streamlined jet tagging network assisted by jet prong structure
role of cross attention

Mihoko Nojiri(IPNS, KEK), with Ahmed Hammad
and Stefano Moretti
arXiv 2401.00452 JHEP 03(2024) 144

Mihoko Nojiri with Ahmed Hammad
arXiv 2404 14677 JHEP 06 (2024) 176

- Recent improvement of jet classfication is achieved direct use of low level variable with flexible network(Transformer, GNN)

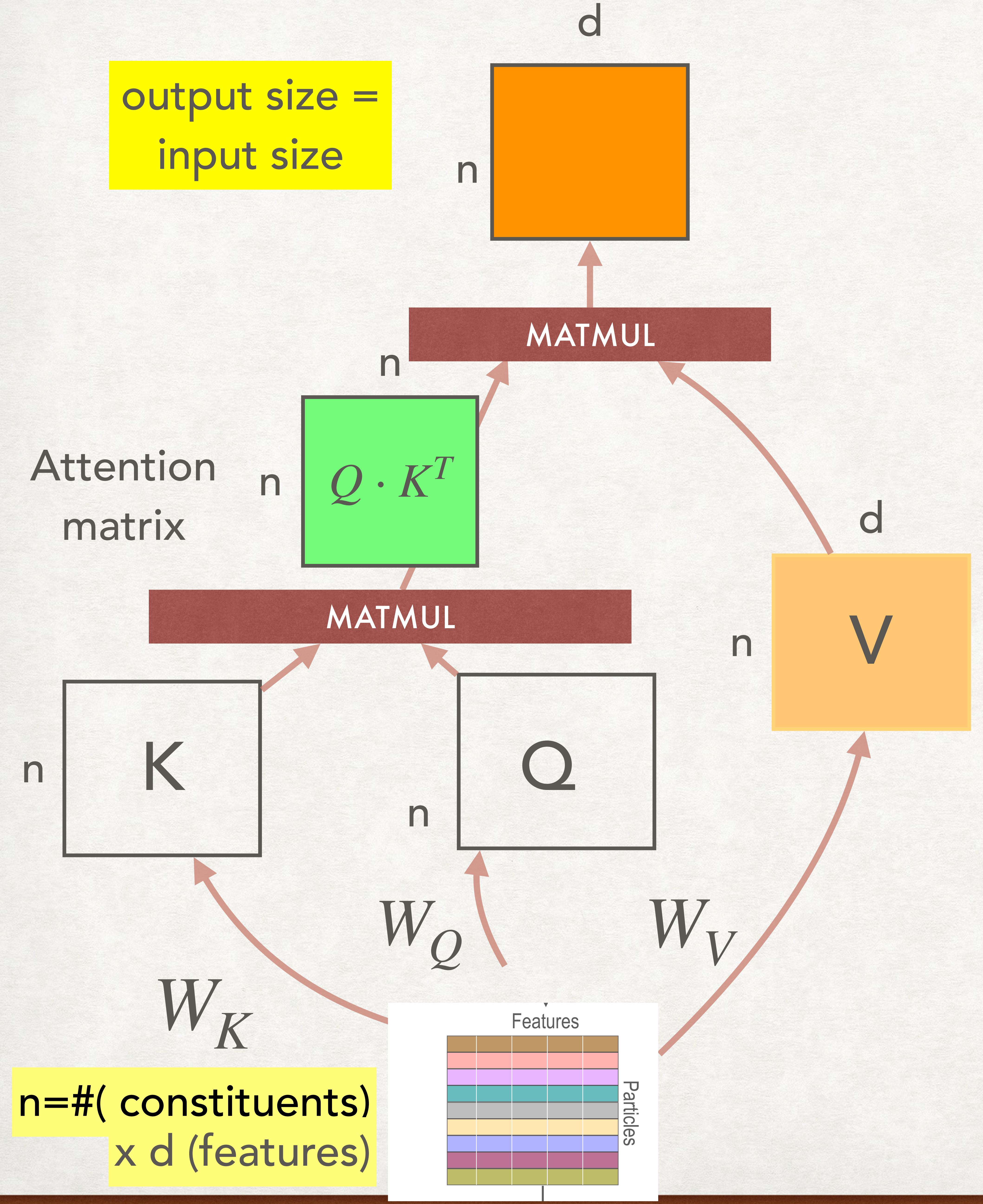
- Con: they suffer low interpretablity.

