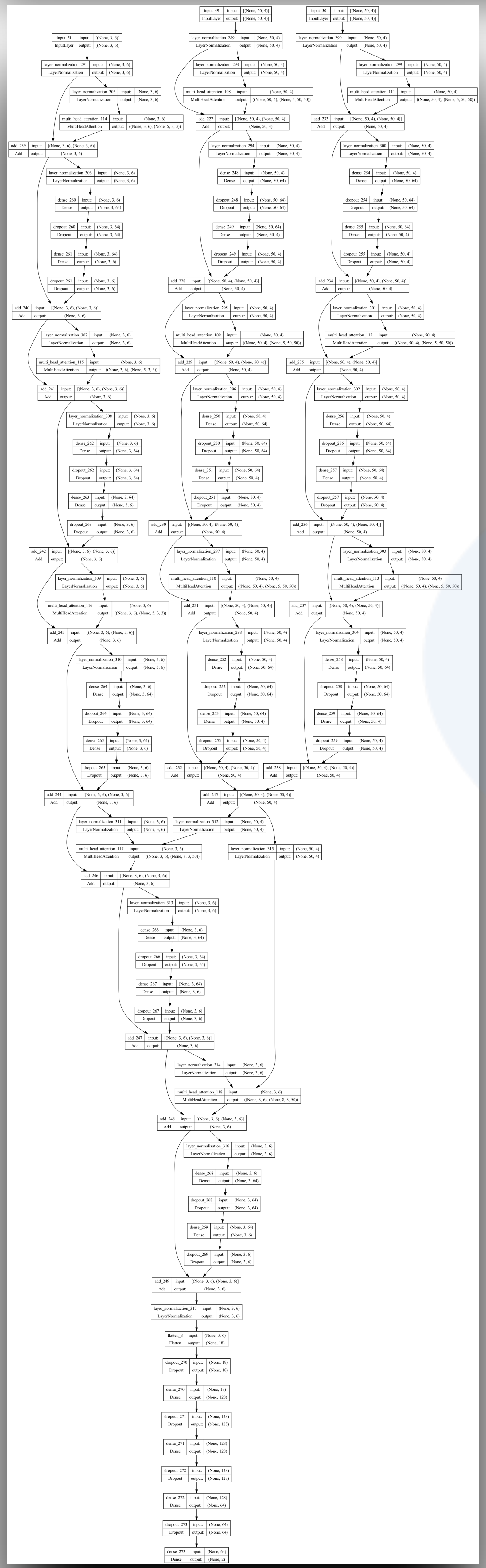
$$\mathrm{Attention}_{(i,j)} = \mathrm{softmax}\left(\frac{Q.K^T}{\sqrt{d}} \cdot V\right)$$

