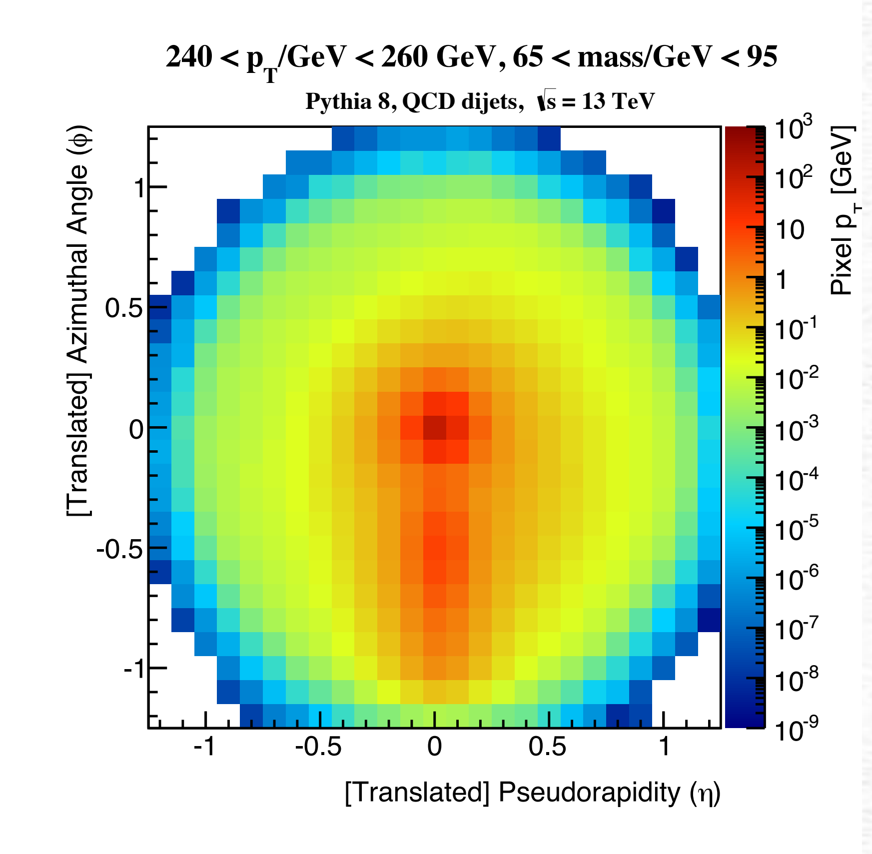
streamlined jet tagging network assisted by jet prong structure -role of cross attention-

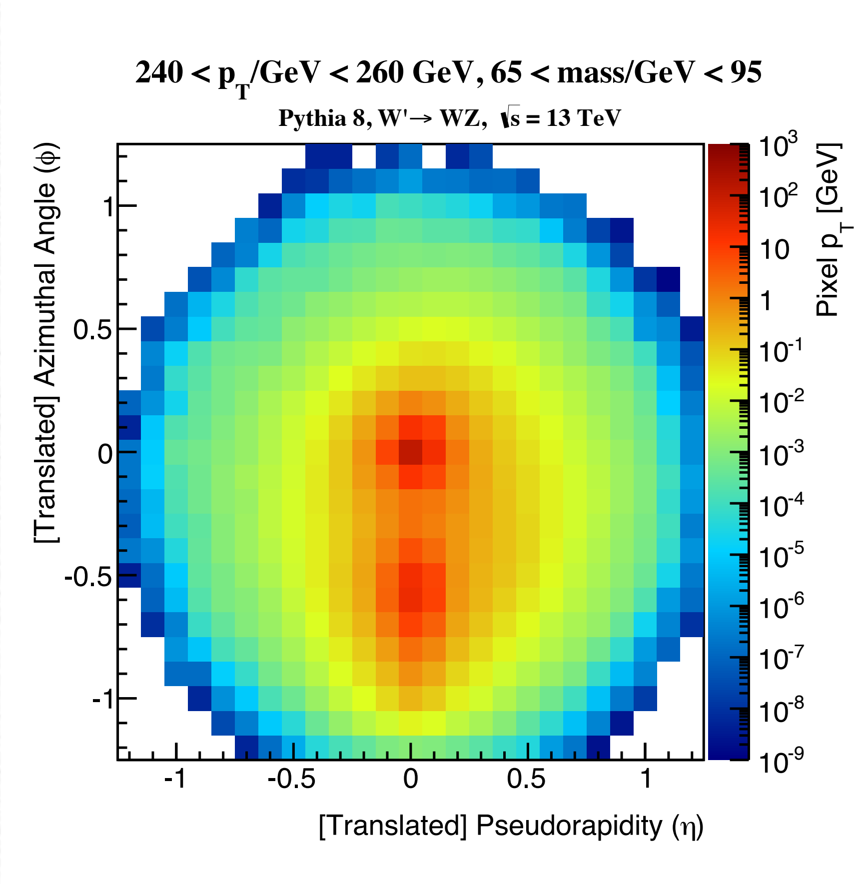
Mihoko Nojiri(IPNS, KEK), with Ahmed Hammad and Stefano Moretti
arXiv 2401.00452 JHEP 03(2024) 144
Mihoko Nojiri with Ahmed Hammad
arXiv 2404 14677 JHEP 06 (2024) 176

