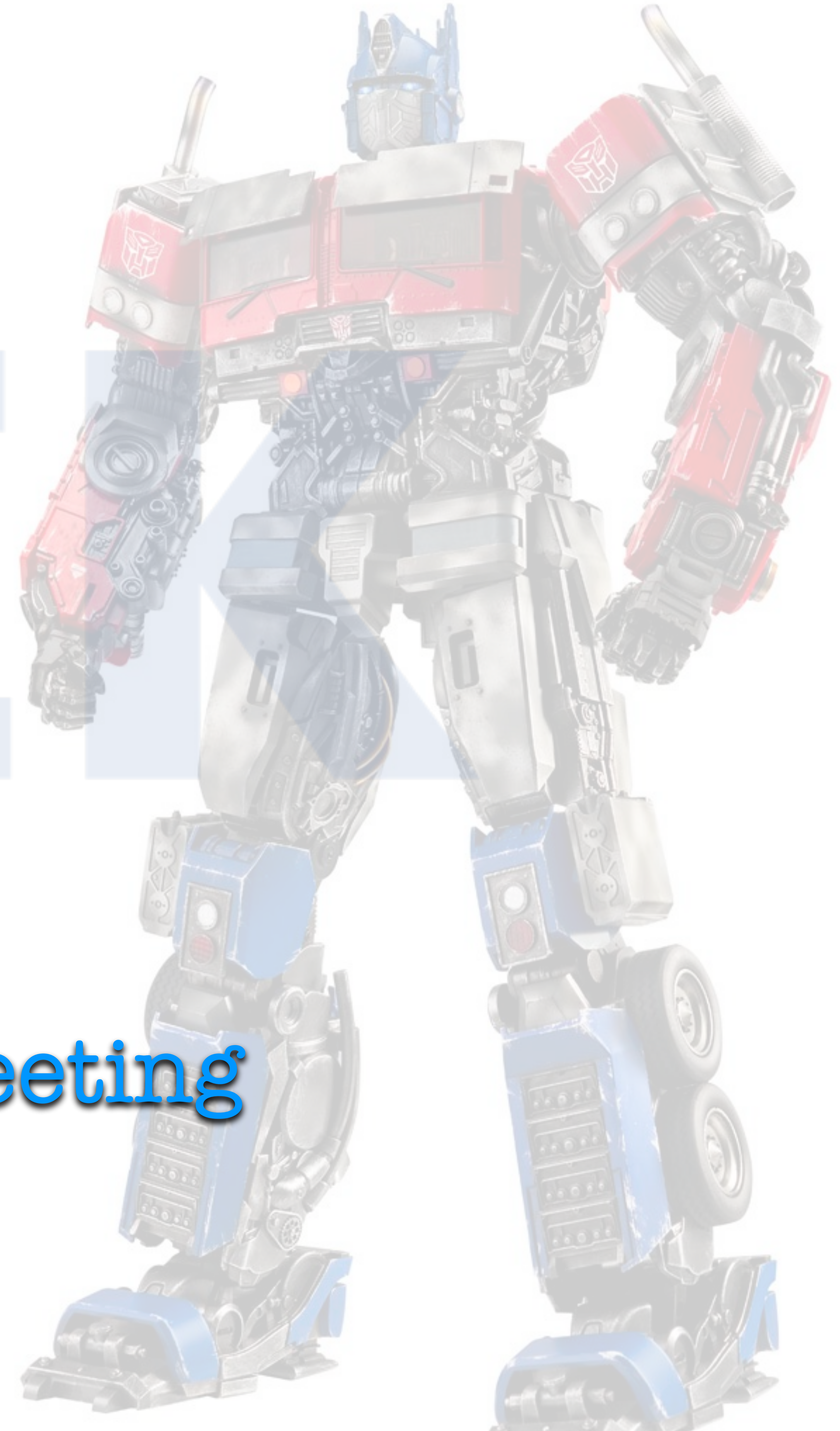
Multi-Scale Cross-Attention Transformer Encoder for Event Classification

Speaker: *Ahmed Hammad*

Collaborators: *Stefano Moretti and Mihoko Nojiri*

Theory center, KEK, Japan



Extended Higgs Sector subgroup meeting

19th November 2024

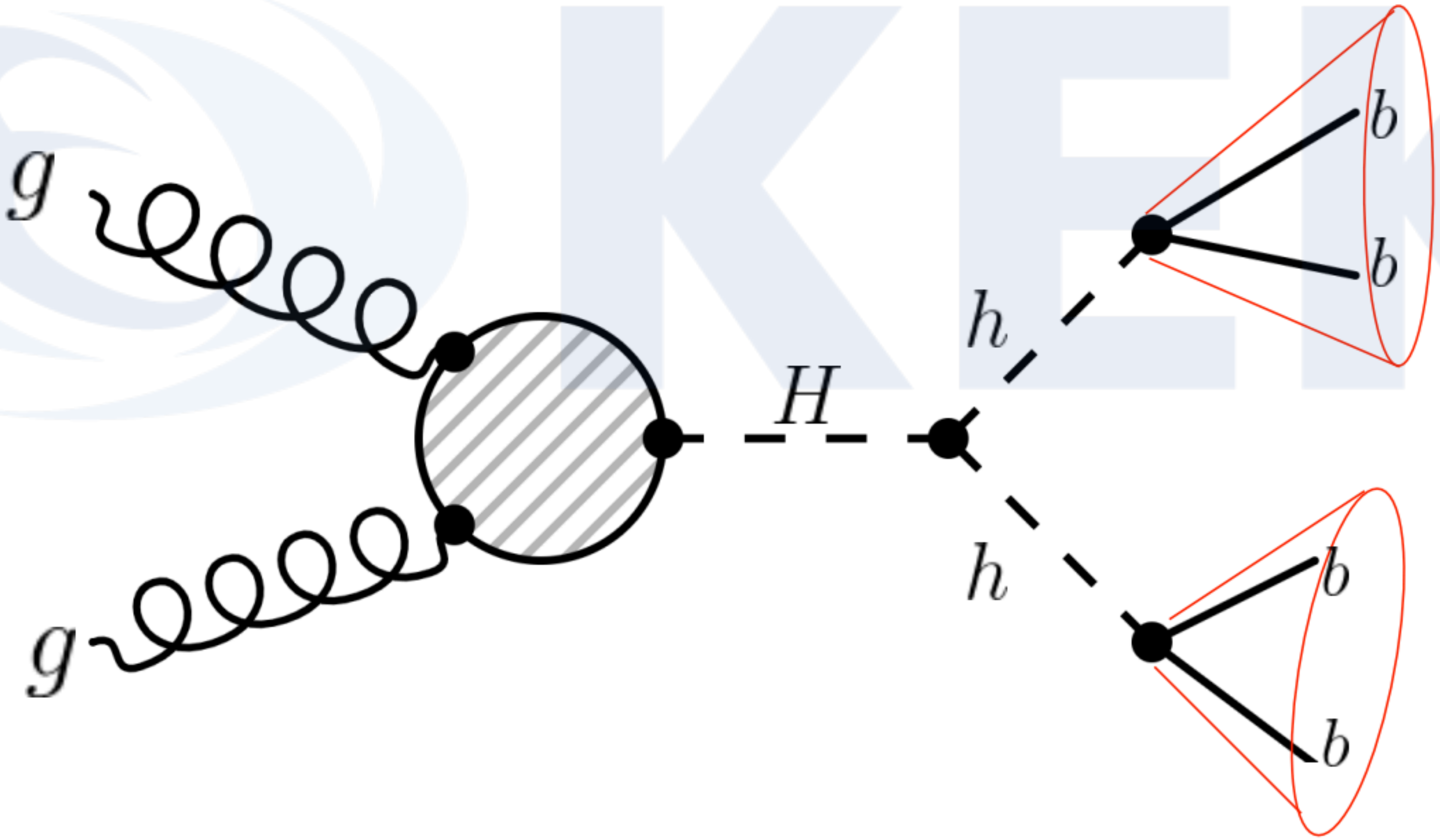
Outline

- *Multi-scale transformer for di-Higgs analysis*
    - *Multi-heads self-attention*
    - *Multi-heads cross-attention*
- *Interpretable AI methods*
    - *Attention Maps*
    - *Gradient Weighted Class Activation Mapping (Grad-CAM)*

Introduction

These slides based on: arXiv:2401.00452 [JHEP 03 (2024) 144]

In which we utilized Transformer encoders for resonant di-Higgs analysis at the HL-LHC



Boosted Higgs regime

Introduction

$$pp \to e^- e^+ jj$$

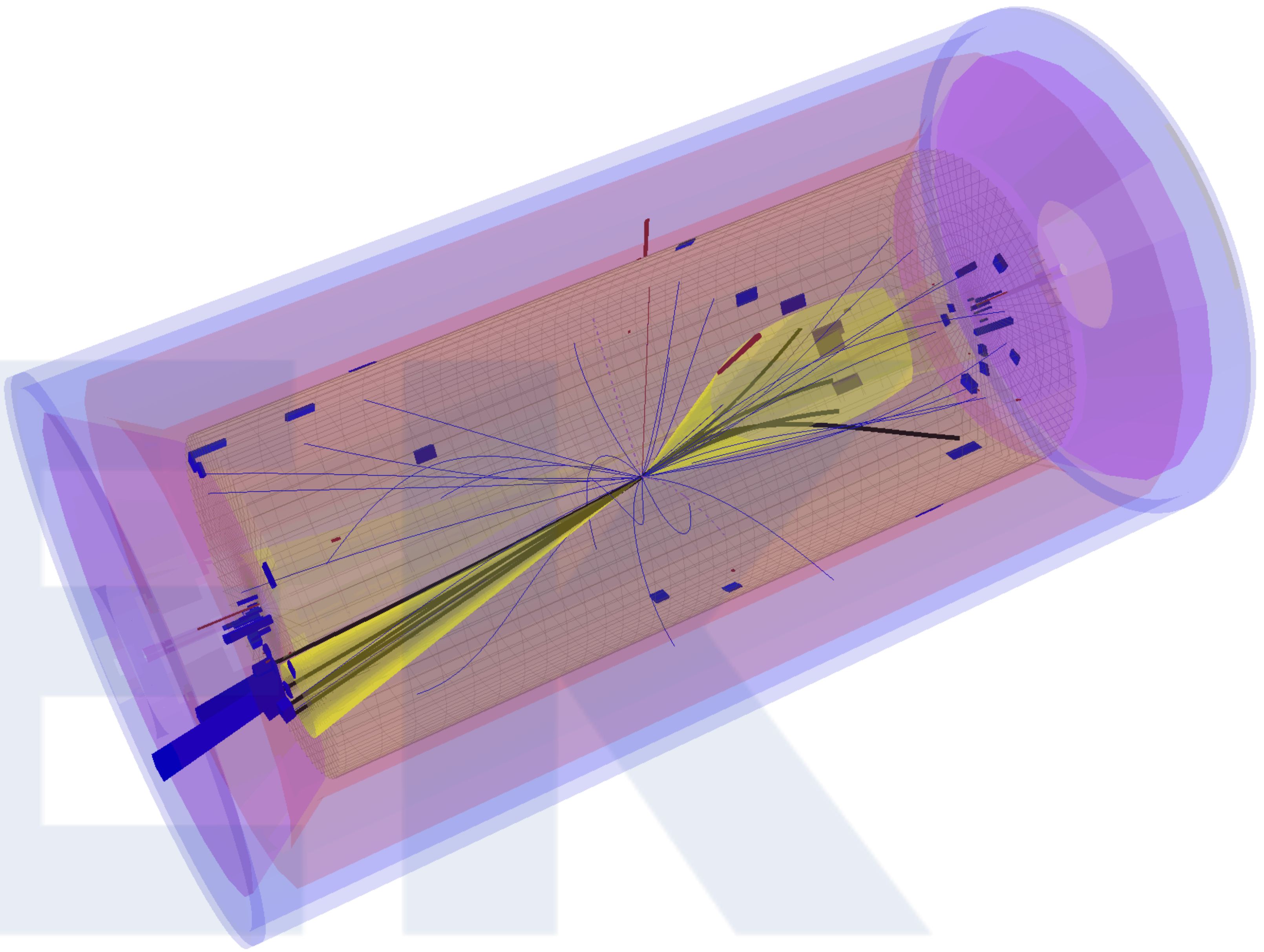
Events collisions at the LHC are characterized via their features:

- ## *Global features:*

  Kinematics of the hard process as masses, momenta, helicity correlations, etc, encode the global features for the final state particles. Together with the kinematics for the resonantly mediated particles, global features span the entire phase space.