2010.11929

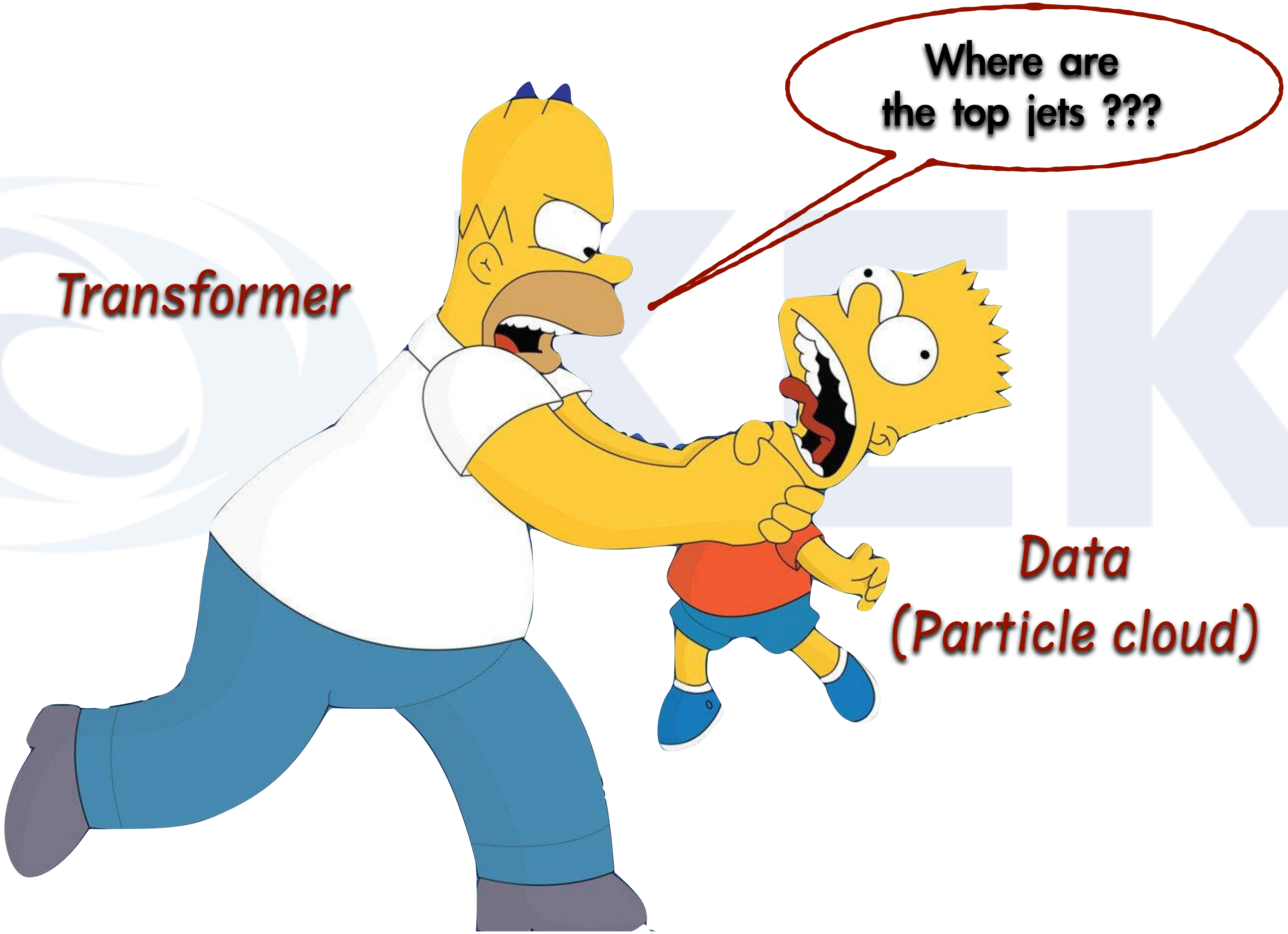


Transformer network

Transformers have the best performance but with high computational complexity.