- Todays talk: →Building the network that respect energy scale of LHC process” using cross attention

1. “streamline jet classifiers”   subjets x jet constituents

2 toward global event analysis fatjet x jet constituents

"TRANSFORMER" :SELF ATTENTION LAYERS

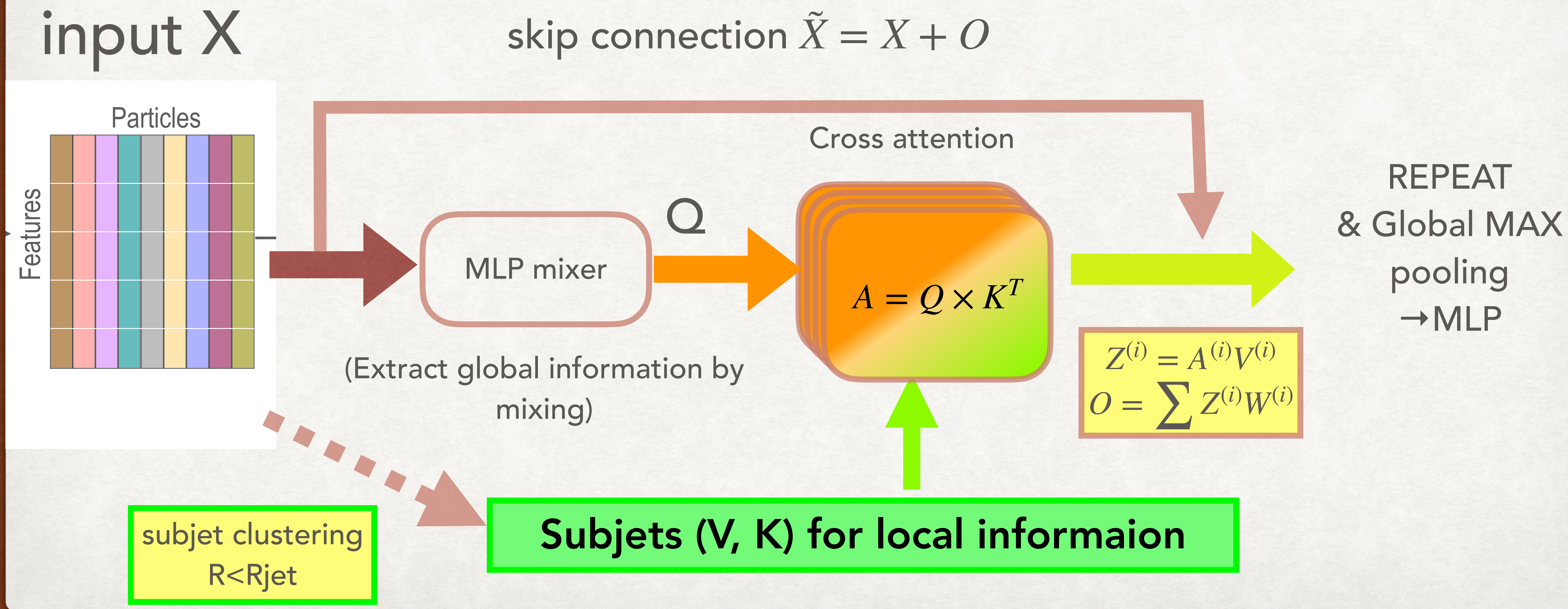


$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Data is matrix of n(#constituent) x d(feature) → $K(n \times *), Q(n \times *), V(n \times d)$
- Attention Matrix evaluate the correlation of constituents taking into account all features. Higher attention elements indicates important correlations
- tStructure of data retained for the next transformation.