- ## *Local features:*

Local features are extracted from the properties of the particles confined by the jet boundaries, e.g. momentum of jet constituents, pseudo rapidity of the jet constituents, etc. This information doesn`t span the entire phase space and localized inside the jet boundaries



Simulated with Delphes Event display

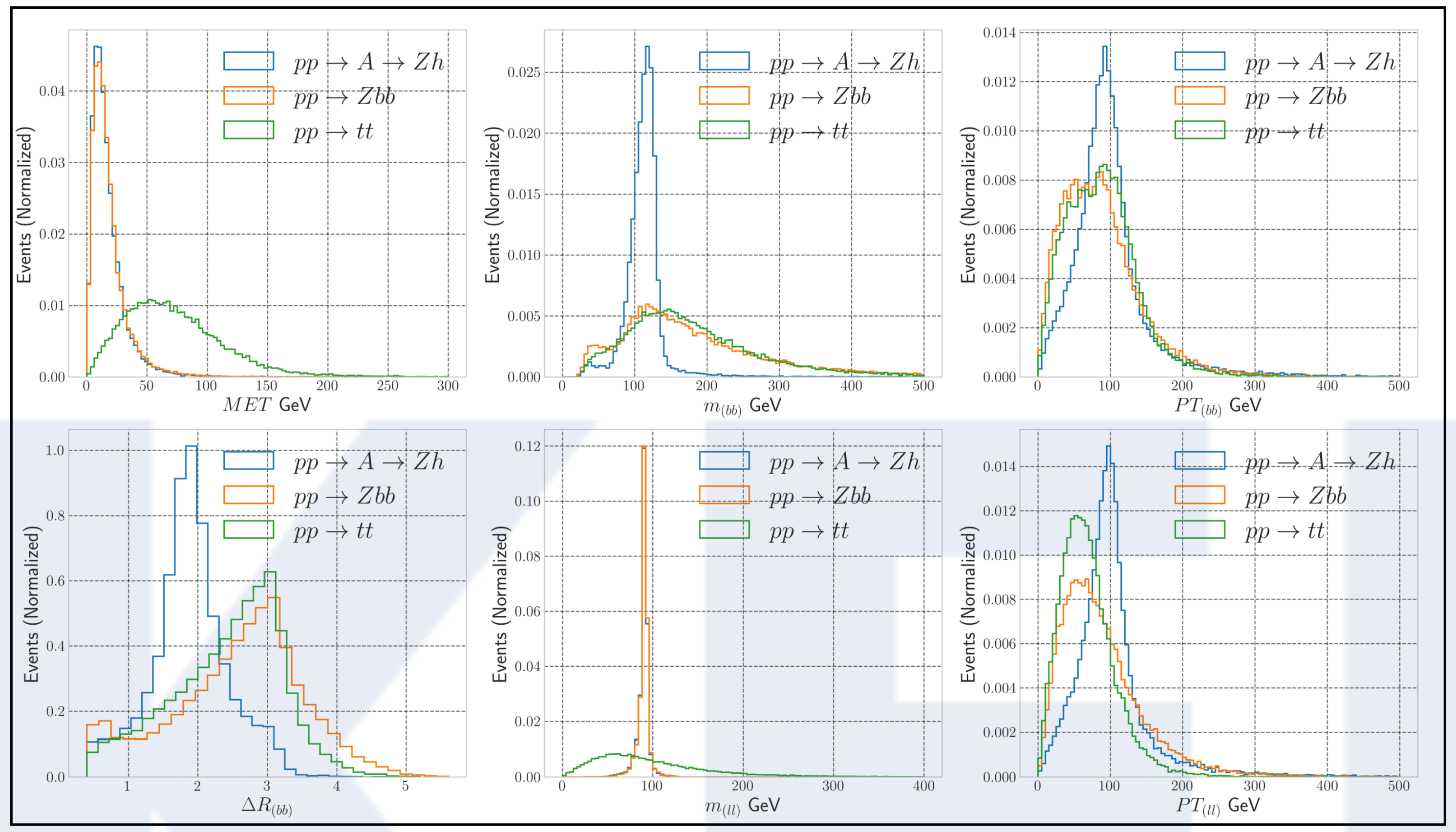
Introduction

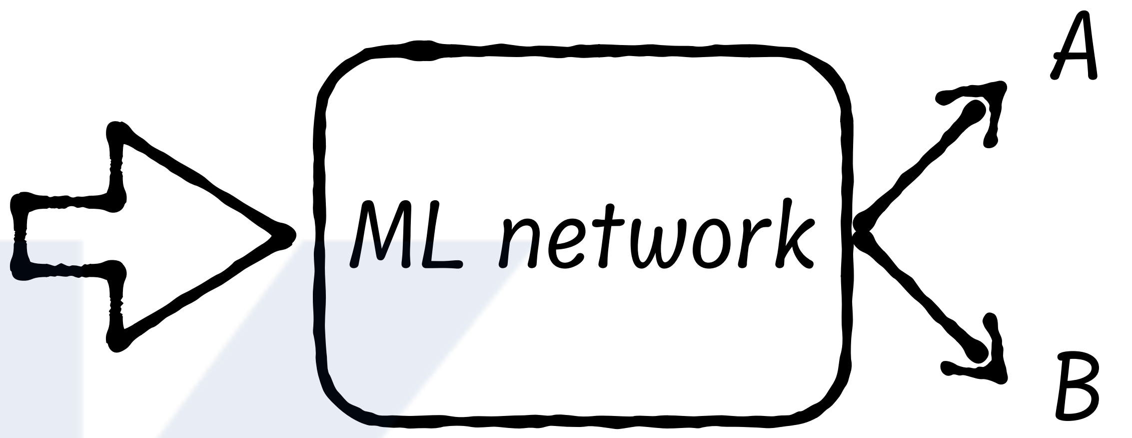
For Event classification one can use:

- ***Kinematics:***

High-level reconstructed kinematics of the hard process can be used to Classify signal from background events

Arxiv:2305.13781

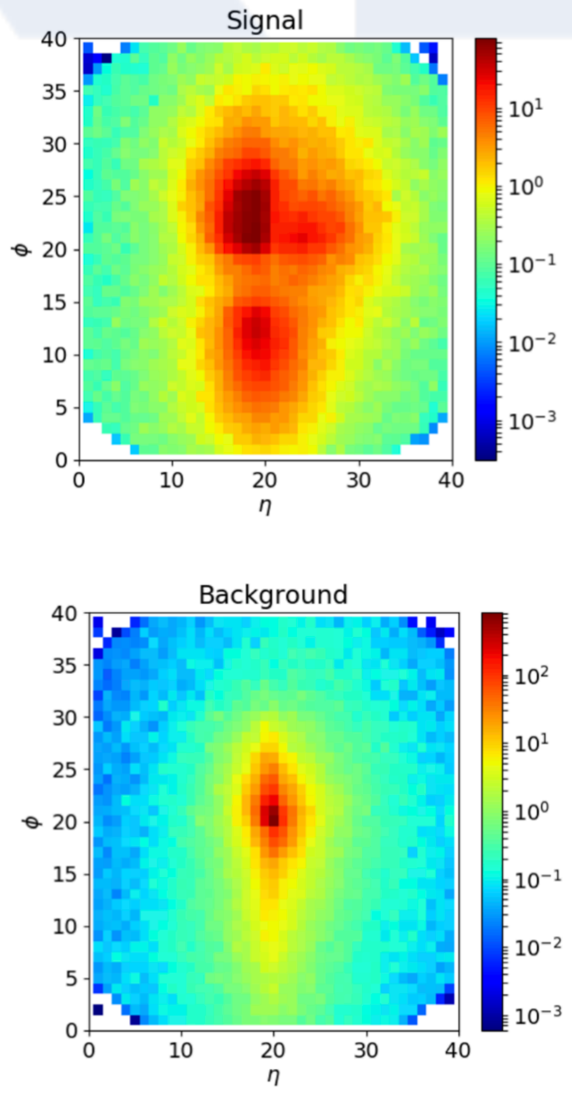




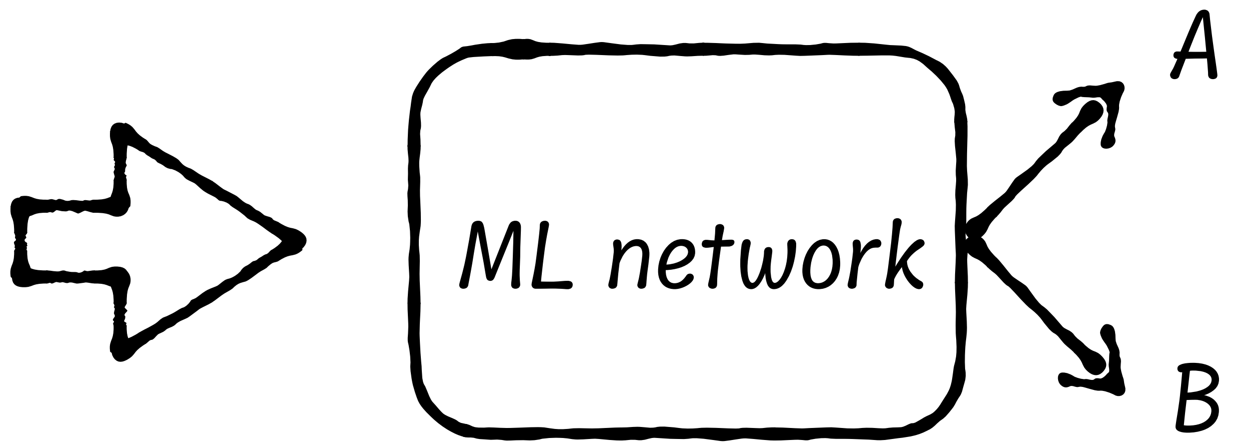
- ***Jet sub-structure:***

Multi-prong structure of jets can be used for classification.

The 3 prong structure of the top jet can be used to distinguish events with top jet from QCD jet processes



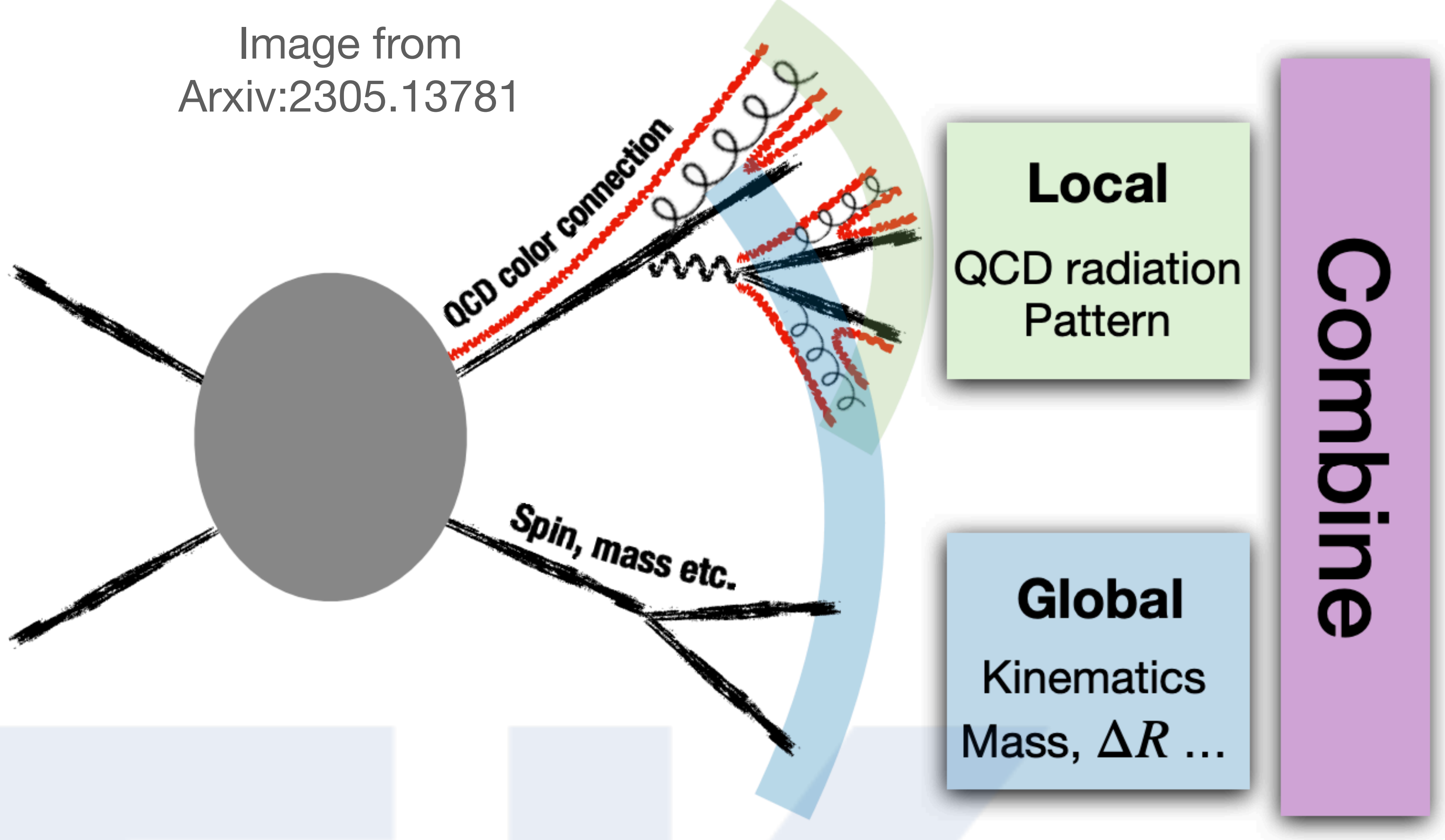
Arxiv:1902.09914



Introduction

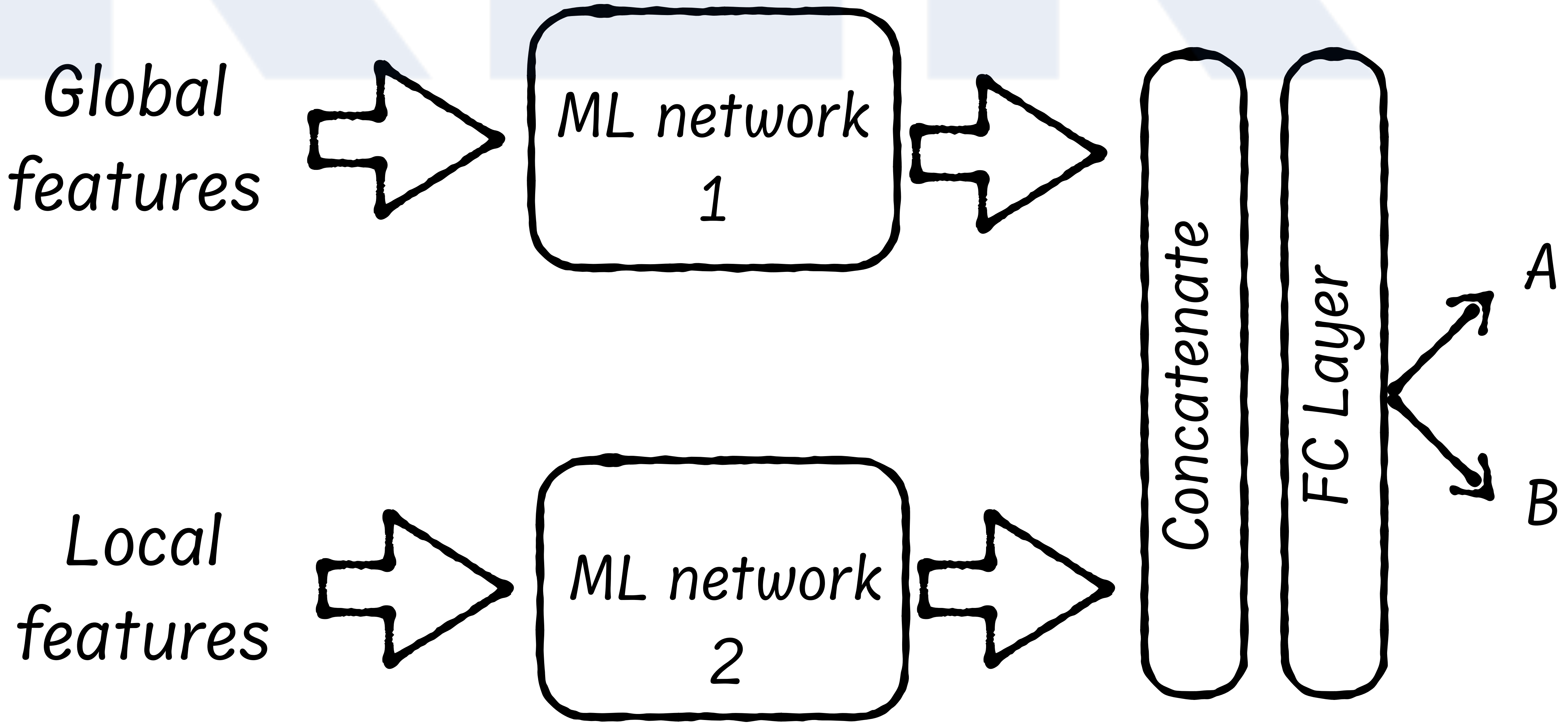
To improve the classification performance we need to use both information in the same time.