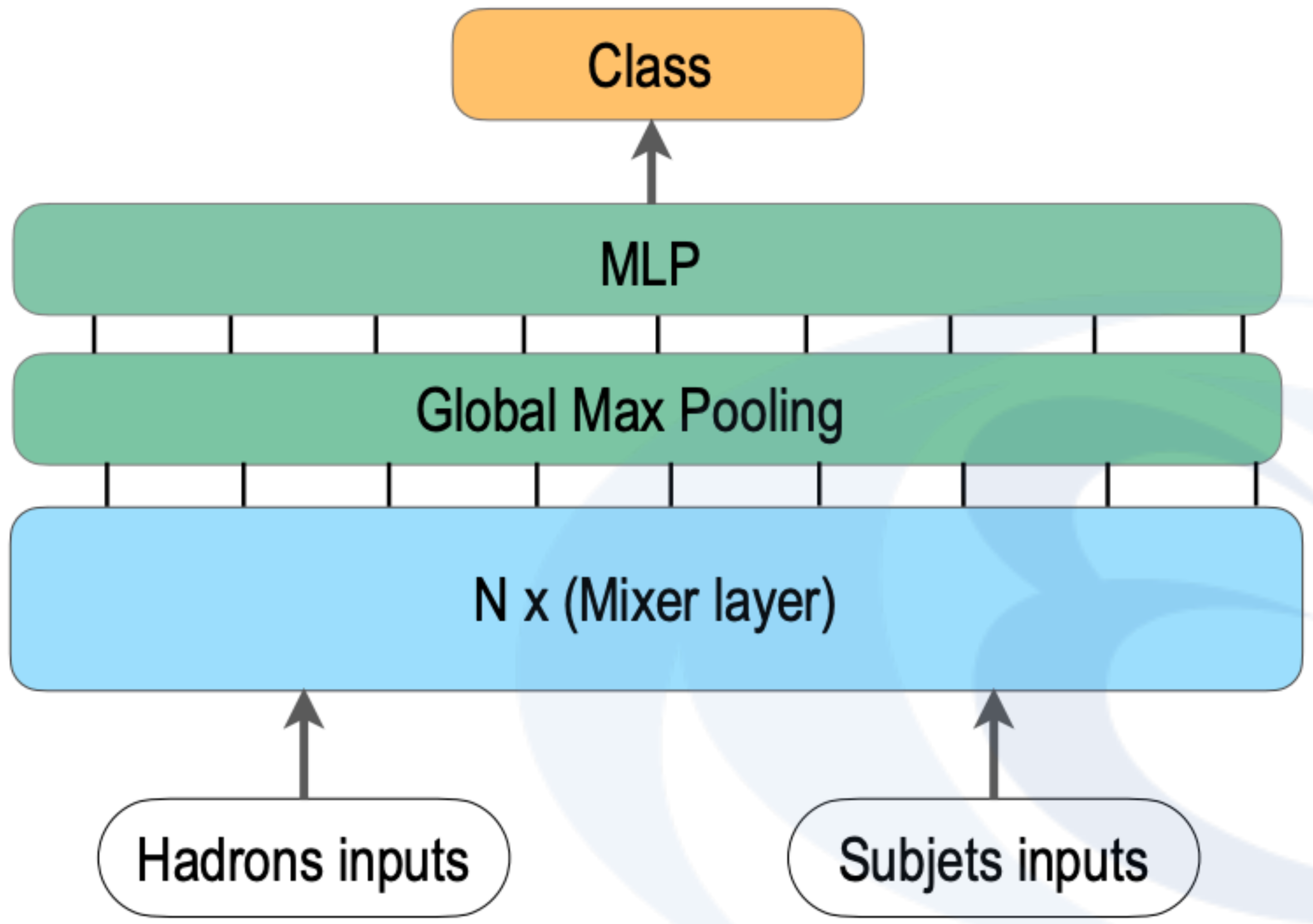


A.H, S.M, Nojiri
JHEP 03 (2024),2401.00452

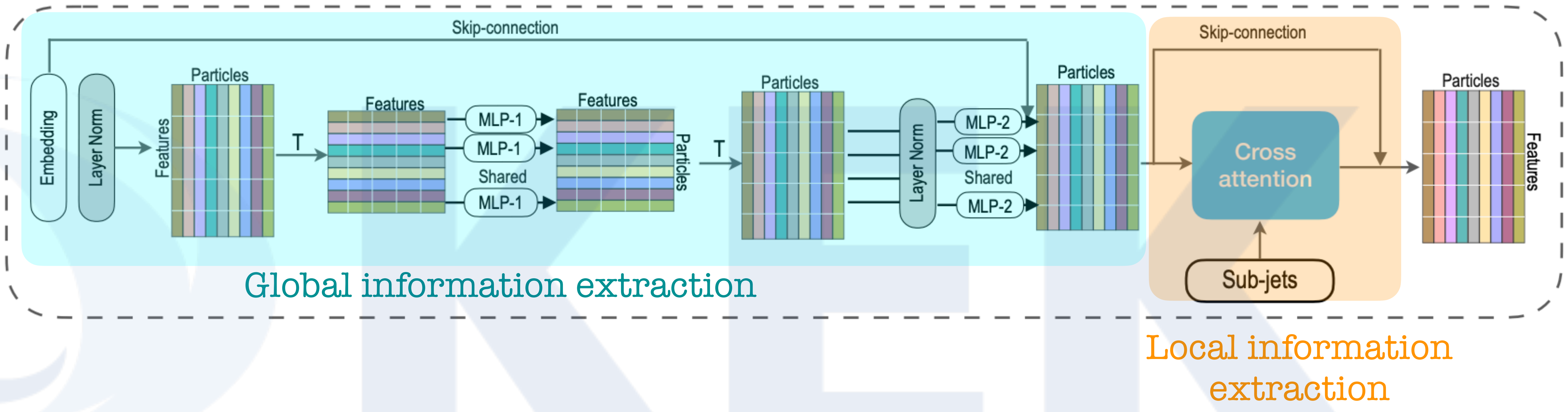


Mixer network

Mixer layer has two MLP that mix both features and Particle tokens (similar to the transformer) which allow for fast extraction of the global features of the event. Local information is extracted from the subjects via Cross-attention layer.



Mixer layer



Jet tagging task can be divided into two main parts:

- *Global information extraction*

The network learns how important each jet constituent to all other constituents via two MLPs.

$$Y_{i,j} = X_{i,j} + \left[\left(W_2 \sigma W_1 (\mathrm{LayerNorm}(\mathbf{X})^T)\right)^T\right]_{i,j},$$
$$\widetilde{X}_{i,j} = Y_{i,j} + \left(W_4 \sigma W_3 (\mathrm{LayerNorm}(Y_{i,j}))\right)$$

- *Local information extraction*

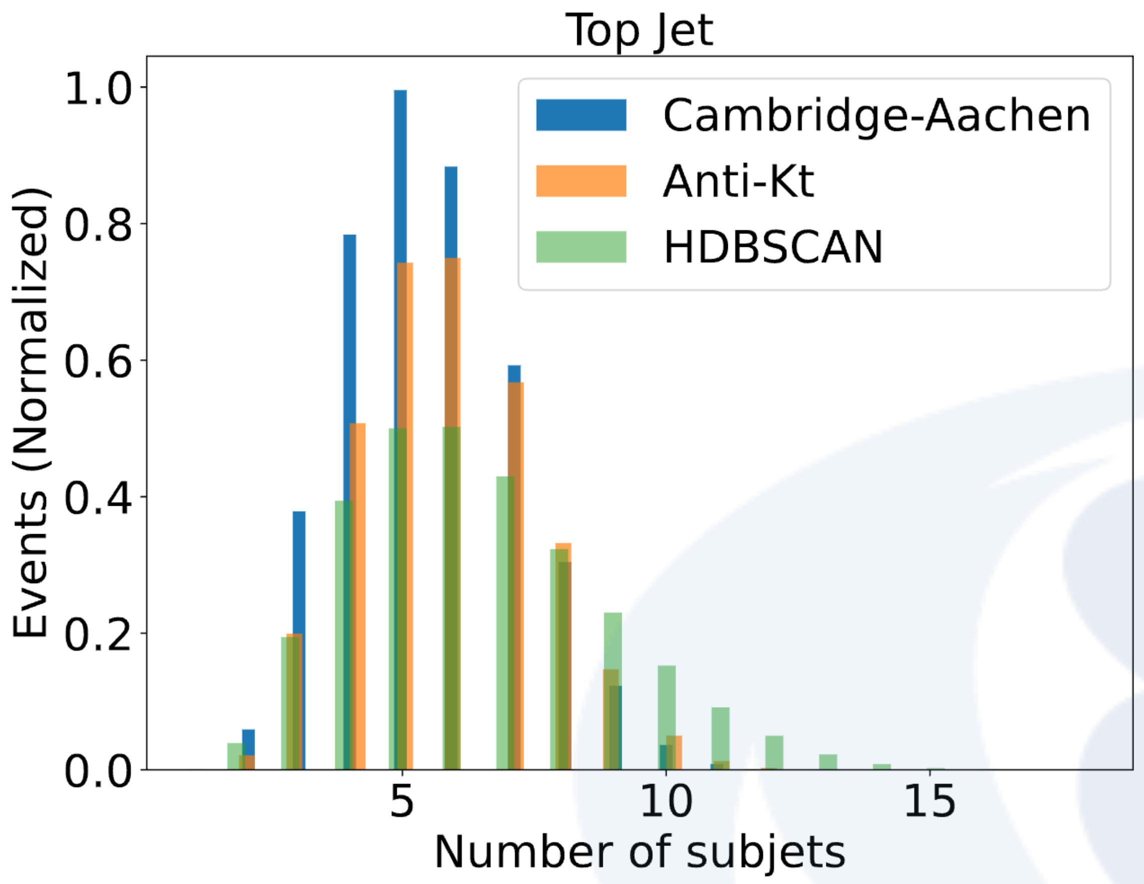
The network learns how important each jet constituent to the sub-jet it belongs to