Jet classification using ML

QCD jet
(in W mass region)

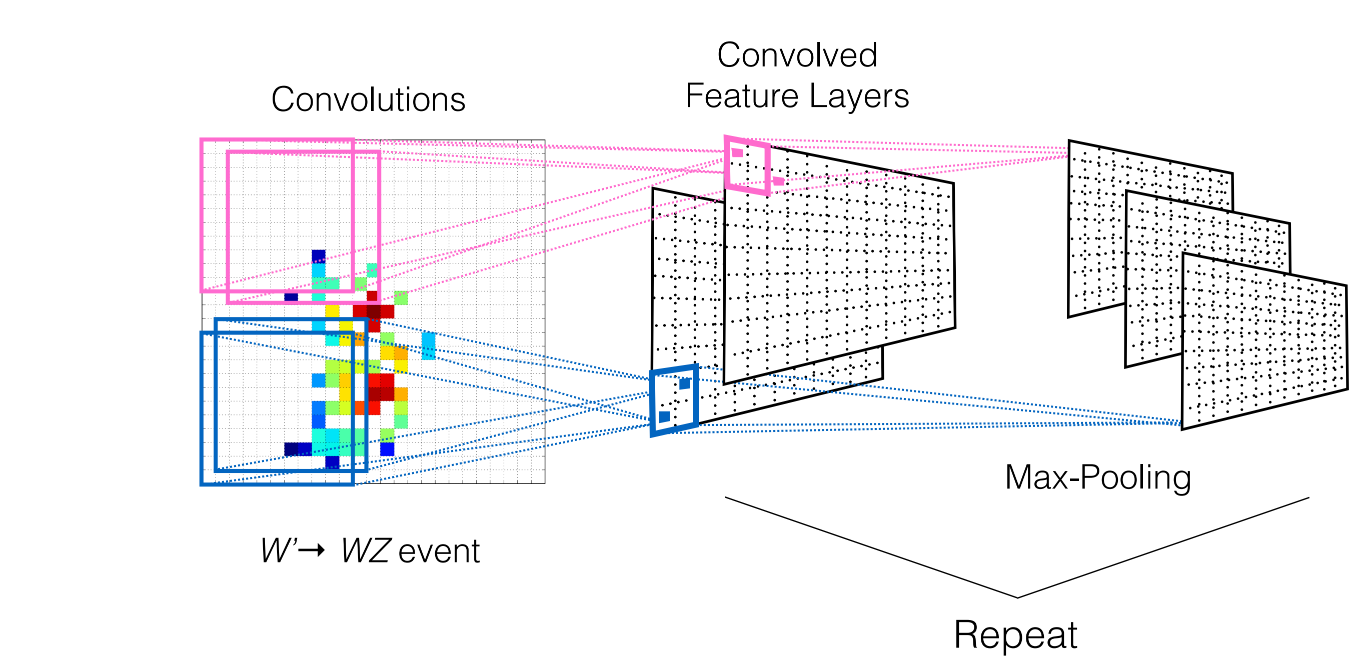


W jet

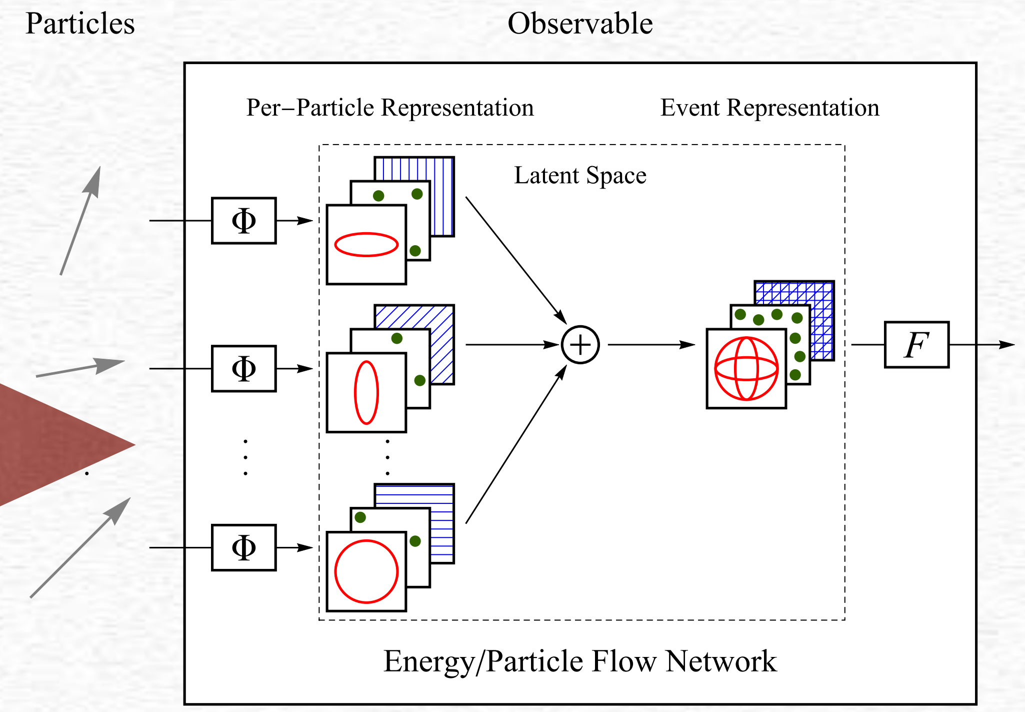


from Schwartzman et all