Image from Arxiv:2305.13781



- *Global + local features:*

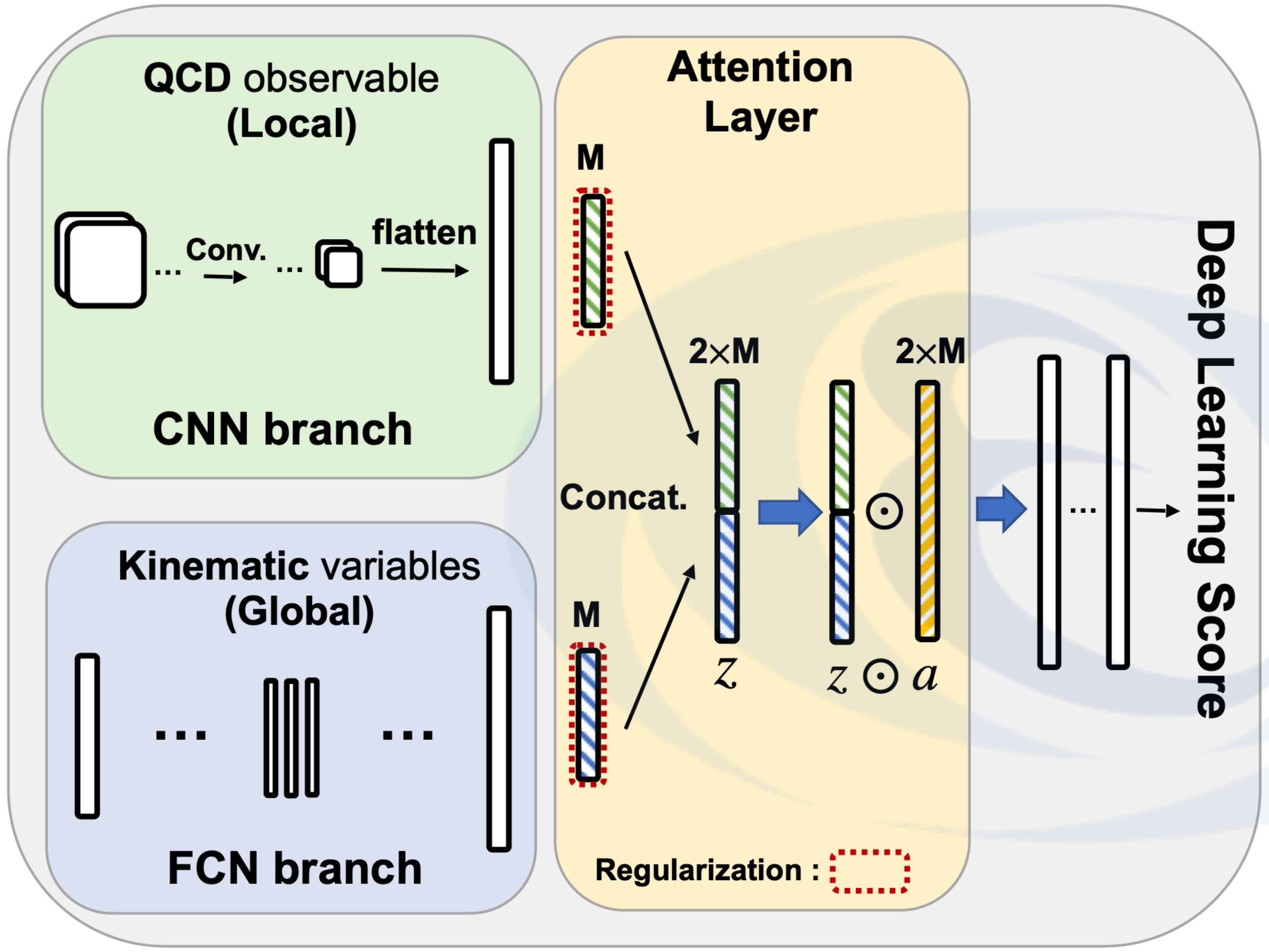
Both local and global information can be feed to multi-modal network with two streams.
Each network extract the characteristic features of each input data before they concatenated in one dense layer.
The concatenated features is then analyzed with one fully connected layer before the output layer.



Introduction

For a simple concatenation the global information encoded by the high level kinematics dominates over the local information. Thus, no much improvement from incorporating the jet information!!

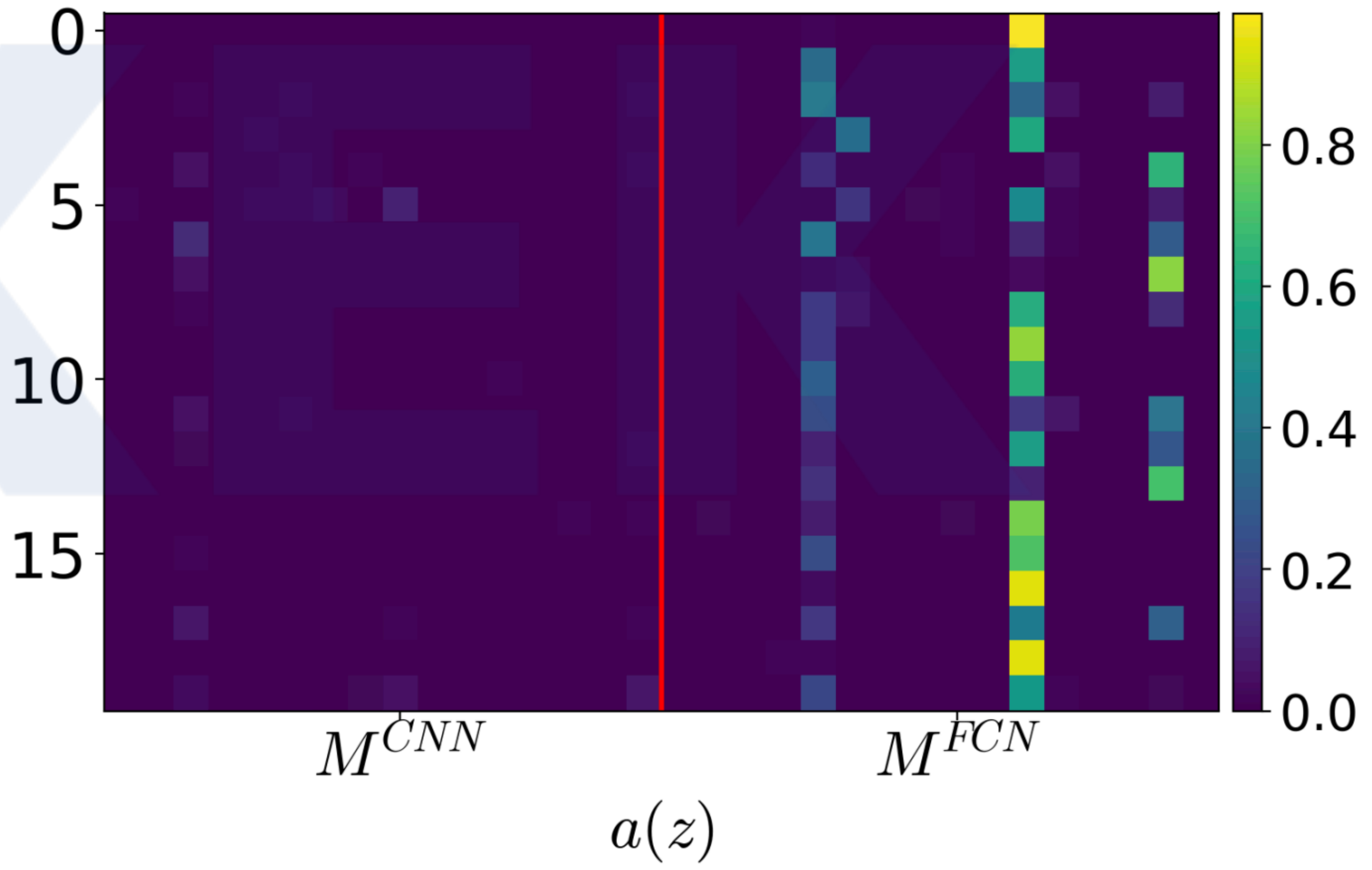
Arxiv:2305.13781



In this paper, the authors used CNN to extract the information from the QCD and MLP to analyze the high level reconstructed kinematics. The output is then concatenated in a single layer, Z.

$z \odot a$ *indicates the important information the model focuses on to make predictions*

Arxiv:2305.13781



For simple concatenation the model totally ignores the extracted information form the QCD and focuses only On the global information extracted from the kinematics

Transformer Encoders

An alternative way is to use Multi-scale Transformer.

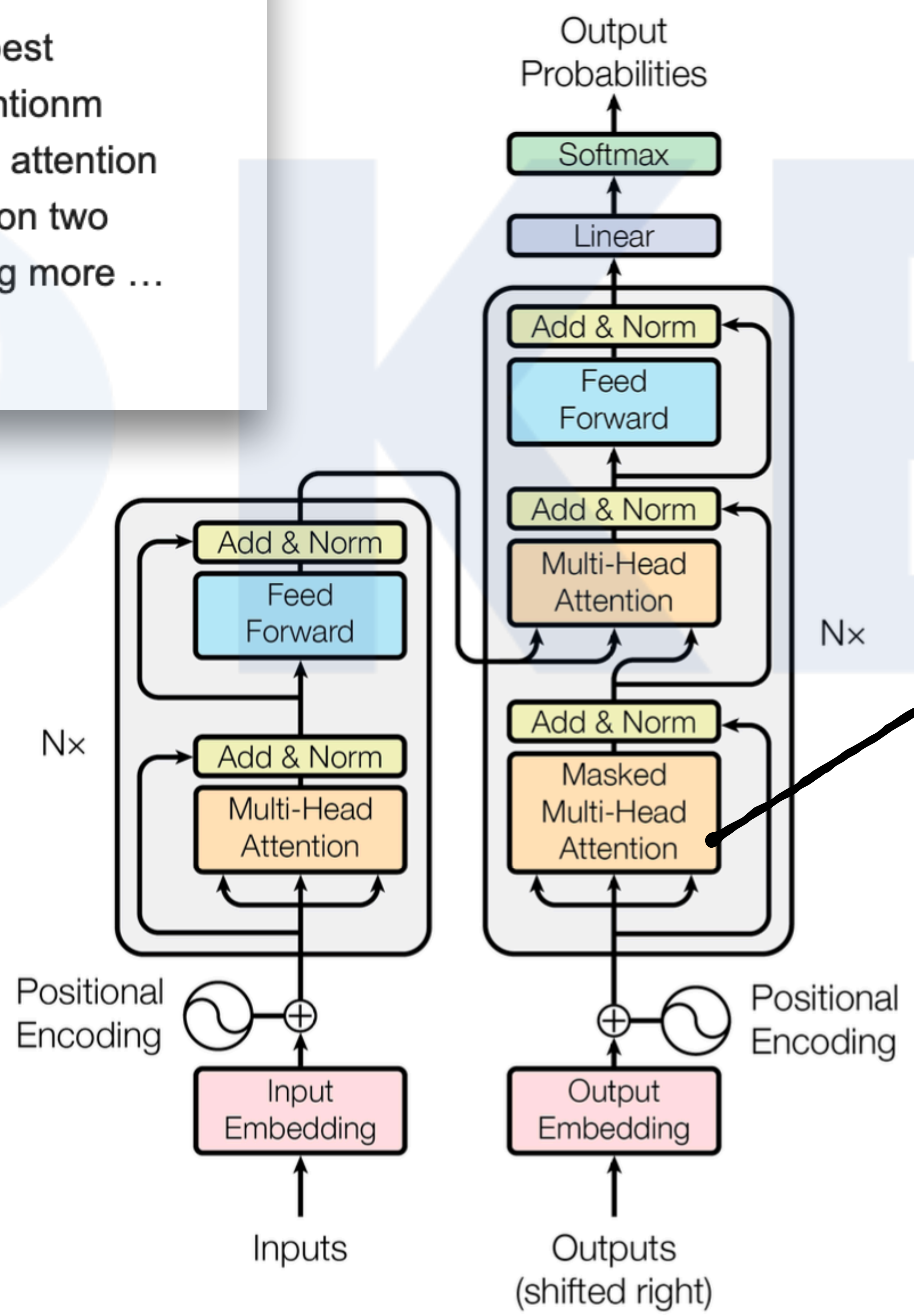
[PDF] Attention is all you need

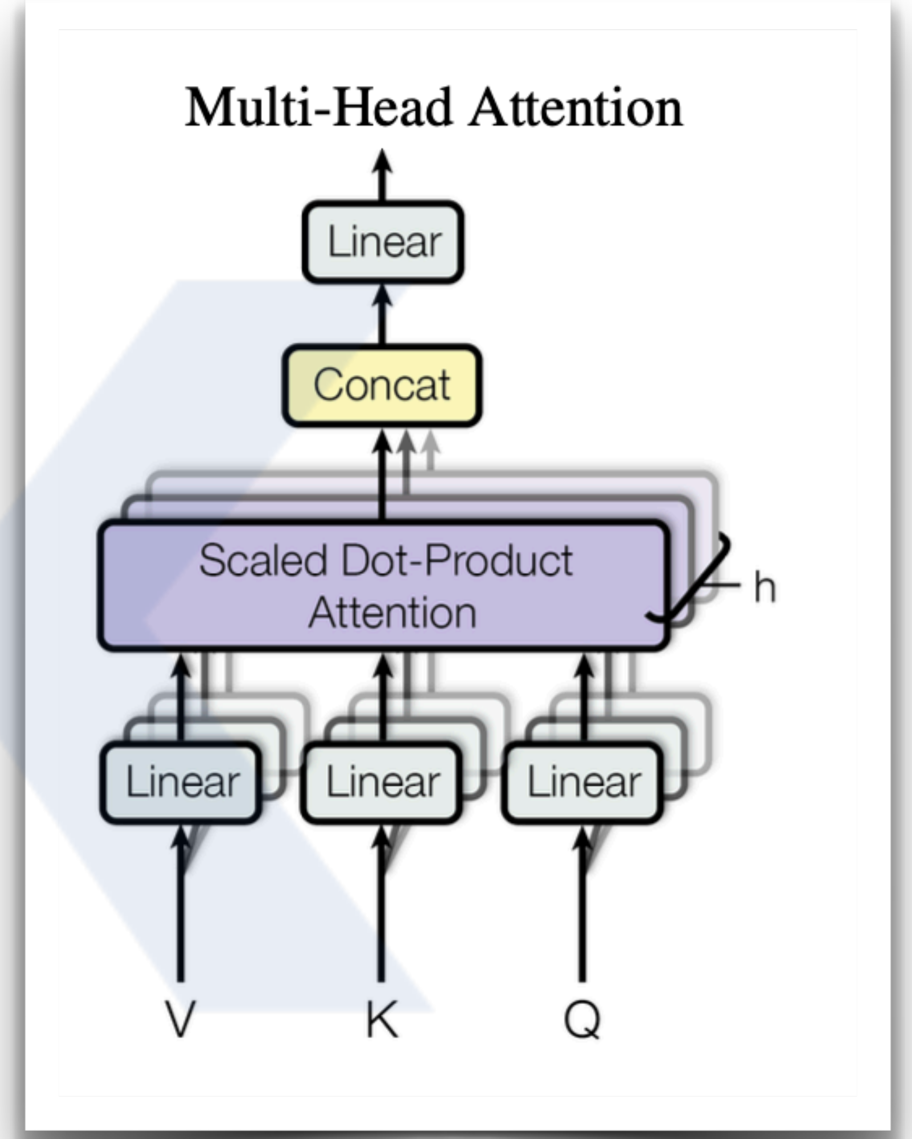
A Vaswani - Advances in Neural Information Processing Systems, 2017 - user.phil.hhu.de