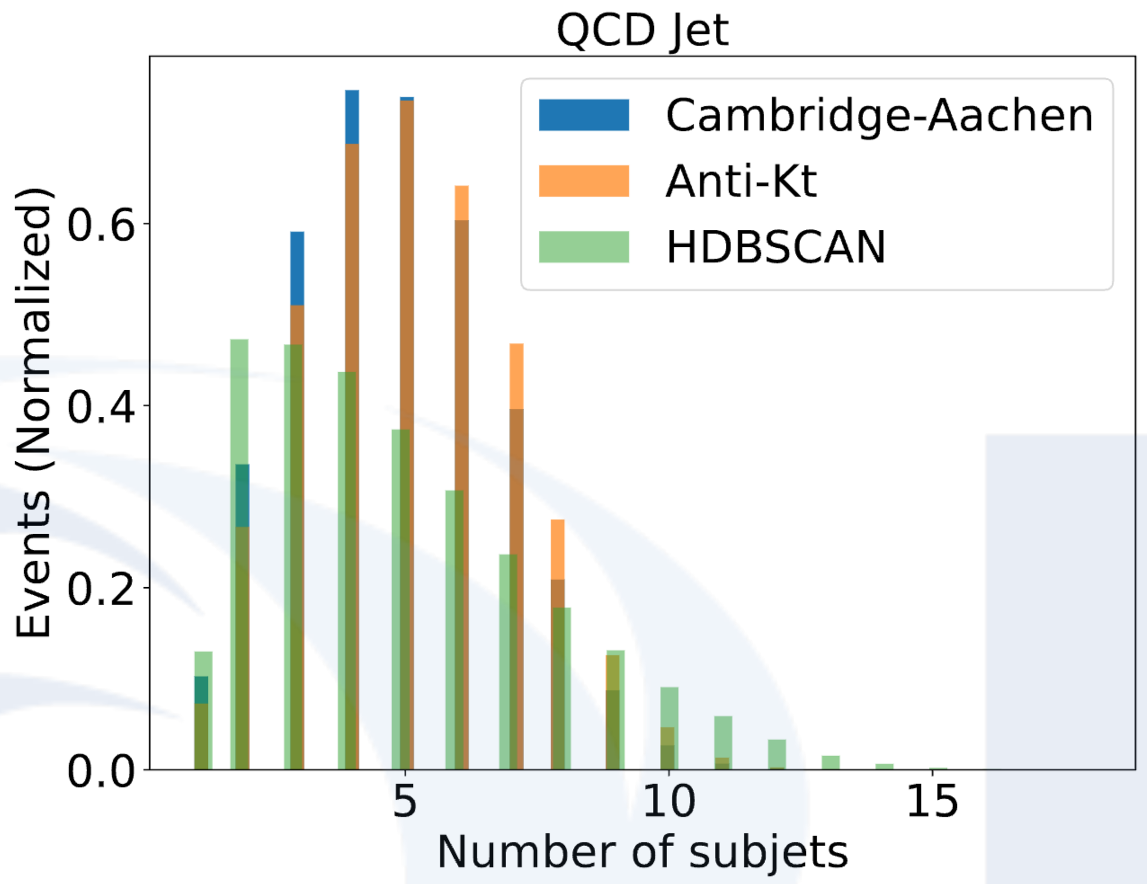
$$P\,(\text{Hadrons inside jet} \mid \text{subjet cluster}) = P\,(x_i | y_\alpha)$$

Validation on Top dataset

The network is validated on TopLandscape community dataset
https://zenodo.org/records/2603256

R = 0.3 for CA and anti-KT





For subjets clustering, three methods are used:

- Cambridge-Achen
- Anti-Kt
- Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

IRC safe, Non parametric clustering with dynamic R

A.H, M.Nojiri
arXiv:2404.14677, JHEP 06 (2024) 176

TABLE I. performance of the Mixer network for top quark tagging compared with other models. Results for PFN [36], ParticleNET [37], and ParT [38] are quoted from their published results. Transformer(subjet) model is trained from scratch using the CA subjets dataset only. Training time is per epoch with a batch size of 1024. The GPU training time is measured on an NVIDIA RTX A6000 card.

	AUC	$Rej_{50\%}$	Parameters	Time (GPU) [s]
PFN	0.9819	247	86.1K	30
ParticleNET	0.9858	397	370K	729
ParT	0.9858	413	2.14M	612
Transformer(subjets)	0.9640	186	398k	129
Mixer(Anti-kt)	0.9854	375	86.03k	33
Mixer(CA)	0.9856	392	86.03k	33
Mixer(HDBSCAN)	0.9859	416	86.03k	33

Mixer network can achieve high performance as the transformer network with low computational cost

Results interpretation

Do the two components of the mixer layer capture different information?

Central kernel alignment

Analysis of the hidden Layers representation

CKA computes the the similarity of the hidden layers representations independent on the size of each layer.