https://iopscience.iop.org/article/10.1088/1742-6596/762/1/012035

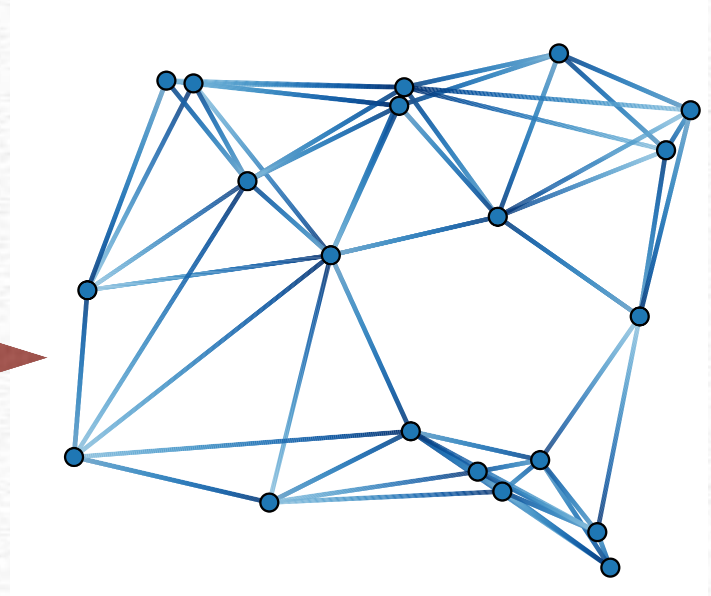
Jet Image



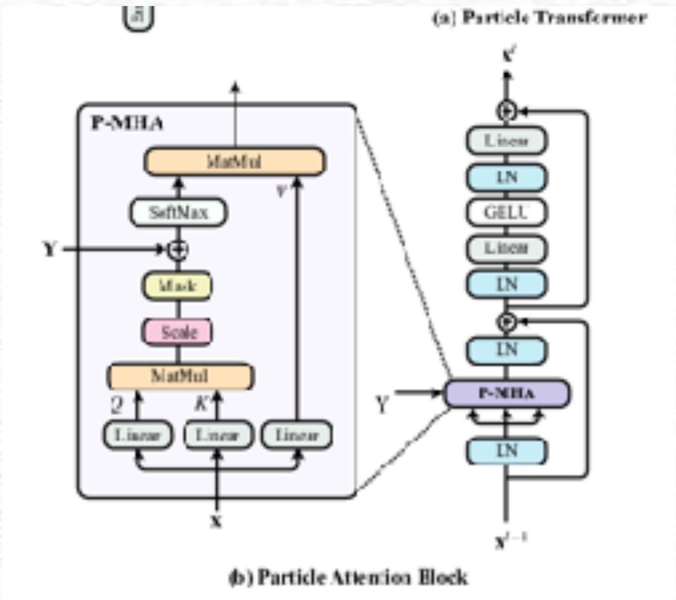
as sets



as graphs



transformer



CNN Oliverira et al
(1511.05190)