The dominant sequence transduction models are based on complex recurrent orconvolutional neural networks in an encoder and decoder configuration. The best performing such models also connect the encoder and decoder through an attentionm echanisms. We propose a novel, simple network architecture based solely onan attention mechanism, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superiorin quality while being more …

☆ Save 99 Cite Cited by 137673 Related articles

Transformer model is first introduced in "attention is all you need" for language modeling. At this moment, more than 100,000 citation





Inputs are Query, Key and Value

Self Attention

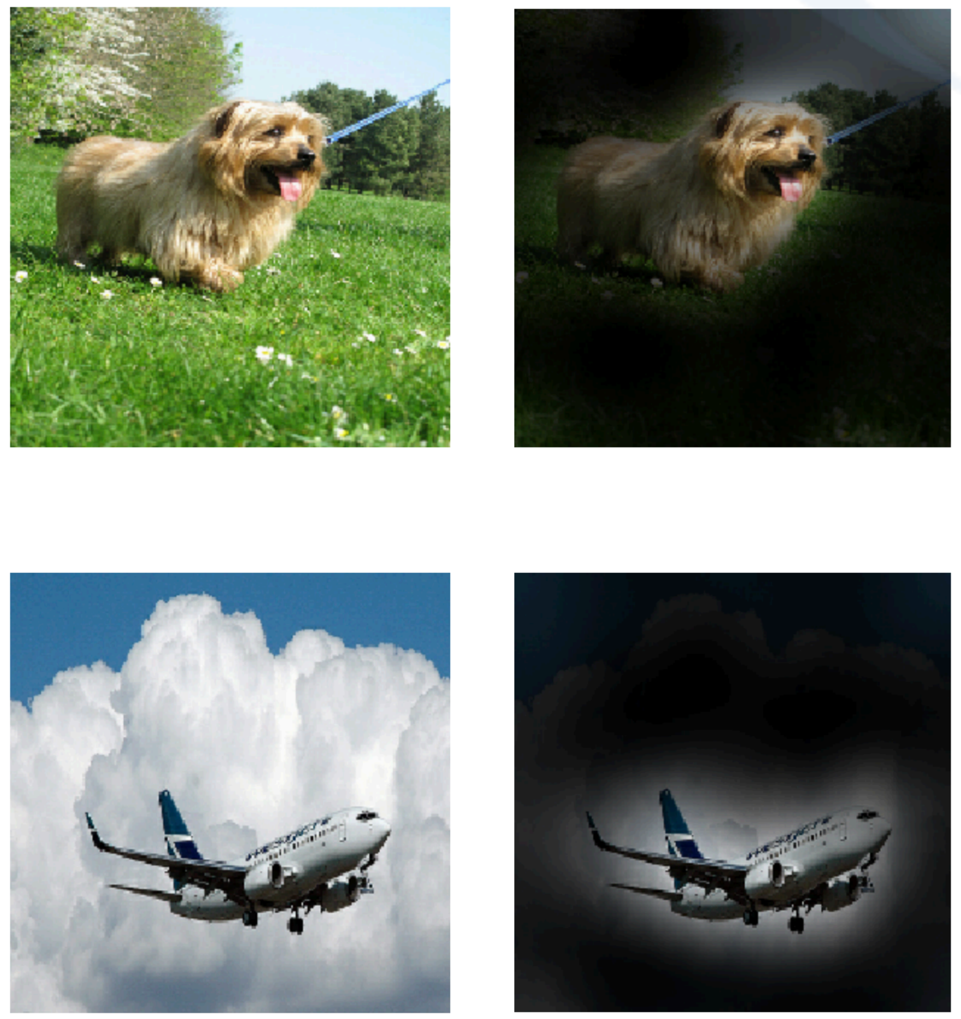
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

○ *Self Attention:*

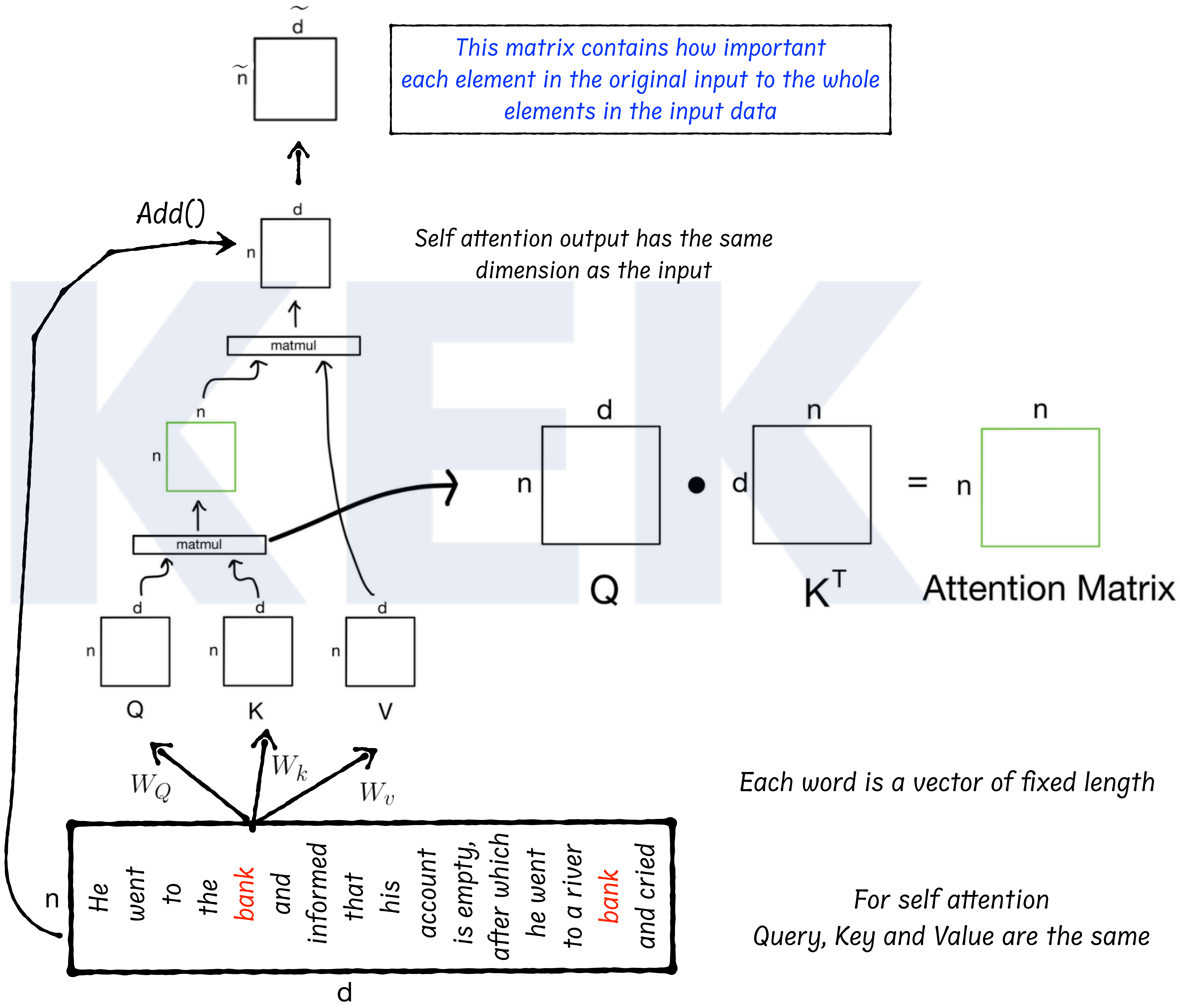
self-attention allows each element in the sequence to attend to all other elements, capturing both local and global dependencies. This is achieved through the calculation of attention scores, which are used to linearly combine the values associated with different positions.

Arxiv:2010.11929

Input Attention



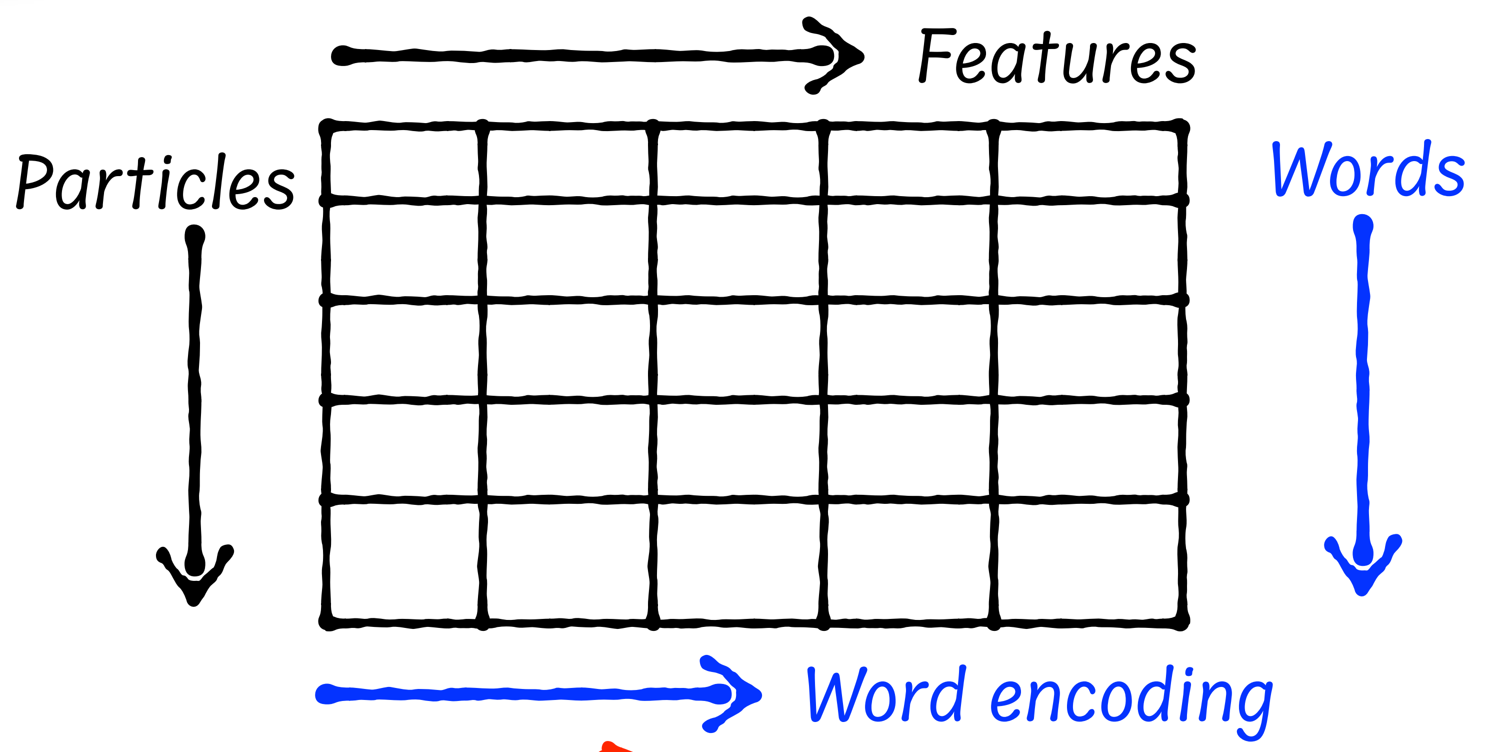
It assigning different weights to different elements in the input sequence, emphasizing the more relevant parts while discarding the less relevant ones



Transformer for particle physics

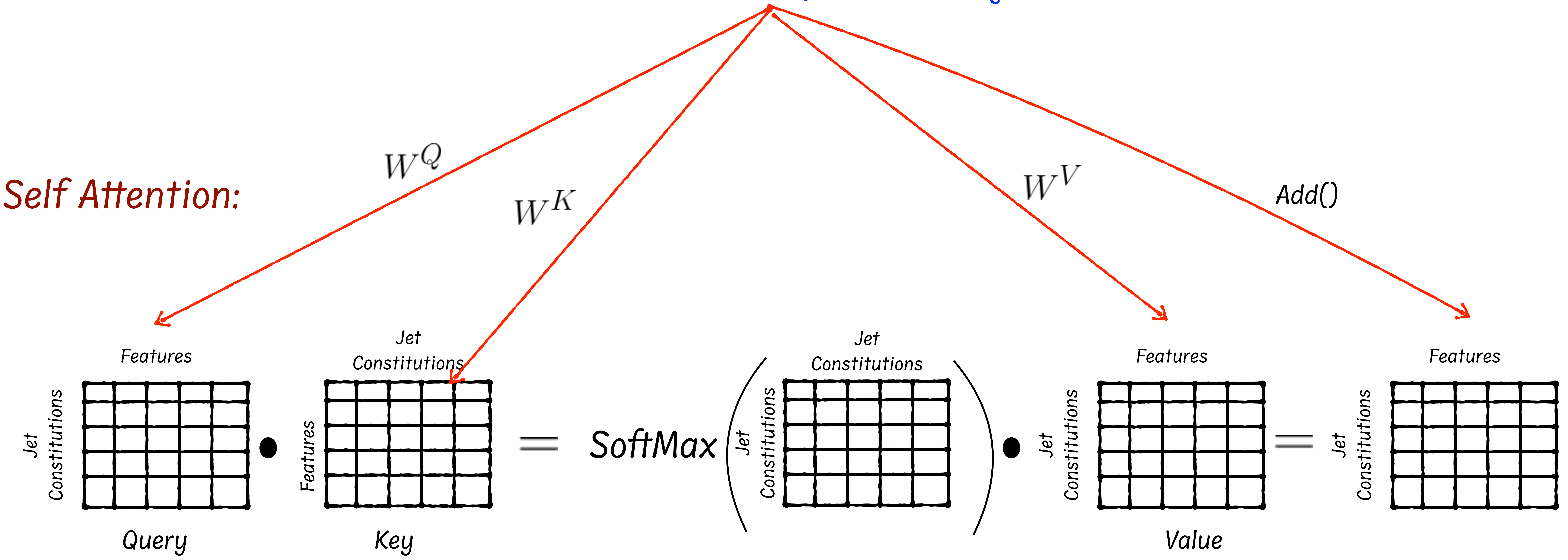
Well, so how it works in particle physics ?