1. Cross attention to focus on the P(h| (sub)jet)

ATTENTION →CROSS Attention for P(h| subjets) estimation

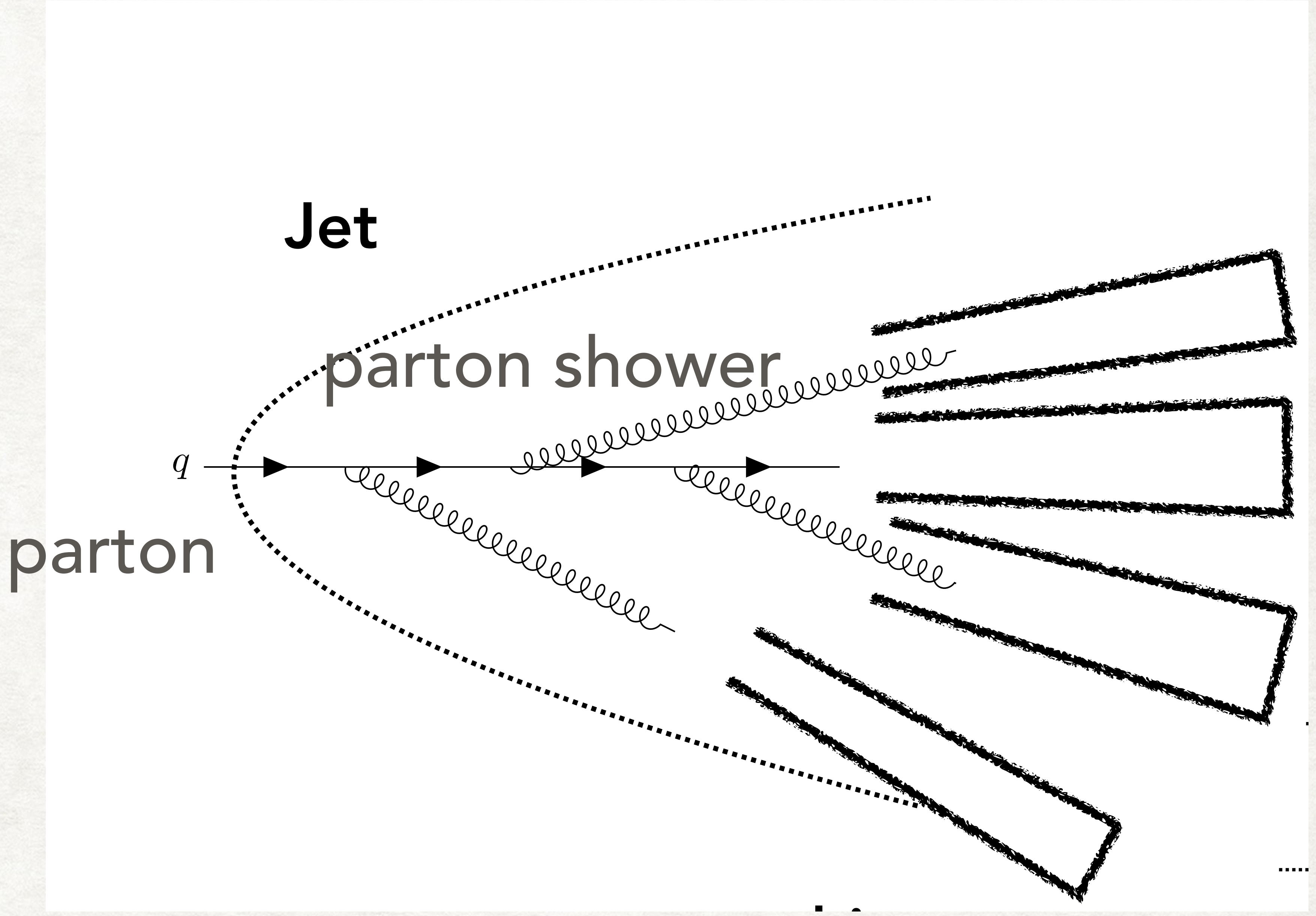


WHY CROSS ATTENTION? (I)

Reason 1 "Physics SCALE"

- Hard Process = Partons y
- Parton shower → hadronization
  - a jet: P(hadrons in jets | parton) = $P(\{x_i\}\,|\,y)$

    jet with substructure $P(\{x_i\}\,|\,\{y_\alpha\})$
  - **Extension: several fatjets in an event →2**



$$P(\{x_i\},\{x'_j\},\{y_\alpha\},\{y'_\beta\}) \sim P(\{x_i\}\,|\,\{y_\alpha\})P(\{x'_i\}\,|\,\{y'_\beta\})\; P(\{y_\alpha\, , y'_\beta\})$$

We need correlation between parton =(sub)jet and particles

WHY CROSS ATTENTION? (II)

LHC process

Hard scattering

Jet function: Parton shower

$$\sigma(pp \to a, b \to N\text{jets}) \sim H_N \left[B_a B_b \prod_{k=1}^{N} J_k \right] \otimes S_N \,,$$

Soft staff!

Cross attention

$$Q(\Phi_{1\theta_1}(\text{constituents})) \cdot K(\Phi_{2\theta_2}(\text{subejts}))$$