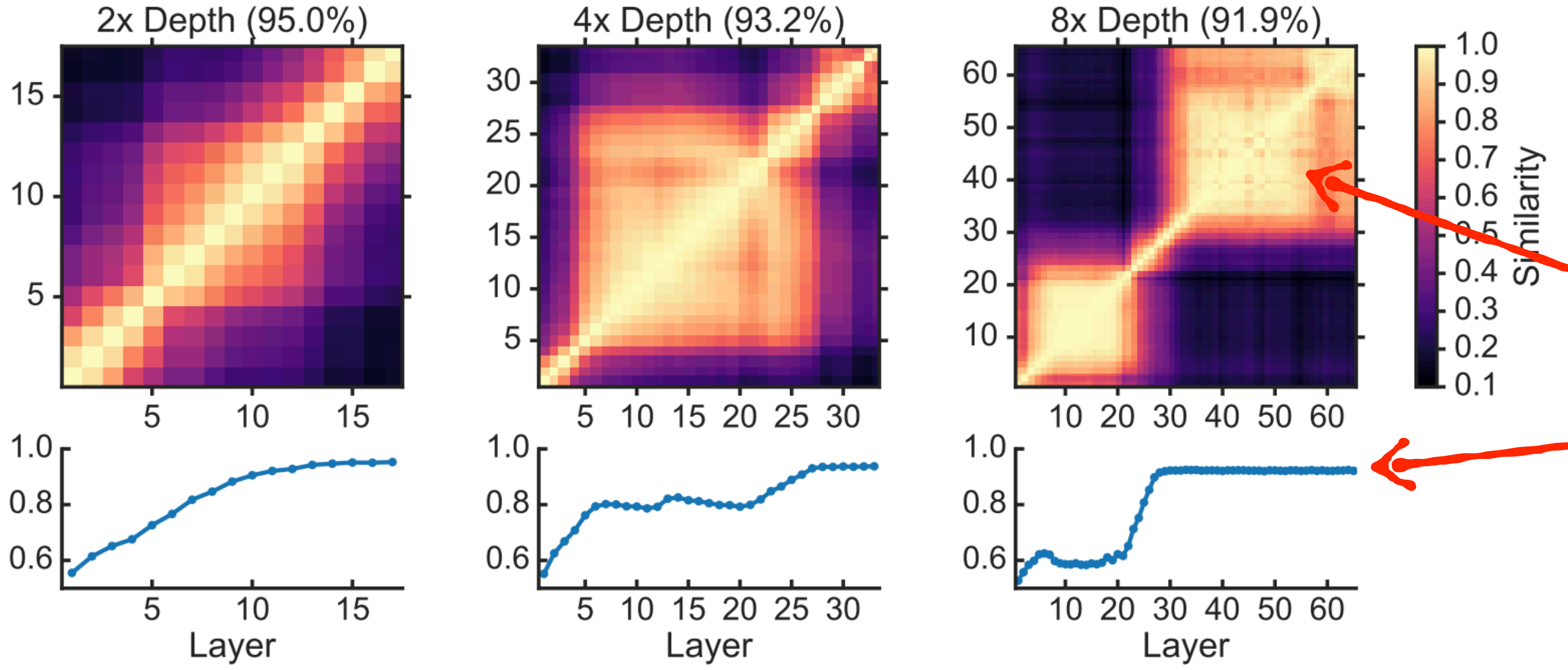
Two inputs from two hidden layers with the form $X \in \mathbb{R}^{d\times P_1}$ $\quad Y \in \mathbb{R}^{d\times P_2}$

Gram matrix can be constructed which is independent on the dimension of the hidden layer $M = XX^\top$ and $N = YY^\top$

$$\boxed{\mathrm{CKA}(M,N) = \frac{\mathrm{HSIC}(M,N)}{\sqrt{\mathrm{HSIC}(M,M)\mathrm{HSIC}(N,N)}}}$$

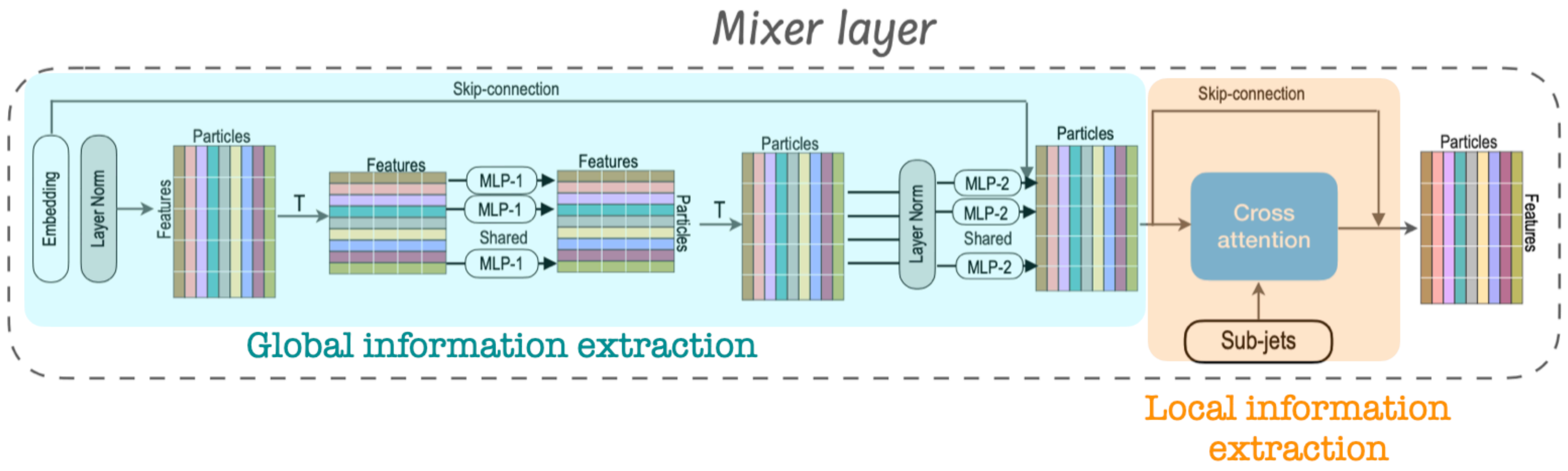
With $\mathrm{HSIC}(M,N) = \frac{1}{(d-1)^2}\, tr(MHNH)$