Input data can structured
as a fixed size unordered grid



Particles can be presented
as words in sentence

- Self Attention:



Transformer for particle physics

Cross Attention:

1- Assign the weight matrices

$$Q^{i\times j} = X^{i\times j}\cdot W_Q^{j\times j}\,,\quad K^{n\times j} = S^{n\times m}\cdot W_K^{m\times j}\,,\quad V^{n\times j} = S^{n\times m}\cdot W_V^{m\times j}$$

For two input data sets $X^{i\times j}\,,\ S^{n\times m}$

2- Attention output

$$\mathcal{Z}^{i\times j} = \mathrm{softmax}\left(\frac{Q^{i\times j}\cdot (K^{n\times j})^T}{\sqrt{d}}\right)\cdot V^{n\times j}$$

Input and output have the same dimensions

3- Concatenate all heads

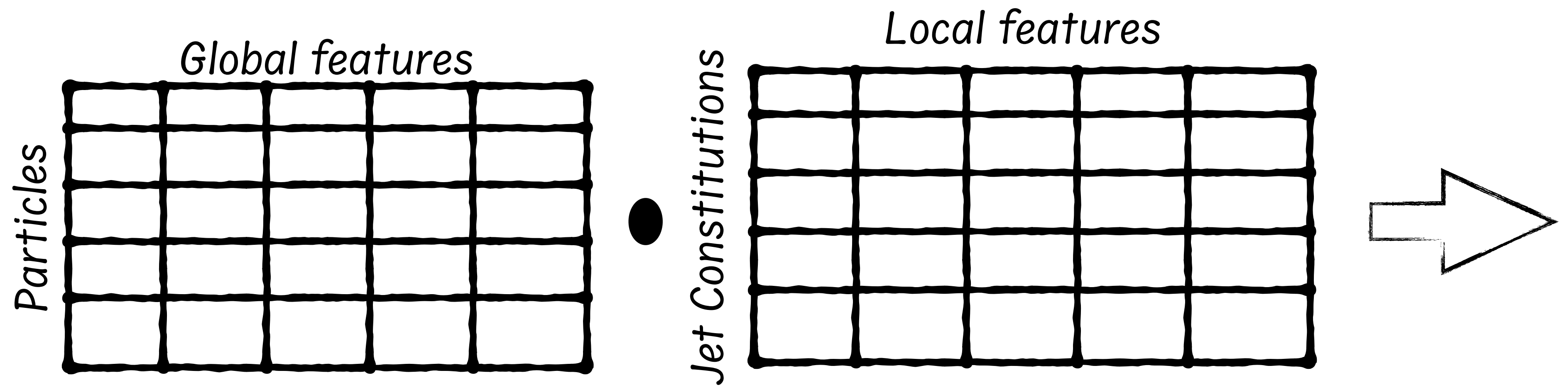
$$\mathcal{O}^{i\times j} = \mathrm{concat}\left(\mathcal{Z}_1^{i\times j}, \mathcal{Z}_2^{i\times j}\cdots \mathcal{Z}_n^{i\times j}\right) W^{(n*j\times j)}$$

Normalizing matrix to preserve the dimension

4- Skip connection

$$\tilde{X}^{i\times j} = X^{i\times j} + \mathcal{O}^{i\times j}$$

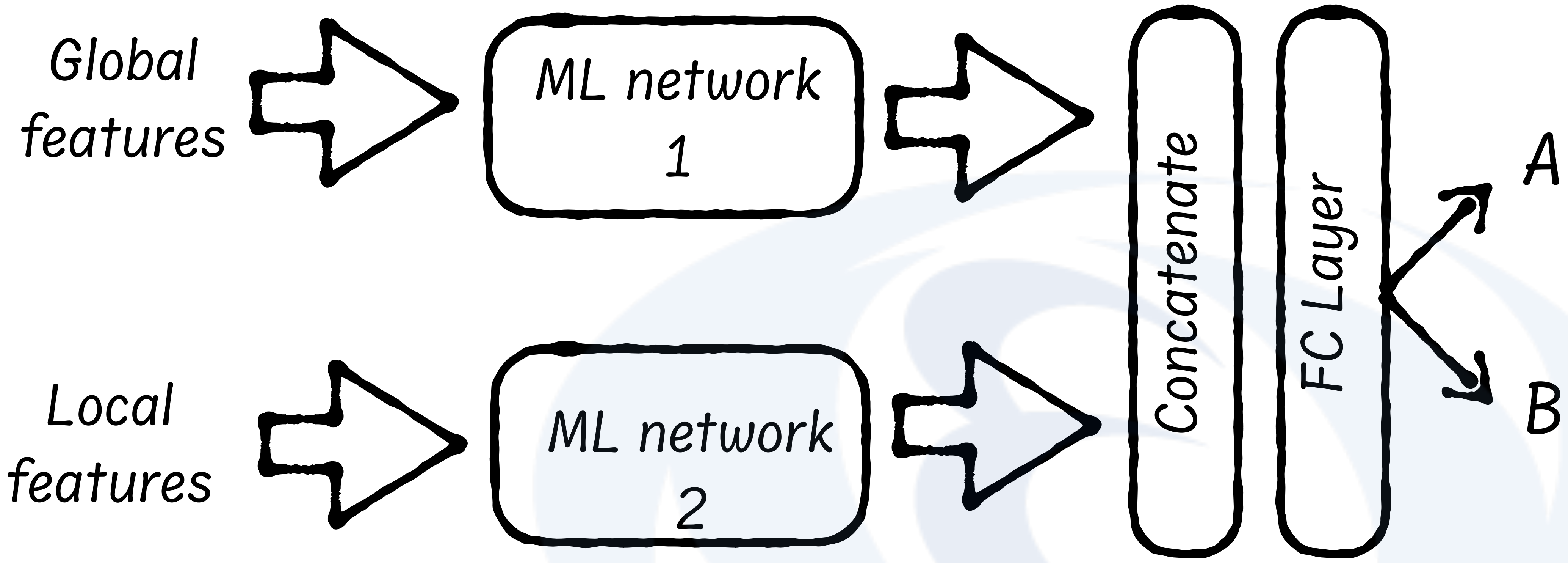
Output has the same dimension of the Input X



One output that encodes the important information extracted from the global and local information Of the event

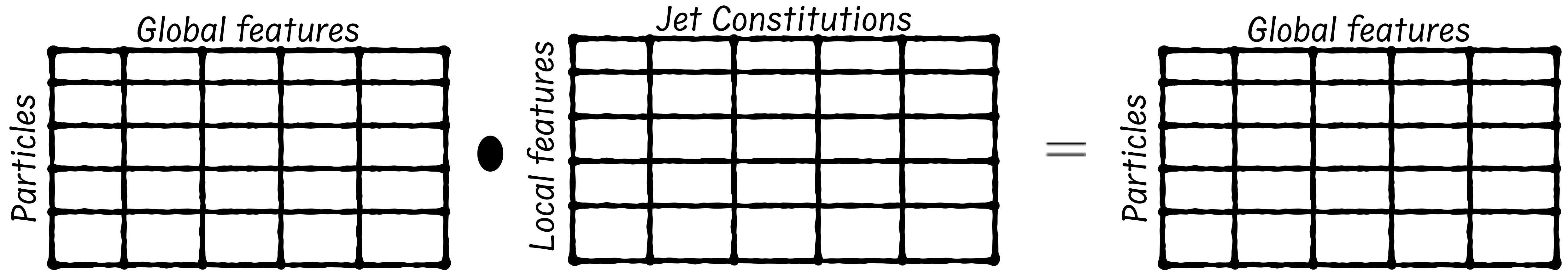
Transformer for particle physics

Simple concatenation



Input data set
with two different scales

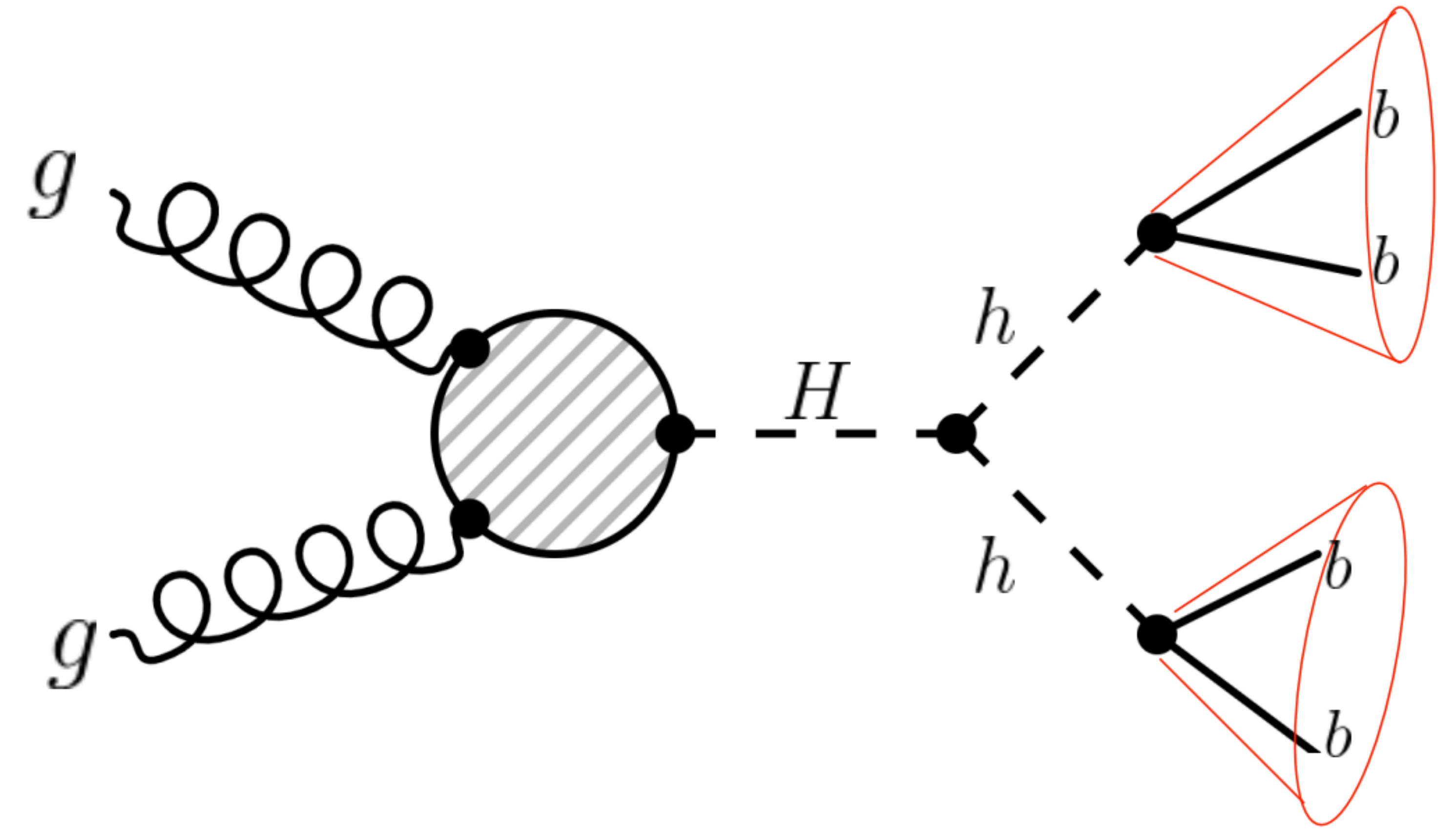
Cross Attention



Input data set
combines two data sets with
one scale

Physics Example

arXiv:2401.00452
JHEP 03 (2024) 144



Data Pre-processing:

- **Centering:**

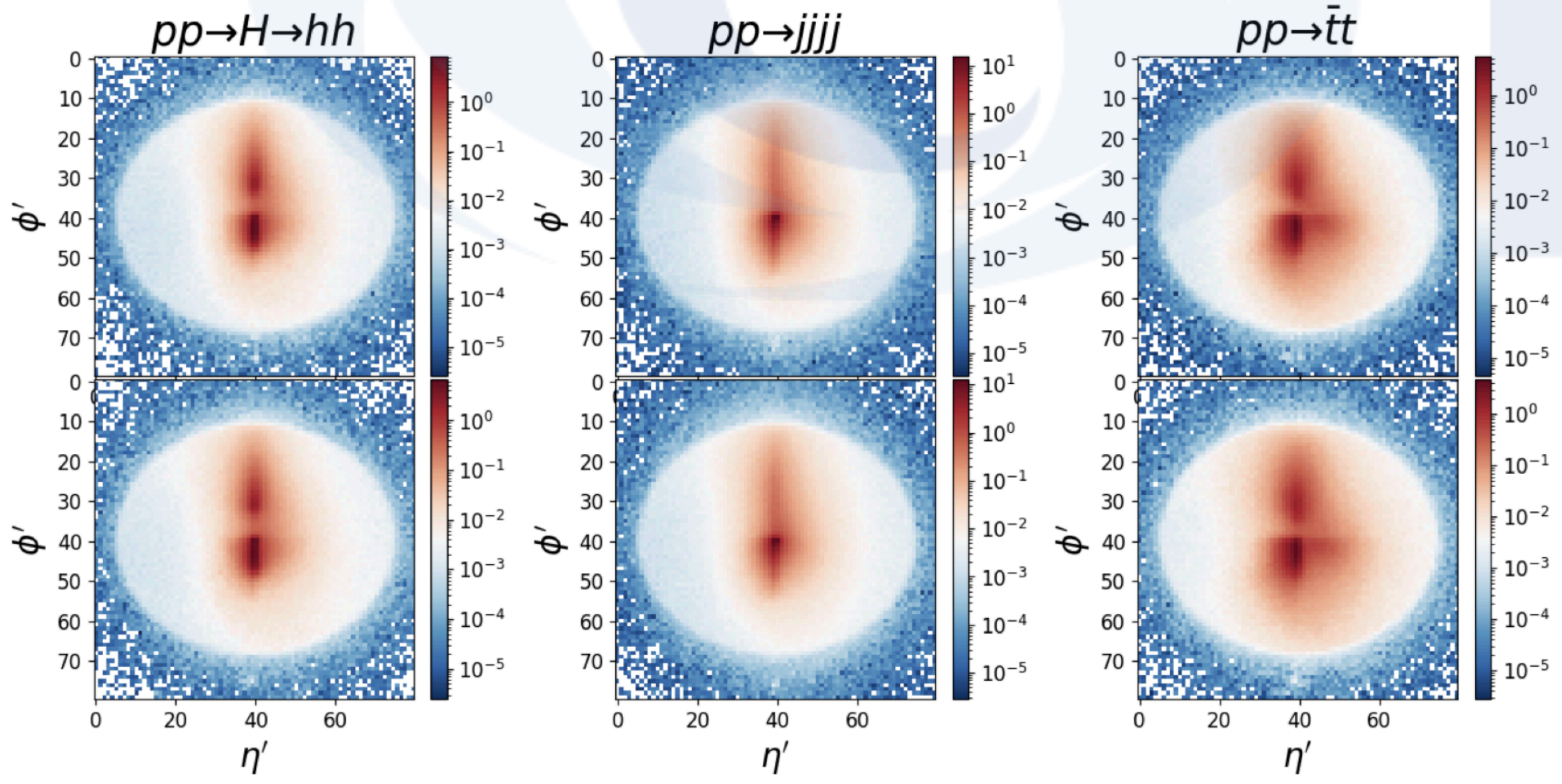
  Jet contents are shifted shifted such that the jet axis in the center

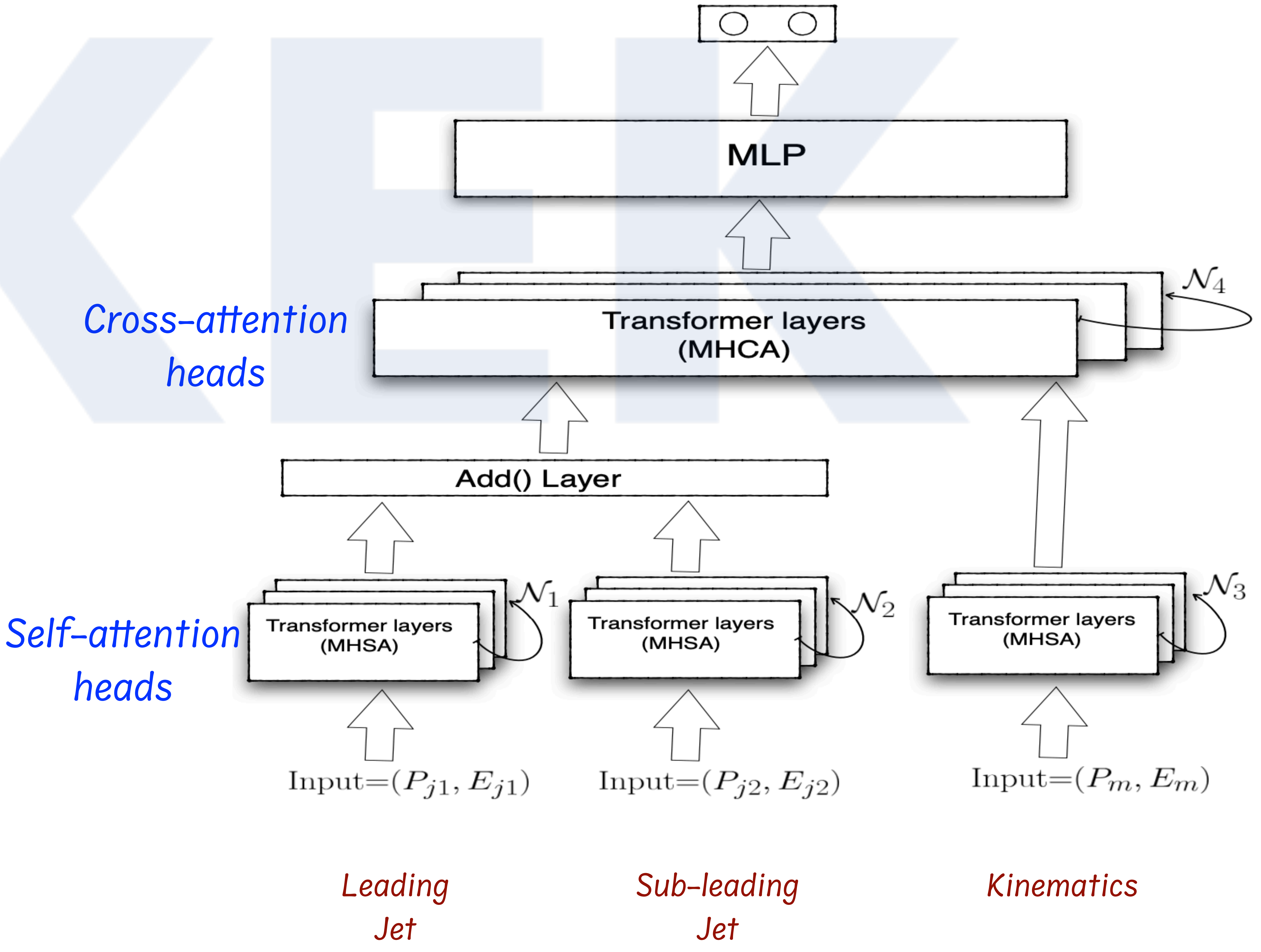
- **Rotation:**

  Rotate the jet contents in the eta-phi plane such that the jet axis is vertical

- **Flipping:**

  Jet contents are reflected over the vertical axis such that the right side contains the hard radiations



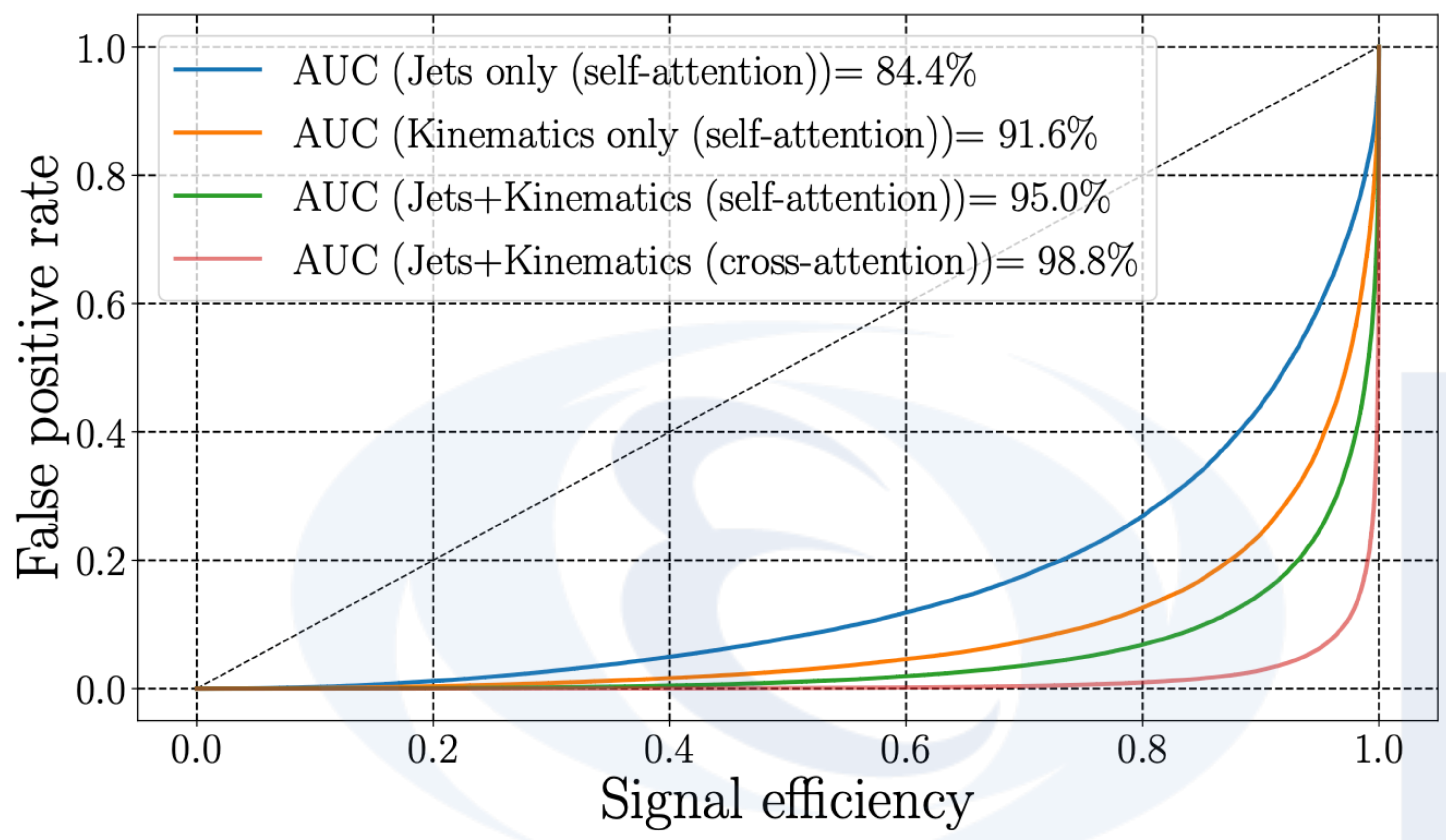


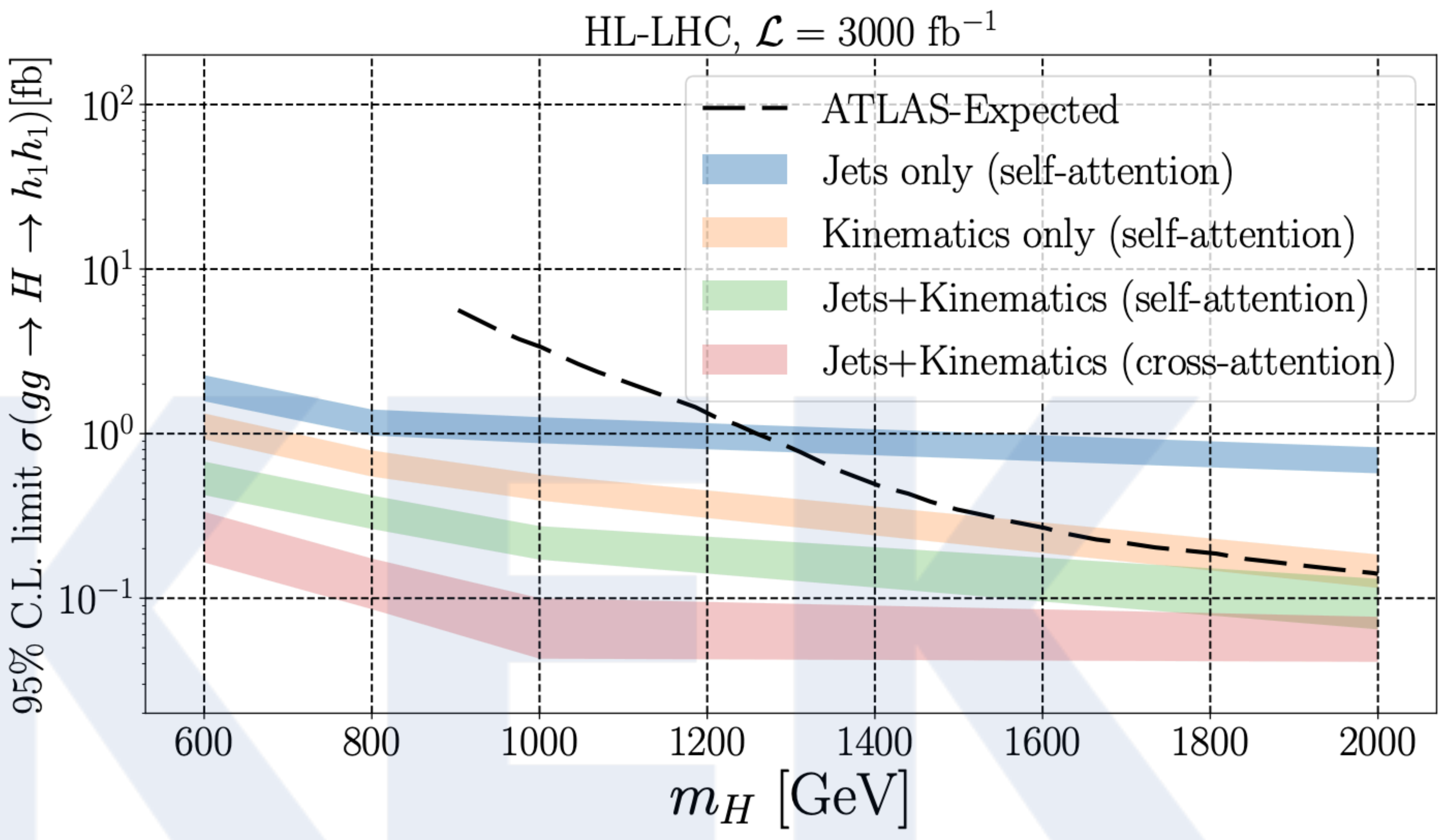
Background consists of 90% of QCD and 10% of ttbar

Results

ATLAS results are taken from ArxiV:2202.07288
and linearly scaled to 3000 1/fb integrated luminosity

ROCs for signal point with mH =1 TeV





Four models are considered:

- Transformer encoder with self attention trained on jets information only
- Transformer encoder with self attention trained on kinematics only
- Transformer encoder with self attention trained on jets information + kinematics
- Transformer encoder with cross attention trained on jets information + kinematics

For high mass range, the kinematics of the signal dominates with no much improvement of the machine learning over the basic cuts. For lower mass our network Is 10 times better than ATLAS analysis.

The bands due to repeating the experiment 5 times with different train and test splitting.