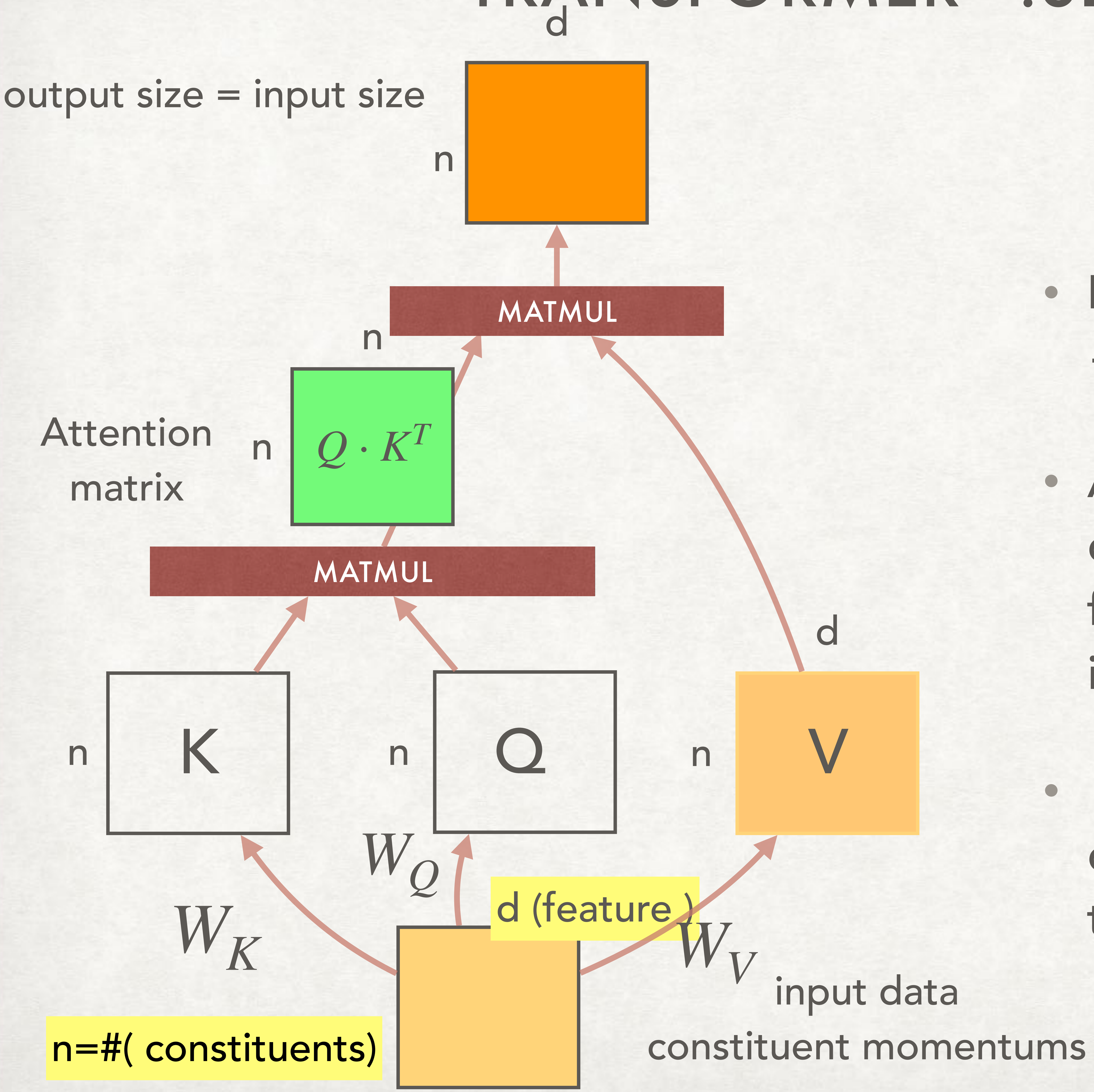
permutation invariance
(Energy Flow Network and
Particle Flow Network 1810.05165)

sparse data
1902.08570 Particle Net
Dreyer et al LundNet (1807.04758)
Gong et al LorentzNet(2201.08187)

building key
and query
2202.03772

Bogatskiy et al PELICAN (2211.00454)