arXiv:1905.00414

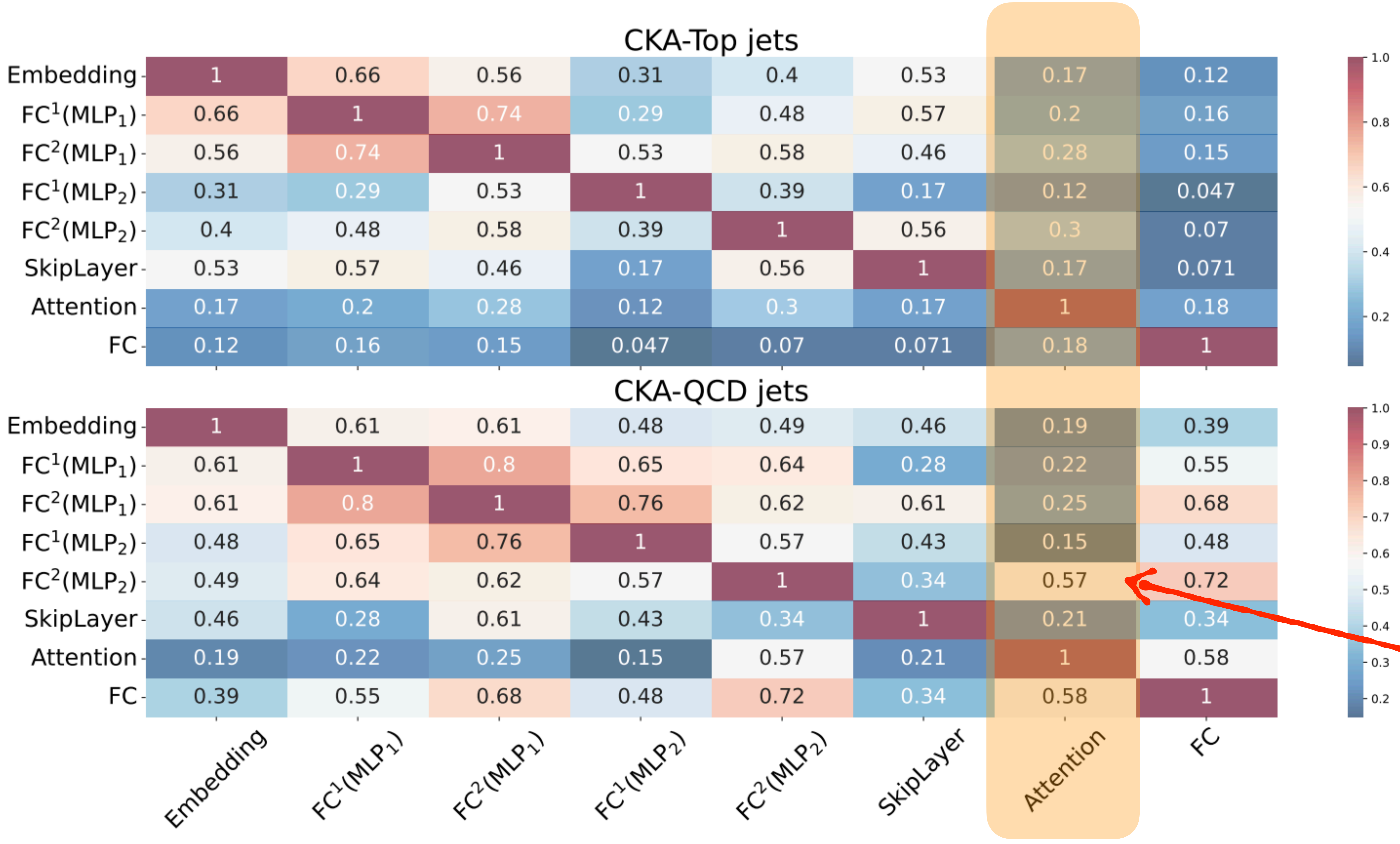


Layers with similar CKA value, learns the similar information

Central kernel alignment



CKA similarity for 5000 test events for the Top and QCD jets



For both data, top and QCD, the cross-attention layer has different CKA value from the two MLPs