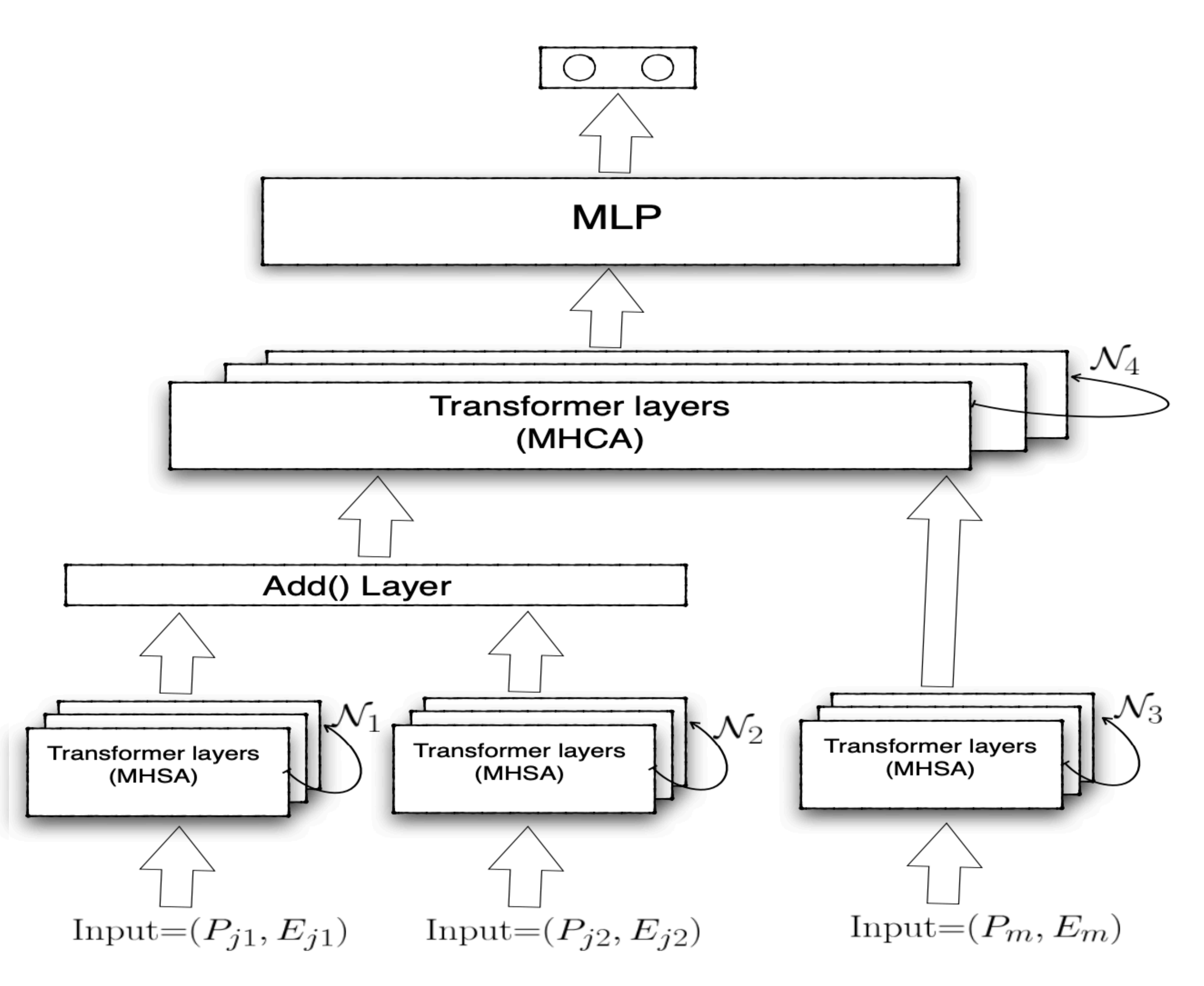
Well, good results! What is next ?

Interpretability

But why do we need interpretation methods?

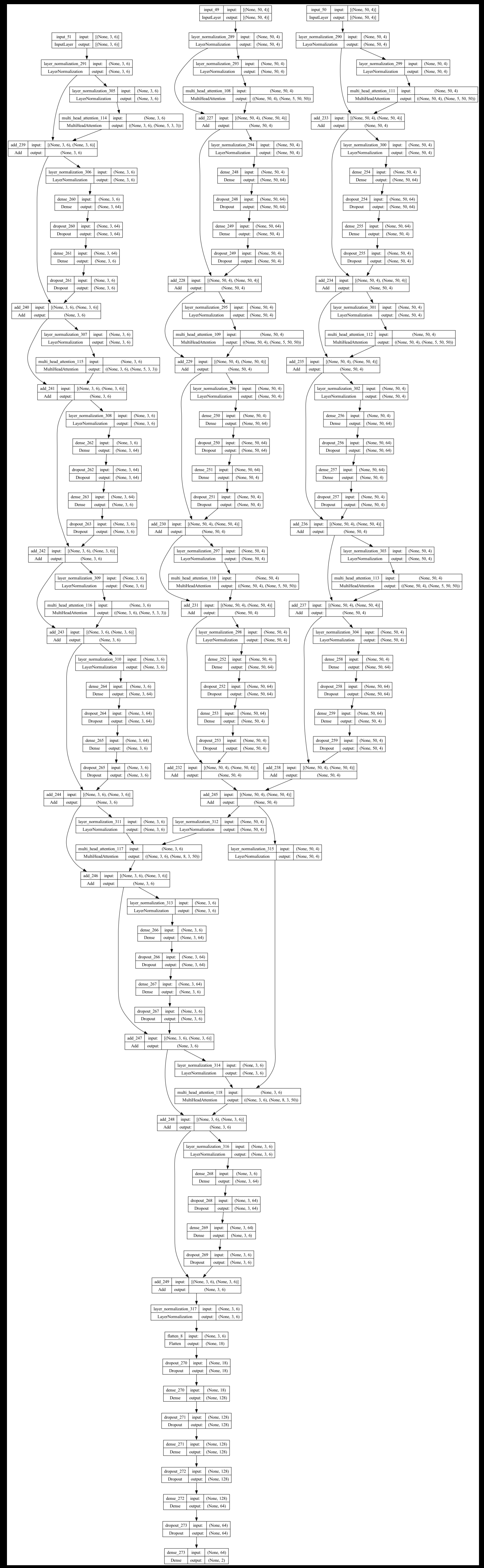
Interpretability

What we present



How the model actually looks like

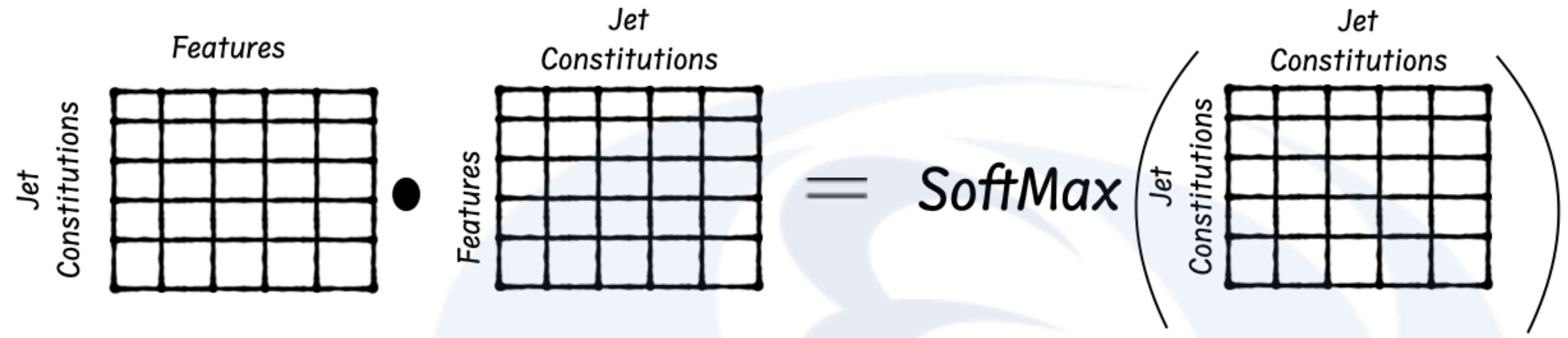
It is dangerous to deal with this complex structure as a black box



Attention Maps

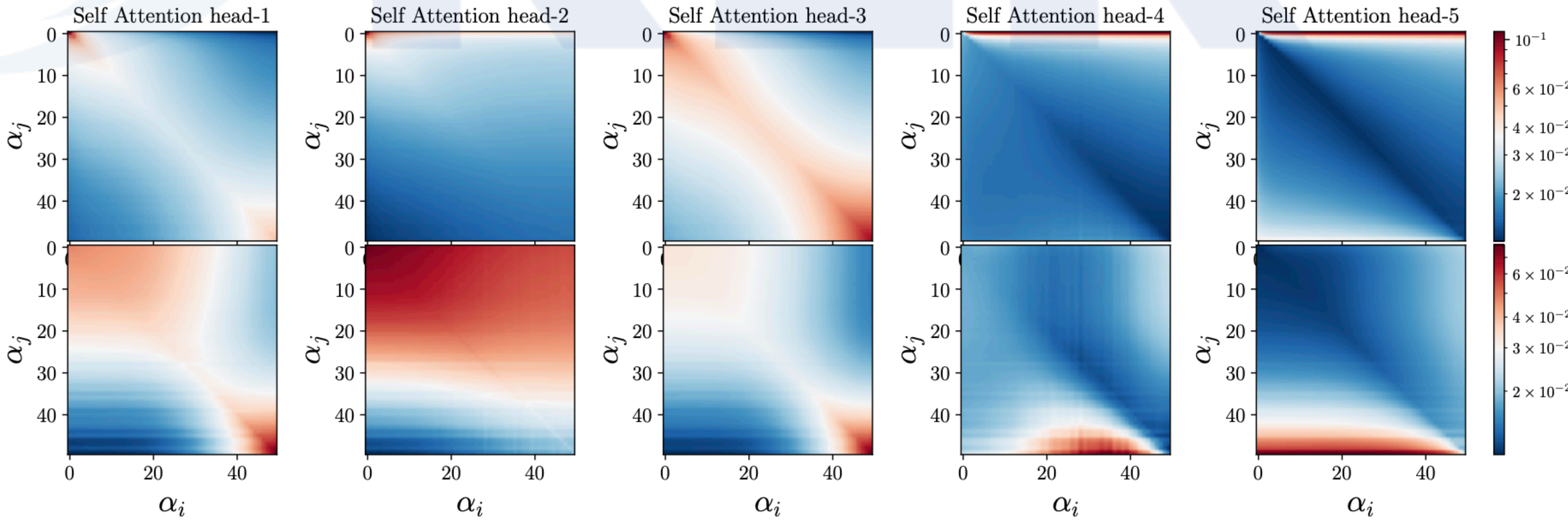
The analysis of the attention maps highlight the particle tokens that receive higher attention scores, indicating their significance in the model's decision. Also, it reveals how particle tokens relate to each other. It highlights the information extracted from the jet constitutes that are relevant to the reconstructed objects.

- *Self Attention:*



We use 5 self-attention heads for the first transformer encoder. Particles are sorted according to their pt, with zero pixel indicates the highest momentum particle.

Attention maps of test 120K samples.
Signal heads (top) background (bottom)

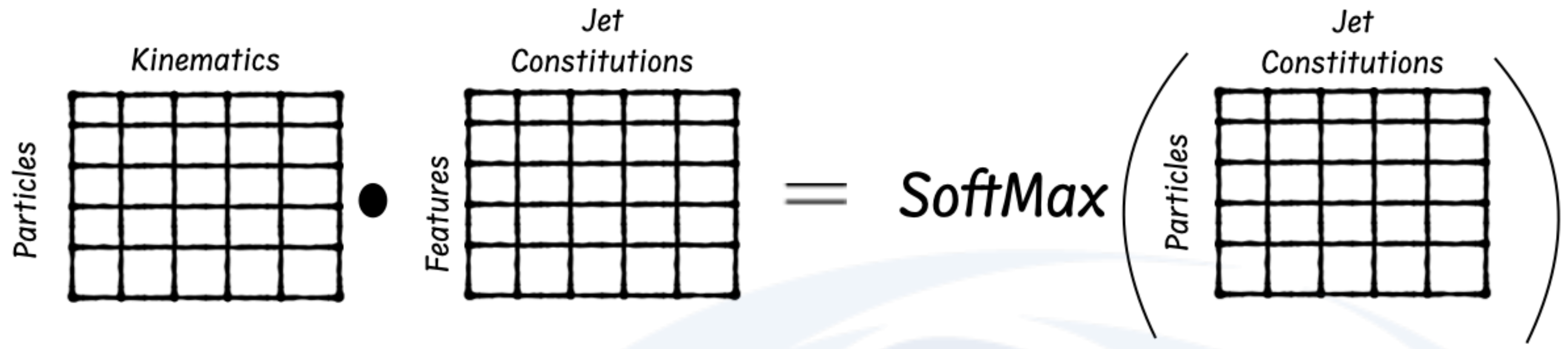


The attention map values reveal that the model concentrates on the leading and second-leading jet constituents to identify events as signal-like

On the other hand, the network assign high attention to wide momentum range of the jet constitutes when identify the input as background event.