"TRANSFORMER" :SELF ATTENTION LAYERS



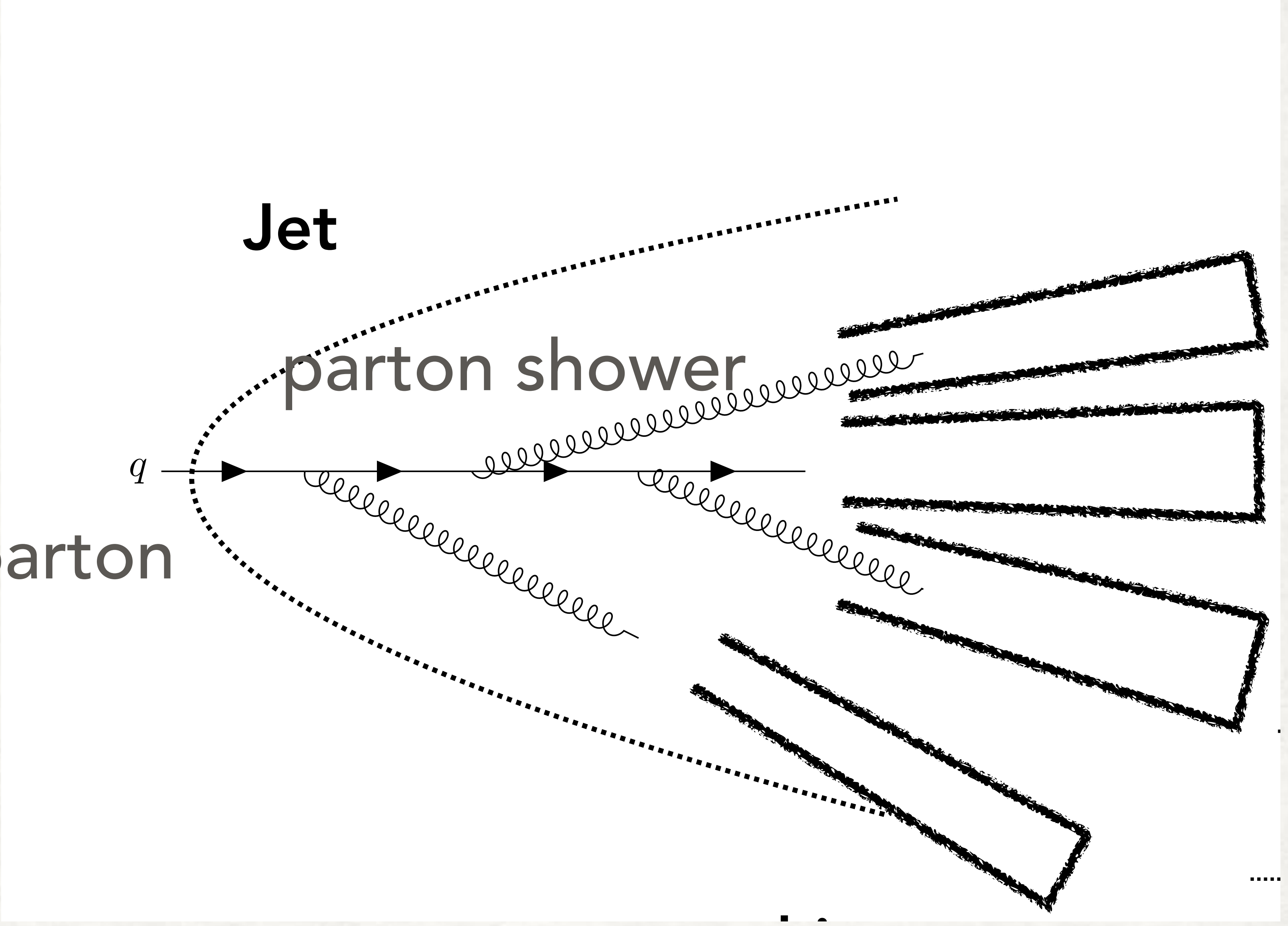
$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

- 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
- transformation does not change the data dimension. Structure of data retained for the next transformation.