Self attention

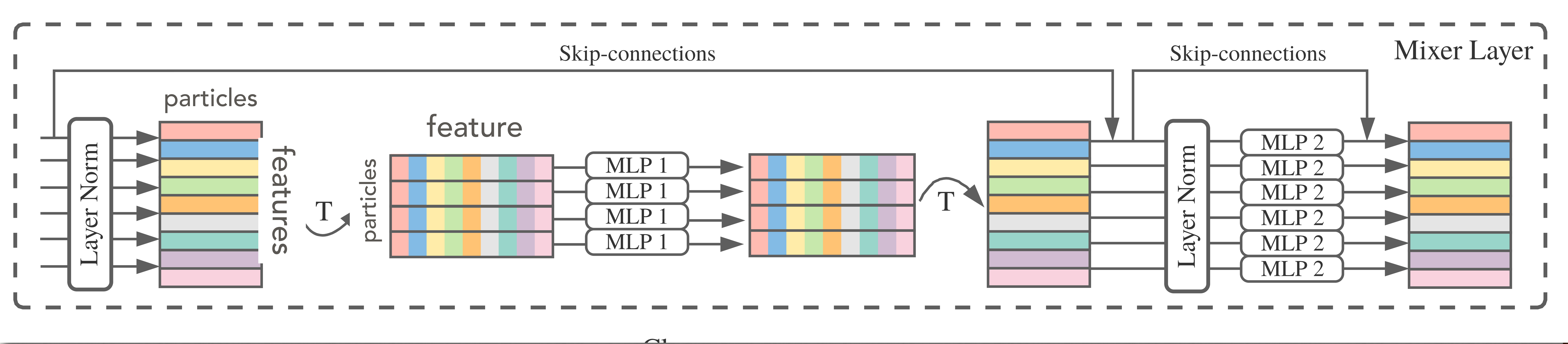
Large gradient

small gradient

$$A\,V = \begin{pmatrix} Q(\text{constituent}) \times K(\text{constituent}) & Q(\text{constituent}) \times K(\text{subjet}) \\ Q(\text{subj}) \times K(\text{constituent}) & Q(\text{subj})\,K(\text{subj}) \end{pmatrix} V = Q(\text{subj})\,K(\text{subj})\,V(\text{subj}) + \text{others}$$

CAPTURE GLOBAL STRUCTURE BY MLP MIXER

The “mixer layer” has only two MLP that mix both features and particle tokens: focus on global feature.



MLP 1 :mix feature only acts for all particles

MLP 2: mix particles acts for all features

any information can be included & apply repeatedly → Transformer like

“subjet information” take care cluster information

Performace comparable to Particle Transformer but much faster and lighter

Models	AUC	R50%	#Parameter	Time (GPU%)
ParT	0.9858	413+-16	2.14M	612
Mixer+subjet (CA)	0.9856	392+-5	86.03K	33
(AK)	0.9854	375+-5	86.03K	33
(HDBSCAN)	0.9859	416+-5	86.03K	33
LorentzNet	**0.9868**	**498+-18**	**224K**	
PELICAN (Lorents Invariance)	0.9869	-	45K	-

*Subjet cone size R=0.3

*HDBSCAN is algorithm without distance measure

Performace comparable to Particle Transformer but much faster and lighter

Models	AUC	R50%	#Parameter	Time (GPU%)
ParT	0.9858	413+-16	2.14M	612
Mixer+subjet (CA)	0.9856	392+-5	86.03K	33
(AK)	0.9854	375+-5	86.03K	33
(HDBSCAN)	0.9859	416+-5	86.03K	33
LorentzNet	**0.9868**	**498+-18**	**224K**	
PELICAN (Lorents Invariance)	0.9869	-	45K	-