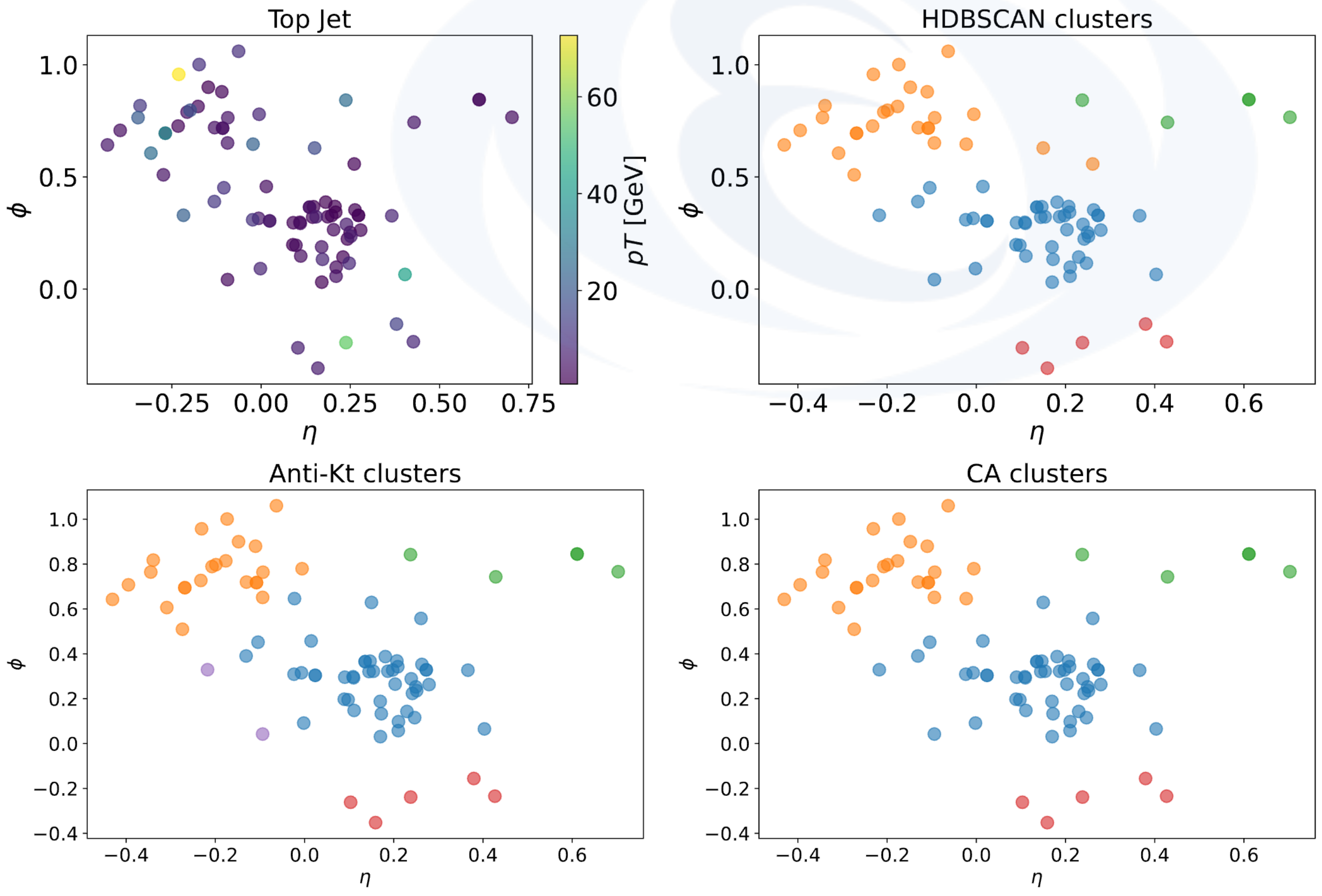
For the QCD, there is no clear prong structure. This is why attention layer capture global information similar to the MLPs