BUT… PHYSICS BEFORE THE NETWORK

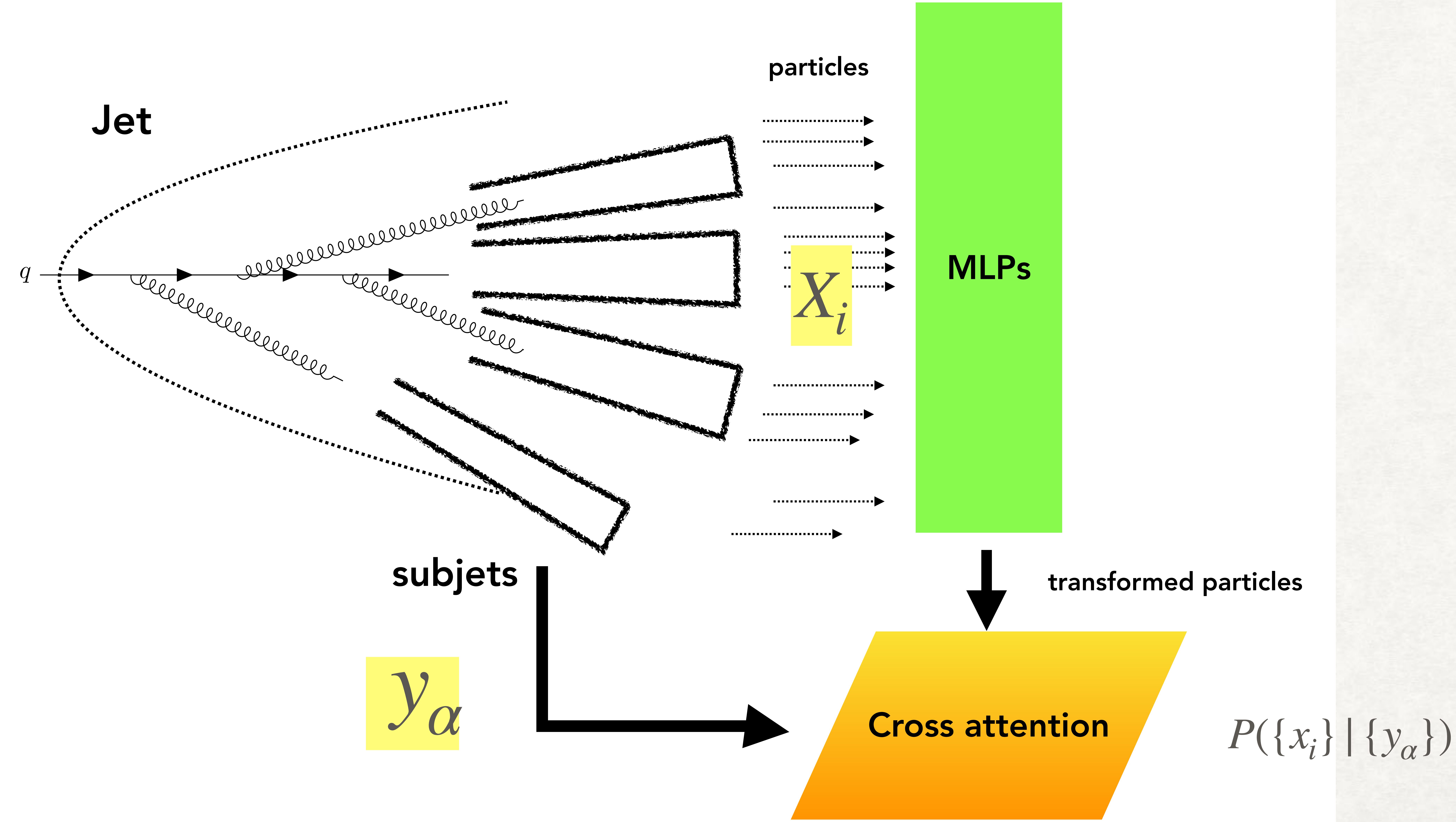
"Physics SCALE"

- Hard Process = Partons y
- Factorization
  - a jet: P(hadrons in jets | parton) = $P(\{x_i\}\,|\,y)$
  - jet with substructure $P(\{x_i\}\,|\,\{y_\alpha\})$
  - Maybe several fatjets in an event



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

Why don't you construct the network focusing on QCD scale structure



particles
Jet
q
x_i
MLPs
subjets
transformed particles
y_α
Cross attention
$P(\{x_i\} \mid \{y_\alpha\})$