*Subjet cone size R=0.3

*HDBSCAN is algorithm without distance measure

SMALL SIZE

FAST

HIGH PERFORMANCE WITHOUT USING LORENTS INVARIANCE

2. GLOBAL EVENT ANALYSIS

A HAMMAD S. MORETTI MN *JHEP* 03 (2024) 144

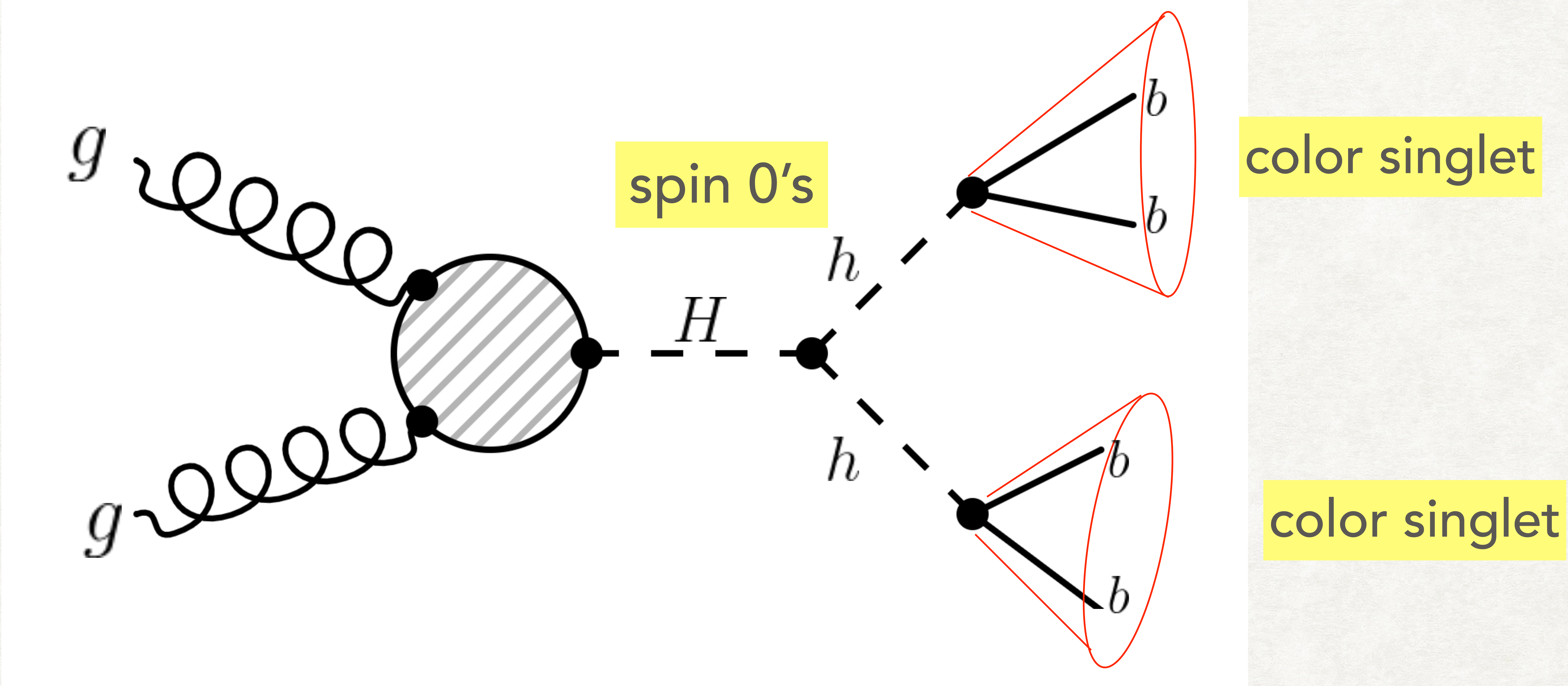
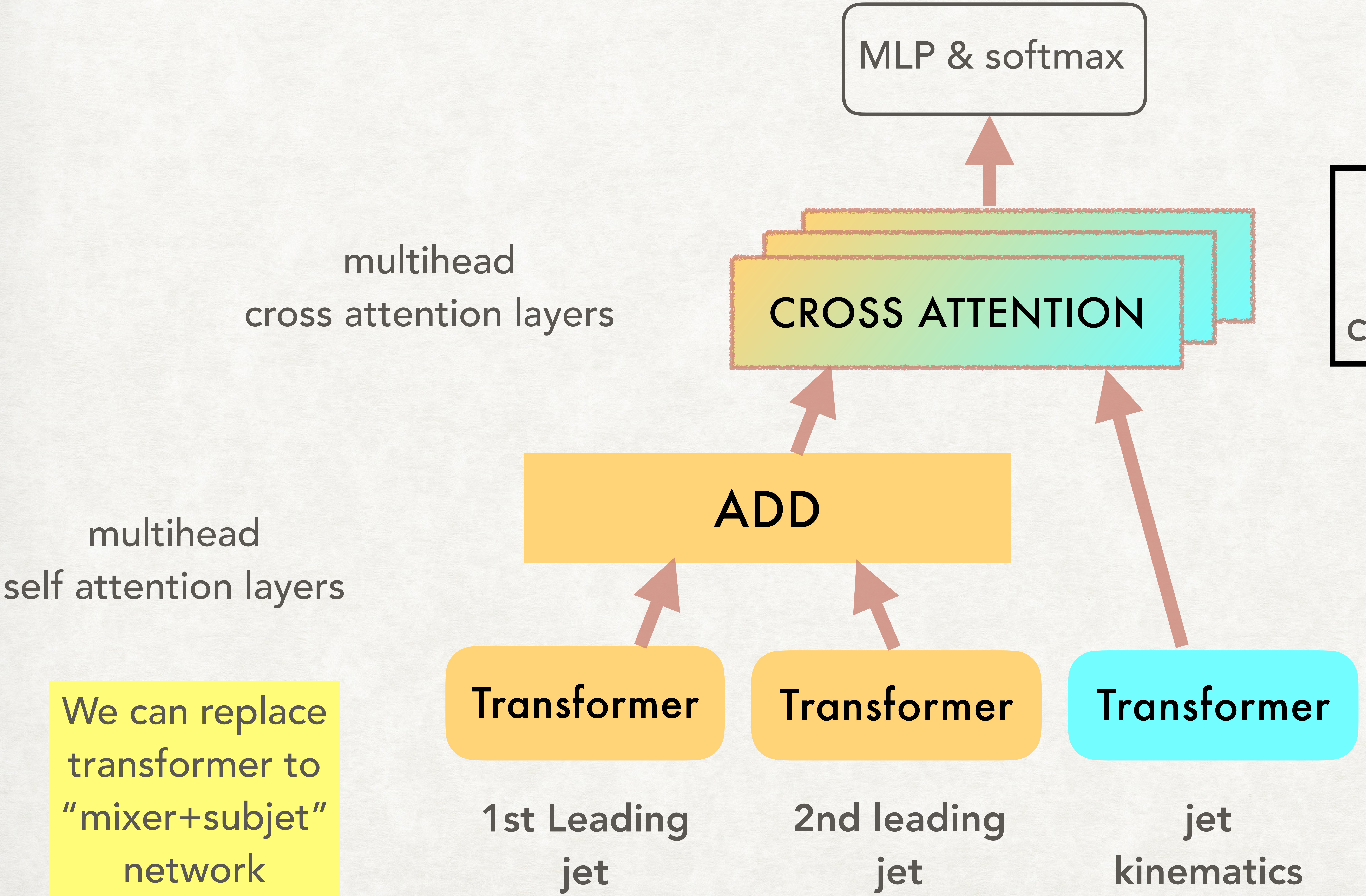