Attention Maps

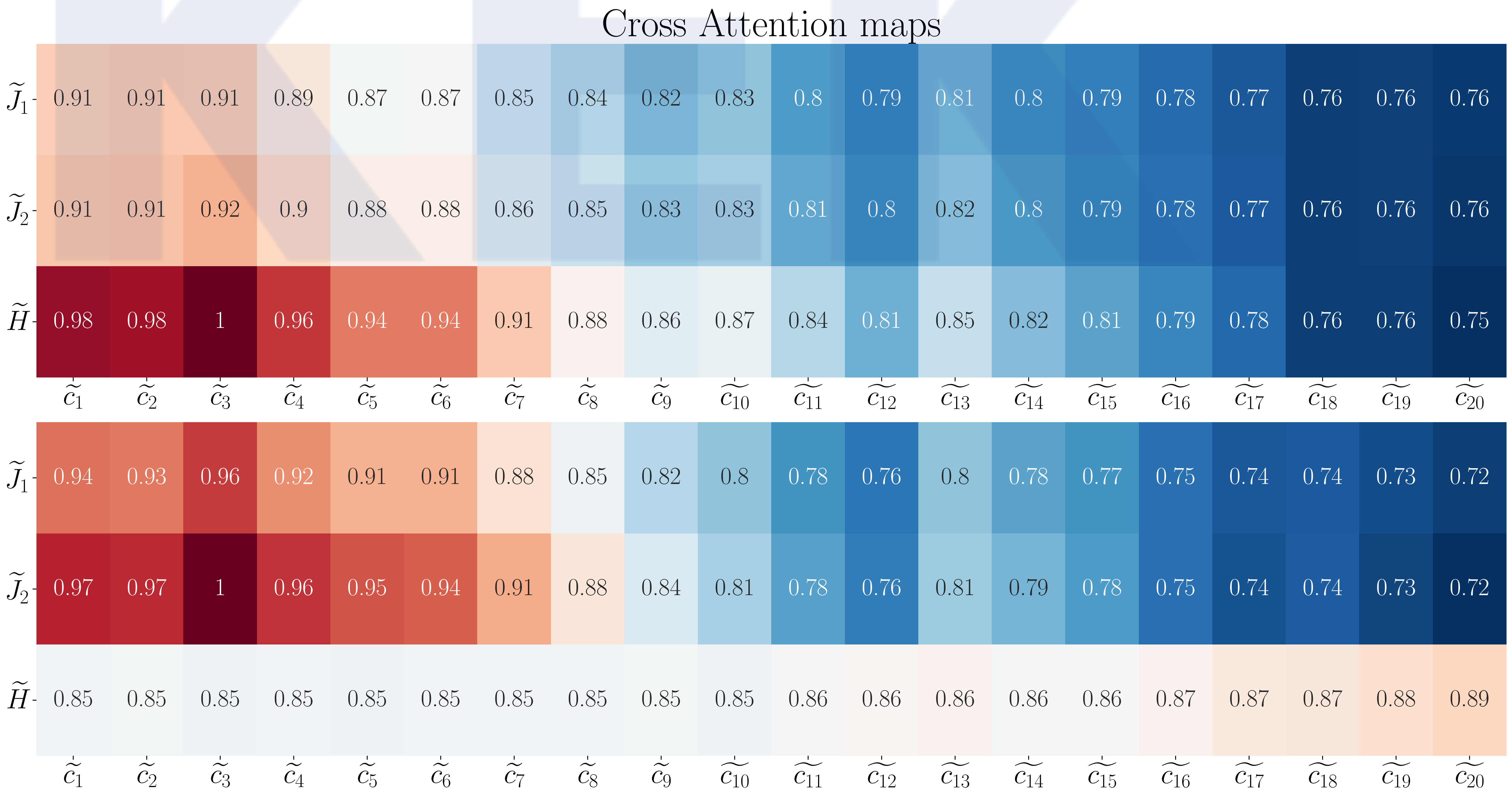
- **Cross Attention:**



Average over the 8 cross attention heads.
X-axis shows the leading 20 jet constituents.
Y-axis shows the reconstructed objects

Attention maps of test 120K samples.
Signal heads (top) background (bottom)

The leading constituents of the signal jets show High attention score to the reconstructed heavy Higgs. While background jets constituents exhibits flat attention to the heavy Higgs



Grad-Cam

Gradient weighted Class Activation Mapping (Grad-Cam) has been first Introduced in CNN model to visualize the most important pixels the model consider for his predictions.

Gad-Cam works as the following:

- *After training split the model from the last convolution layer.*
- *Compute the output of the last convolution layer (A)*
- *Compute the gradient of the class score of the second half of the model*
- *Compute the average of the gradients with resect to the spatial coordinates*

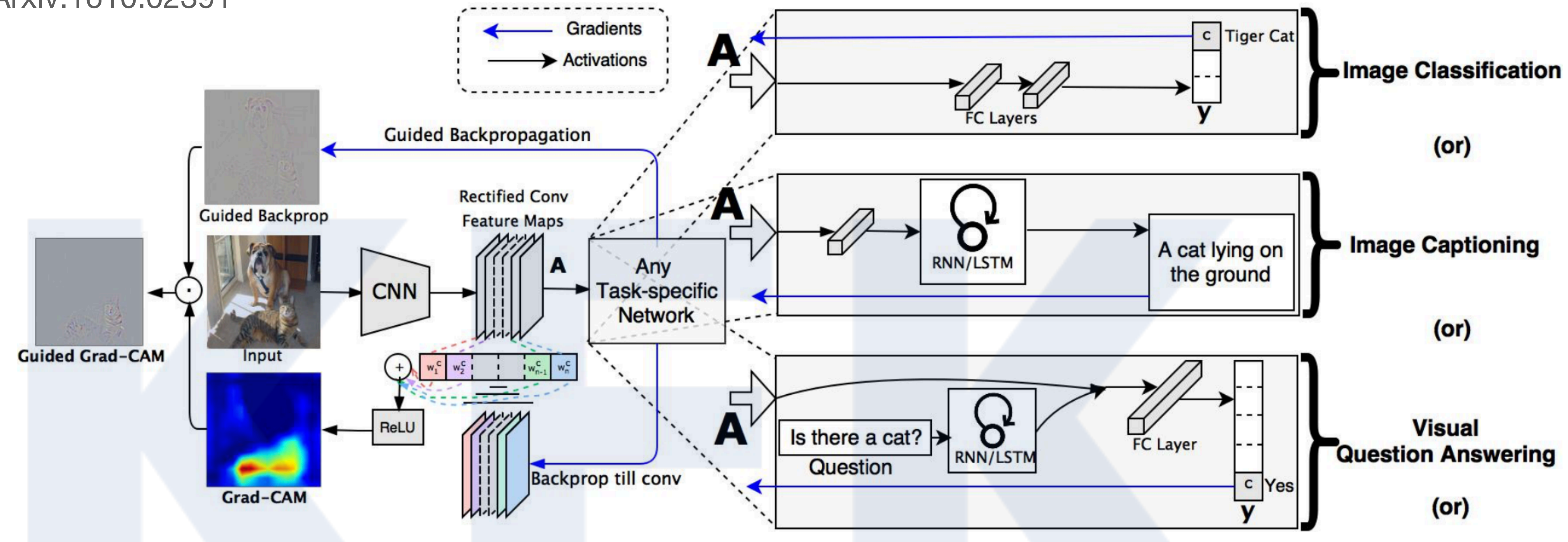
$$\alpha_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y}{\partial A_{ij}}$$

- *Compute the weighted sum of the feature maps output (A)*

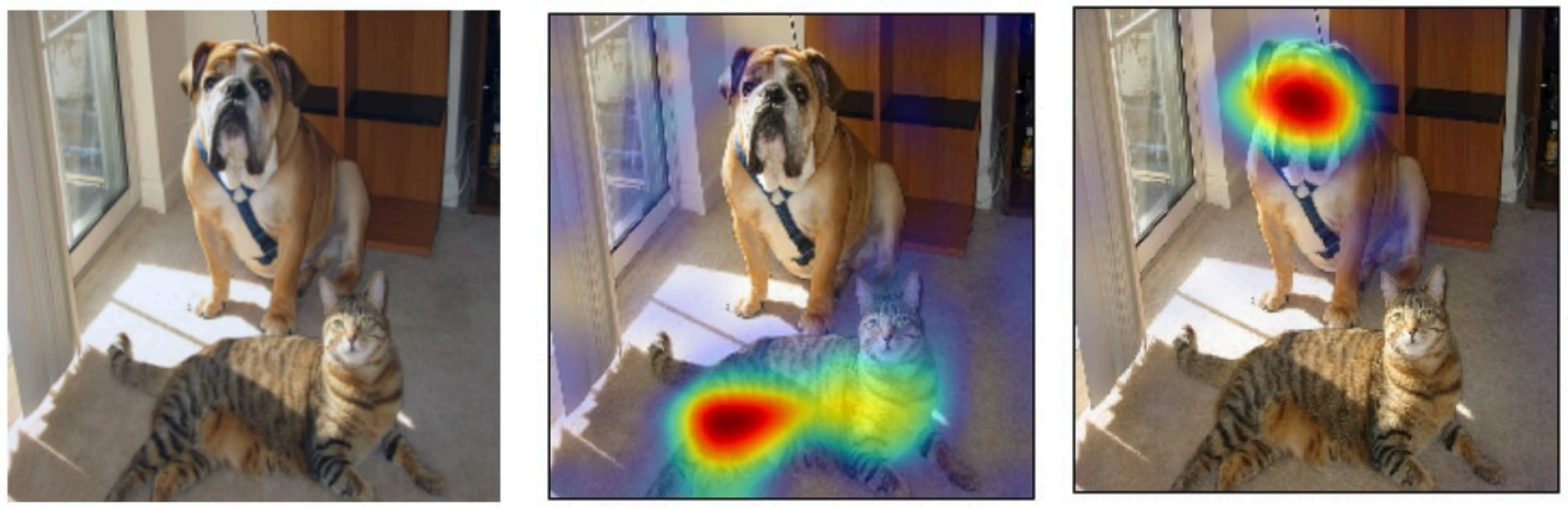
$$\text{Grad-Cam} = \text{ReLU}\left(\sum_k \alpha_k A^k\right)$$

- *The resulting heatmap indicates the spatial region in which the model focuses for predictions*

Arxiv:1610.02391



Arxiv:1610.02391

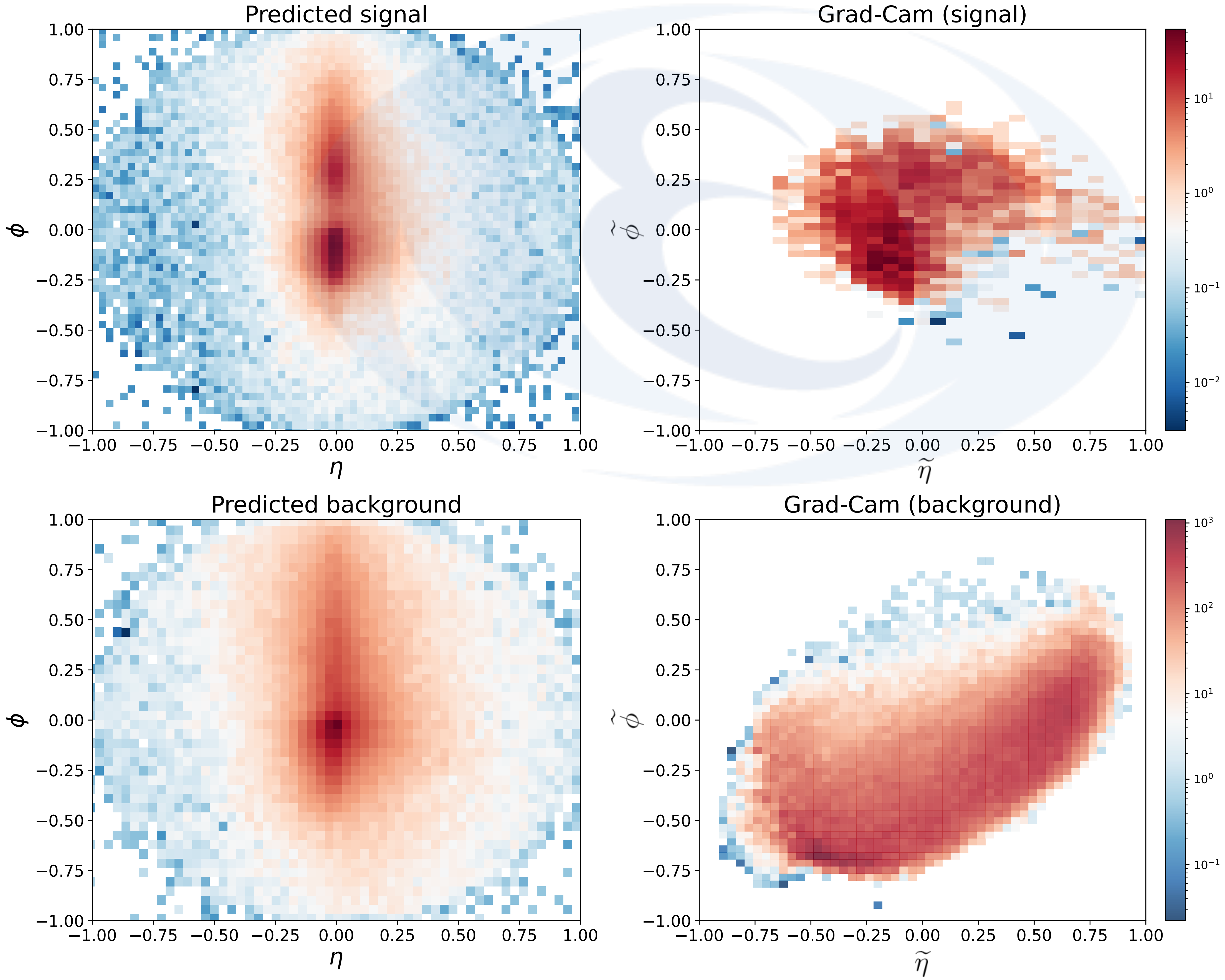


(a) Original Image (c) Grad-CAM 'Cat' (i) Grad-CAM 'Dog'

Grad-Cam

To visualize the region in the features space the model consider to Classify the inputs as signal or background like, we use Grad-Cam

Results for 5000 test images of the last self attention layer of the Jet transformer layer.



Grad-Cam shows that the network focuses on the two prong structure to predict the input as signal like, while it focuses on the radiation pattern of to predict the input as background like

The asymmetric pattern due to the flipping transformation In which all hard radiation are in the positive eta range

Our code

Our code is made for public with no hard coding

https://github.com/AHamamd150/Multi-Scale-Transformer-encoder

```
sig_dir = 'sig/'
bkg_dir = 'bkg/'
outdir = 'out/'
num_classes=2
batch_size= 500
epoch = 15
mlp_units = [128, 64]
masked = False
######Loss functions and optimizer########
loss_func = keras.losses.CategoricalCrossentropy()
optimizer = tf.keras.optimizers.legacy.Adam(learning_rate=0.005)
train_accuracy = tf.keras.metrics.CategoricalAccuracy()
test_accuracy = tf.keras.metrics.CategoricalAccuracy()
#######################################
## paremters of the first transformer#
#######################################
num_heads_1 = 5
num_transformers_1= 2
n_constit_1 = 40
n_channels_1 = 3
input_shape_part_1 = (n_constit_1,n_channels_1)
mlp_head_units_1 = [64,n_channels_1]
########################################
## paremters of the second transformer#
########################################
num_heads_2 = 5
num_transformers_2= 2
n_constit_2 = 9
n_channels_2 = 5
input_shape_part_2 = (n_constit_2,n_channels_2)
mlp_head_units_2 = [64,n_channels_2]
########################################
```

Only one txt file to control the network structure

Run the code via the terminal command:

```
python3 run.py input.py
```

Thank you

for your listening