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

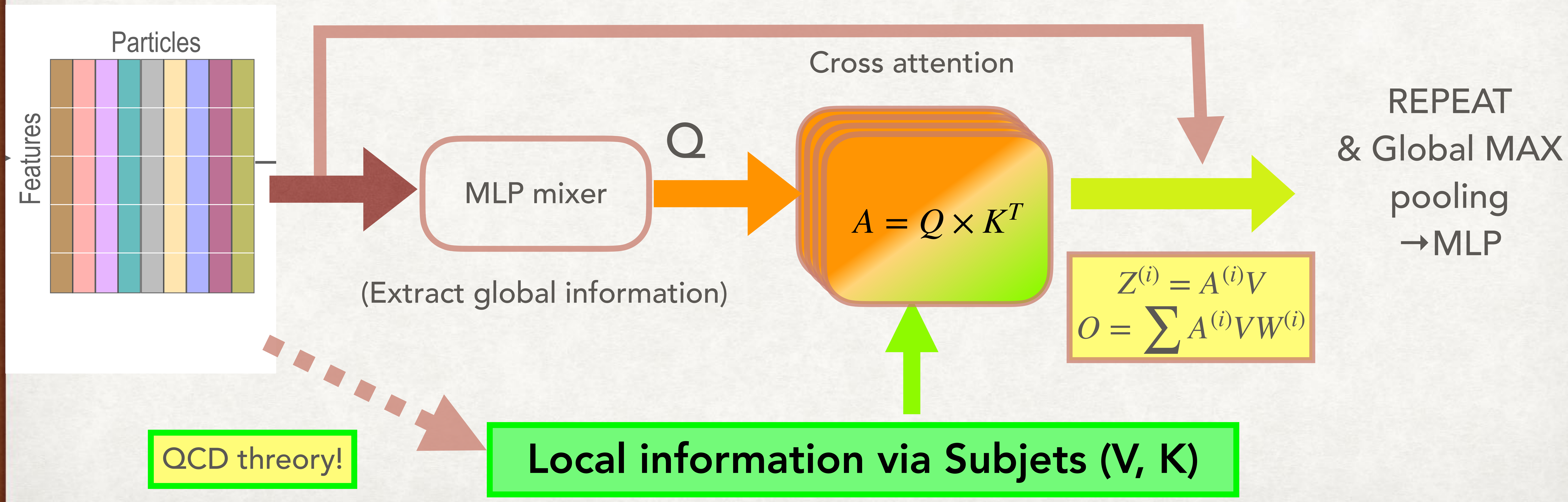
24) 176

24) 176

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

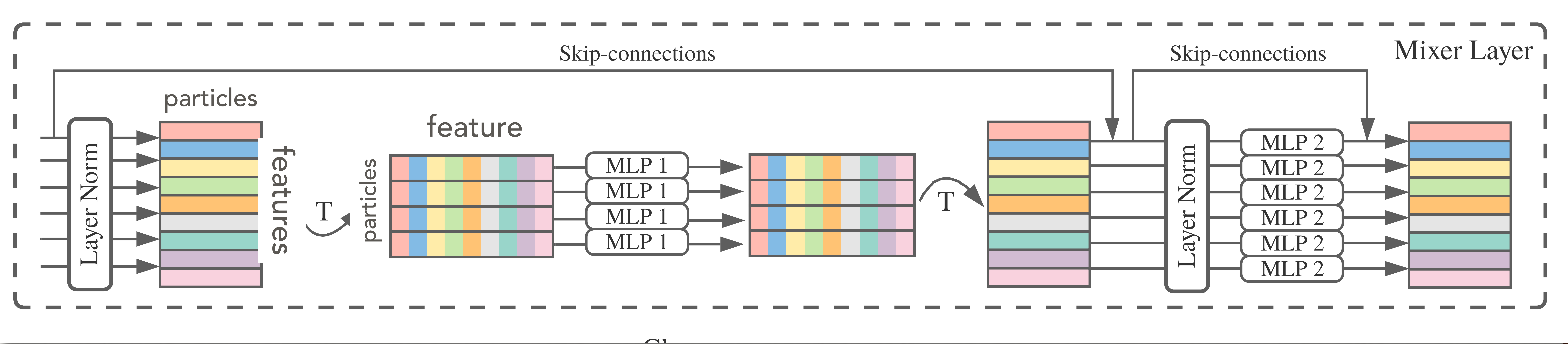
input X

skip connection $\tilde{X} = X + O$



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

transformer-like (add any information, apply repeatedly)

"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

INTERPRETATION USING CKA SIMILARITY

GLOBAL(MIXER) AND LOCAL(SUBJET)

input — n event

Network1

hidden layer — $X(n \times h_1)$

Network 2 — Do we need this?

hidden layer — $Y(n \times h_2)$

output

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

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

$$H = \delta_{ij} - \frac{1}{d}$$

1 if they are same (no improvedment)

matrix X, and Y varies for each training but how about $M = XX^T, N = YY^T$

Efficient!

CKA-Top jets

	Embedding	$FC^1(MLP_1)$	$FC^2(MLP_1)$	$FC^1(MLP_2)$	$FC^2(MLP_2)$	SkipLayer	Attention	FC
Embedding	1	0.66	0.56	0.31	0.4	0.53	0.17	0.12
$FC^1(MLP_1)$	0.66	1	0.74	0.29	0.48	0.57	0.2	0.16
$FC^2(MLP_1)$	0.56	0.74	1	0.53	0.58	0.46	0.28	0.15
$FC^1(MLP_2)$	0.31	0.29	0.53	1	0.39	0.17	0.12	0.047
$FC^2(MLP_2)$	0.4	0.48	0.58	0.39	1	0.56	0.3	0.07
SkipLayer	0.53	0.57	0.46	0.17	0.56	1	0.17	0.071
Attention	0.17	0.2	0.28	0.12	0.3	0.17	1	0.18
FC	0.12	0.16	0.15	0.047	0.07	0.071	0.18	1