Figure 2: Feynman diagram for the signal process.

cross attention for 2 fatjet event



step 2 :Cross attention
transform jet kin by
cross Att. [substracture]x [jet kin]

step 1 : Self attention
[substructure]x[substructure]
[jet kin]x [jet kin]

We can replace
transformer to
“mixer+subjet”
network

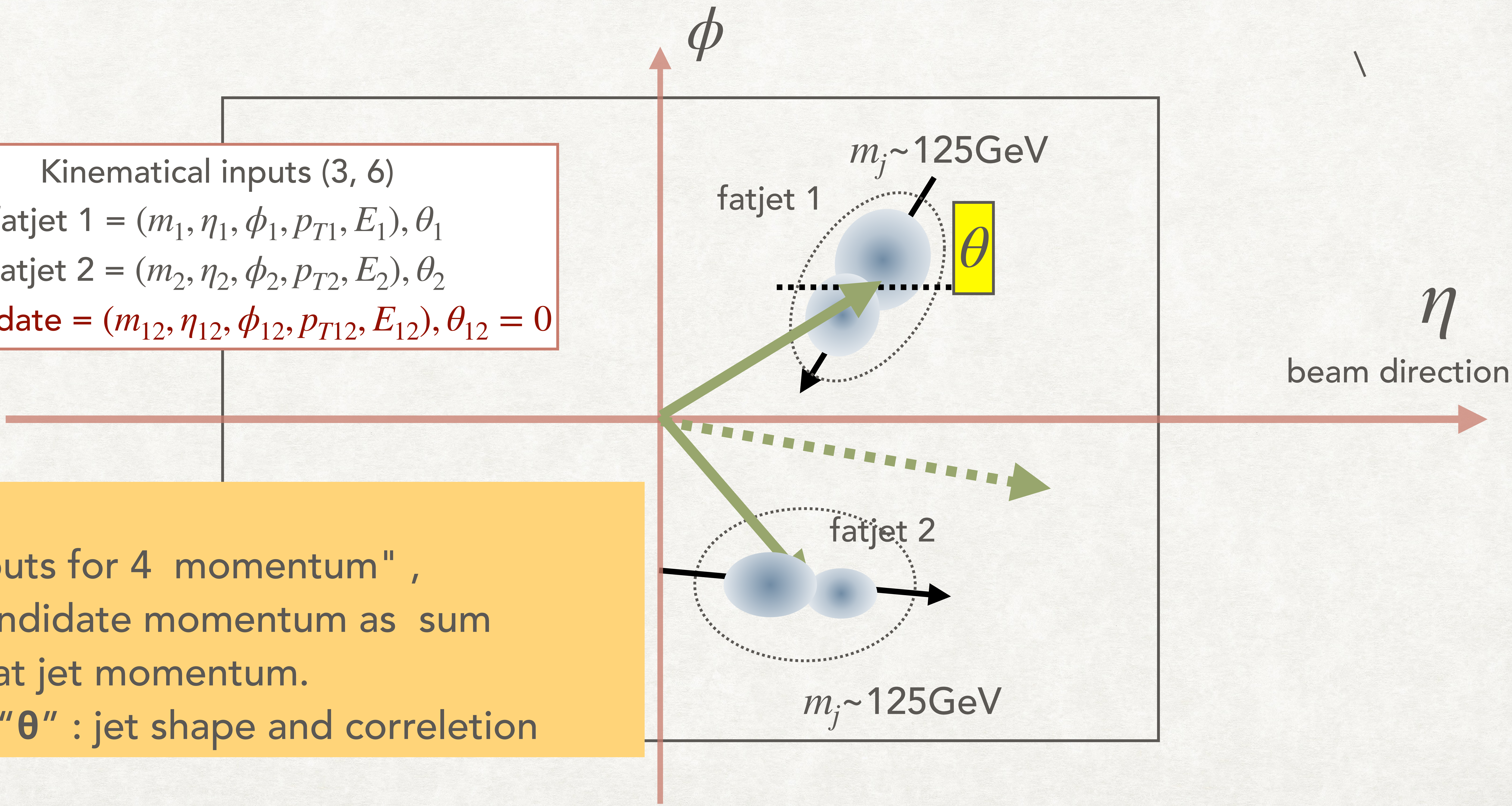
JET INFORMATION

Kinematical inputs (3, 6)