Thank you

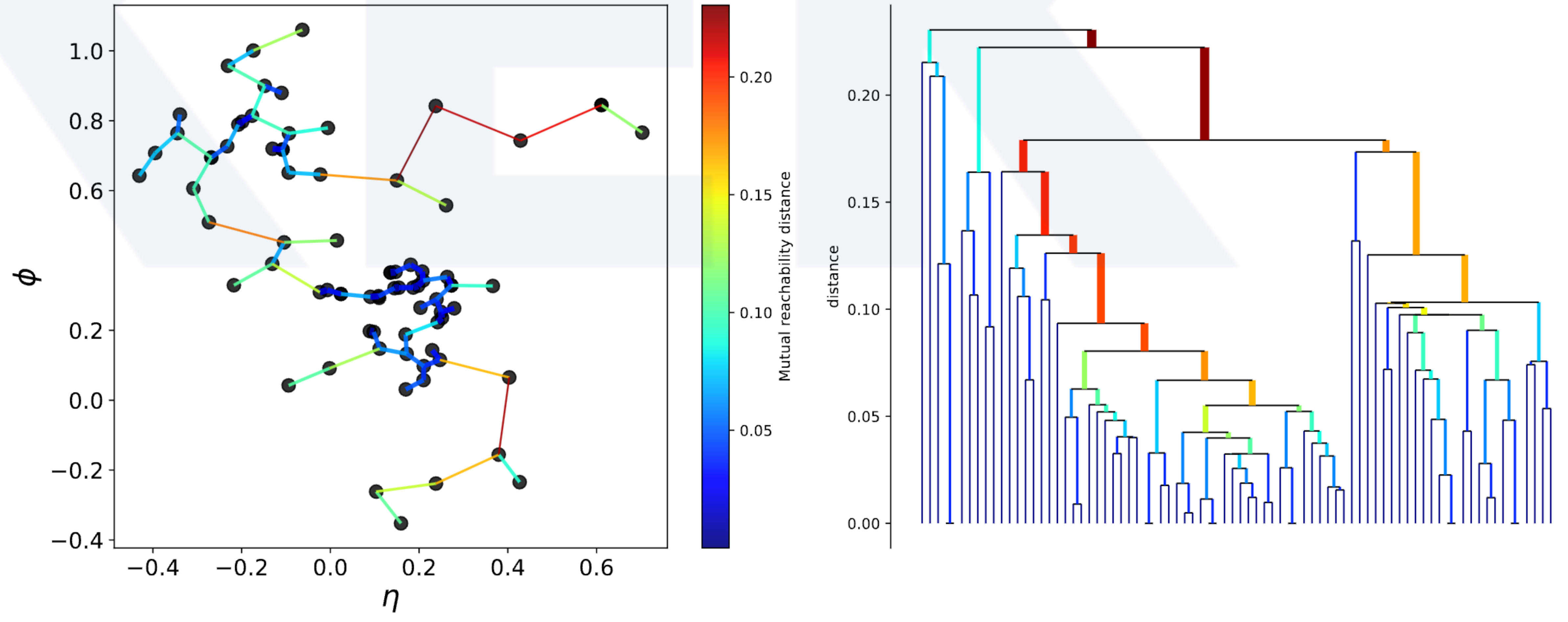
for your attention

PackUP

Example of top jet clustering by different methods

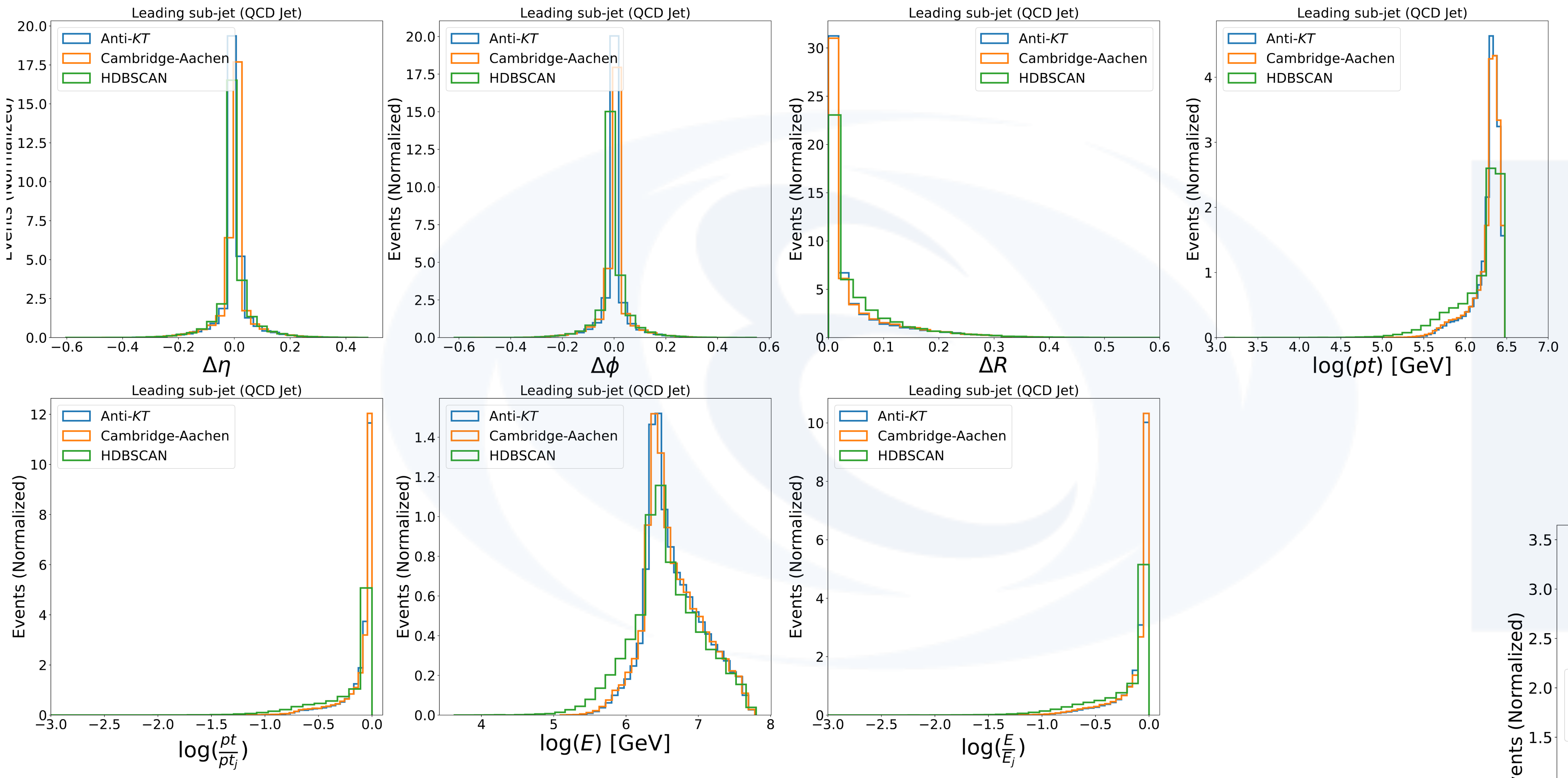


Minimum spanning tree by the HDBSCAN



PackUP

Subjets kinematics for the QCD events



Subjets kinematics for the top events