fatjet 1 = $(m_1, \eta_1, \phi_1, p_{T1}, E_1), \theta_1$

fatjet 2 = $(m_2, \eta_2, \phi_2, p_{T2}, E_2), \theta_2$

H candidate = $(m_{12}, \eta_{12}, \phi_{12}, p_{T12}, E_{12}), \theta_{12} = 0$



NOTE :

1.”5 inputs for 4 momentum" ,

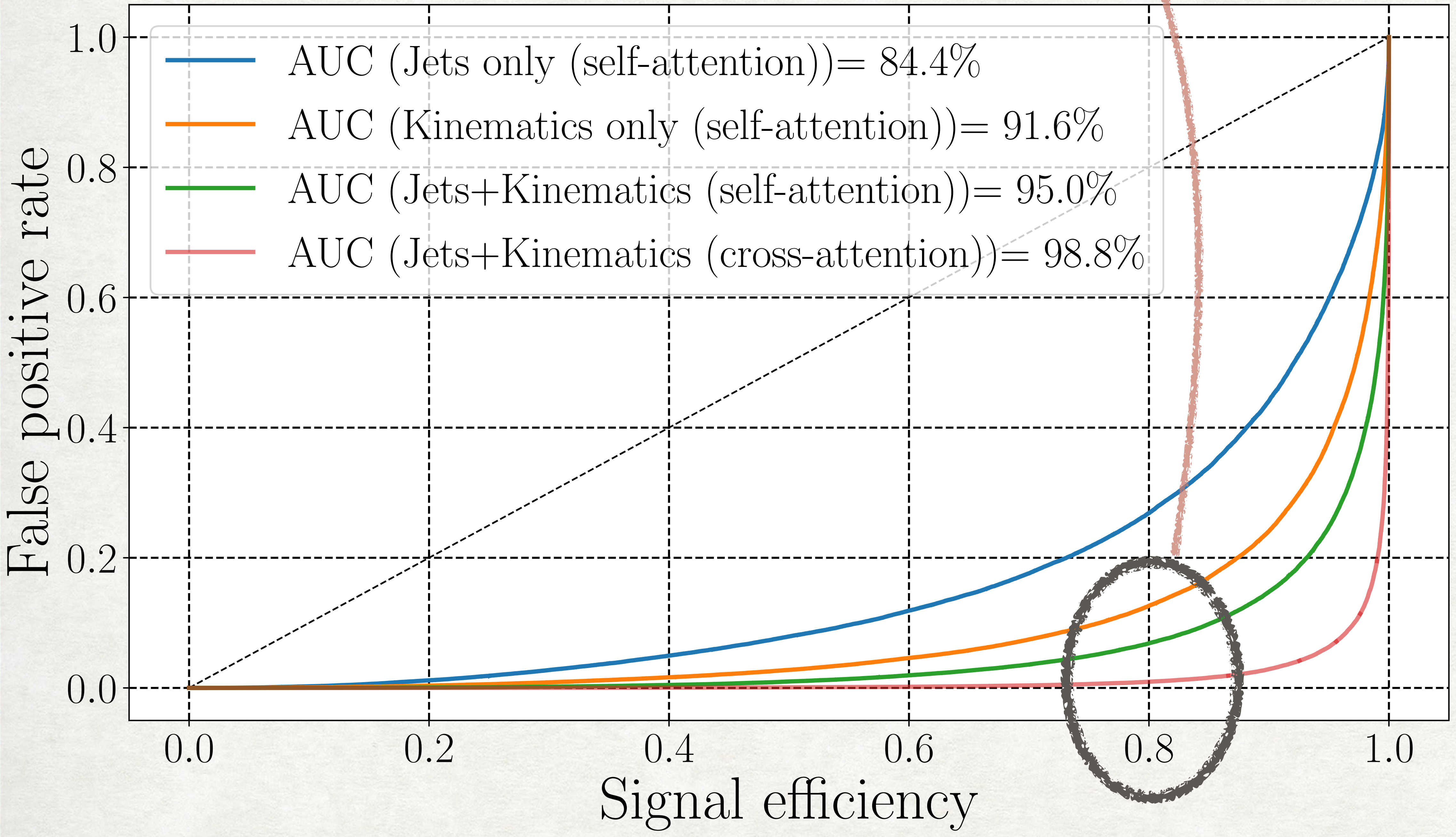
2. H candidate momentum as sum of the fat jet momentum.

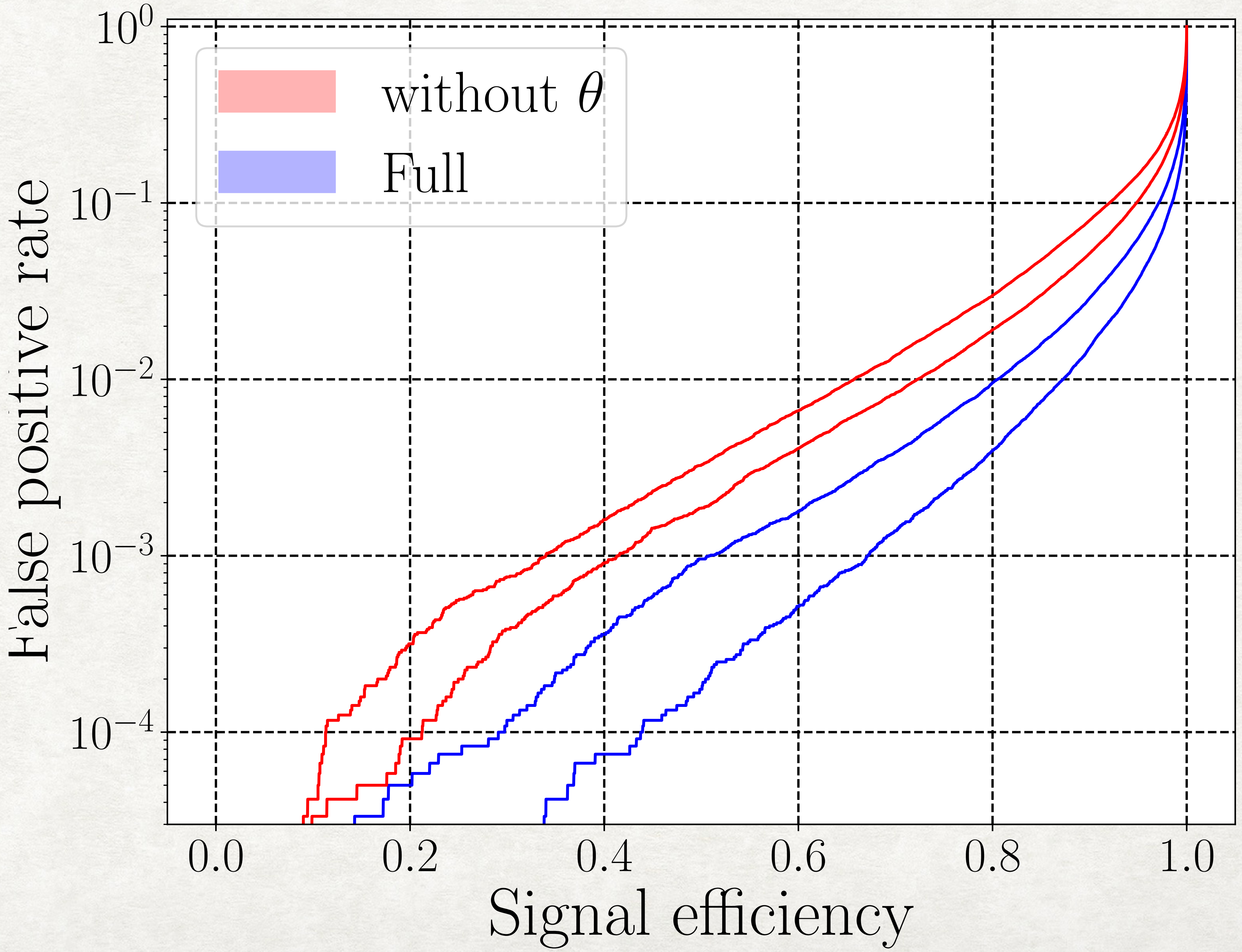
3. add “θ” : jet shape and correletion

IMPROVEMENT USING CROSS ATTENTION

factor 5 improvement at the same acceptance.

Decay correlation is important
(because QCD jets are color connected)





Cross attention improves the rejection efficiency significantly

SUMMARY

LHC process

Hard scattering

Jet function Parton shower

$$\sigma(pp \to a, b \to N\text{jets}) \sim H_N \left[B_a B_b \prod_{k=1}^{N} J_k \right] \otimes S_N \,,$$

Cross attention for P(constituents | (sub)jets~partons)

Something soft

constituent information

$A = QK^T$

Local information via (sub)jets

SUMMERY

Jet classification:Mixer+ Subjet network

- Small, first, and high performance (you can test it on your computer!)

- you can stack all information (vertex, track, etc)

Global event analysis by Deep learning

- correlation beyond a jet