CKA-QCD jets

	Embedding	$FC^1(MLP_1)$	$FC^2(MLP_1)$	$FC^1(MLP_2)$	$FC^2(MLP_2)$	SkipLayer	Attention	FC
Embedding	1	0.61	0.61	0.48	0.49	0.46	0.19	0.39
$FC^1(MLP_1)$	0.61	1	0.8	0.65	0.64	0.28	0.22	0.55
$FC^2(MLP_1)$	0.61	0.8	1	0.76	0.62	0.61	0.25	0.68
$FC^1(MLP_2)$	0.48	0.65	0.76	1	0.57	0.43	0.15	0.48
$FC^2(MLP_2)$	0.49	0.64	0.62	0.57	1	0.34	0.57	0.72
SkipLayer	0.46	0.28	0.61	0.43	0.34	1	0.21	0.34
Attention	0.19	0.22	0.25	0.15	0.57	0.21	1	0.58
FC	0.39	0.55	0.68	0.48	0.72	0.34	0.58	1

TOWARD GLOBAL EVENT ANALYSIS

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

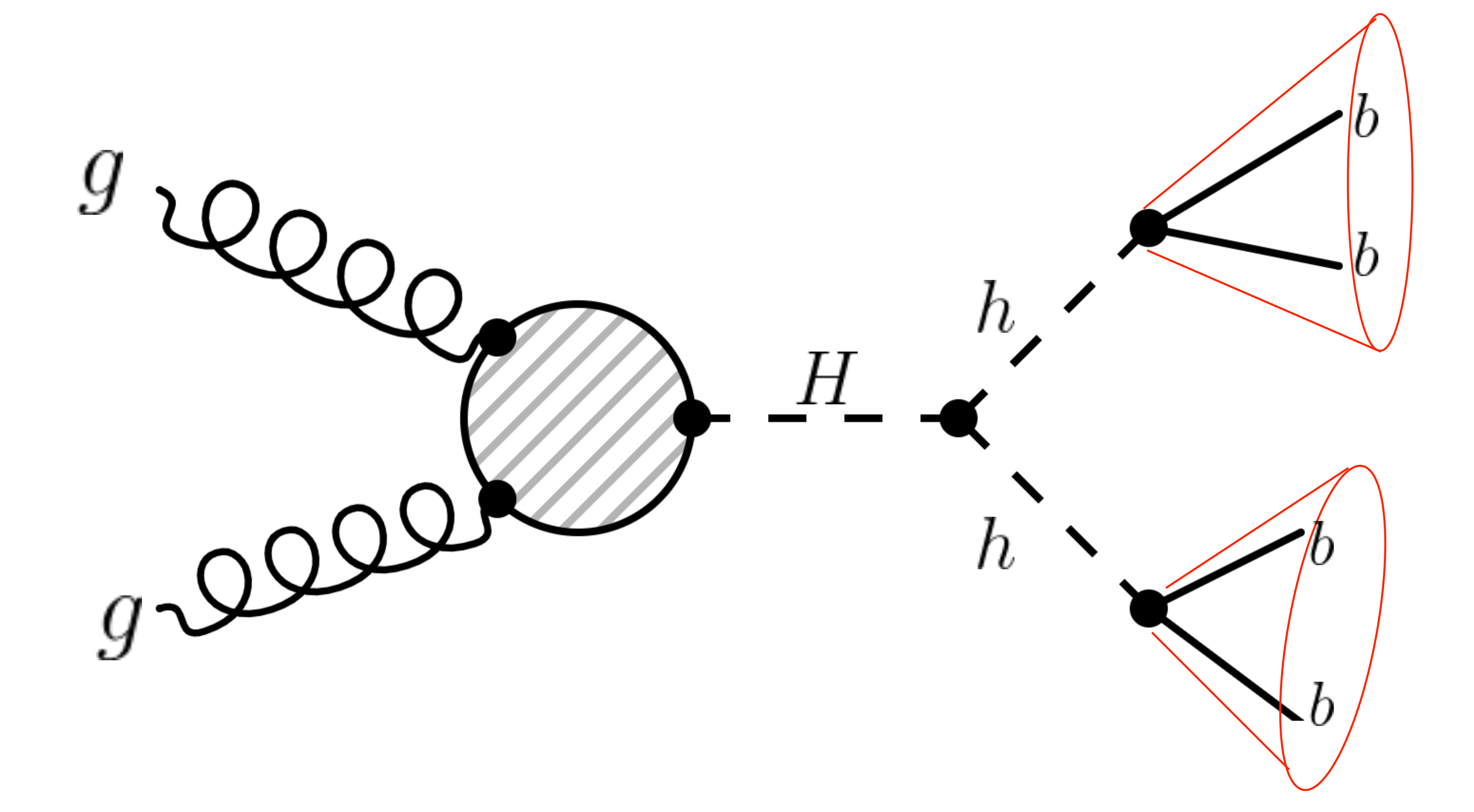
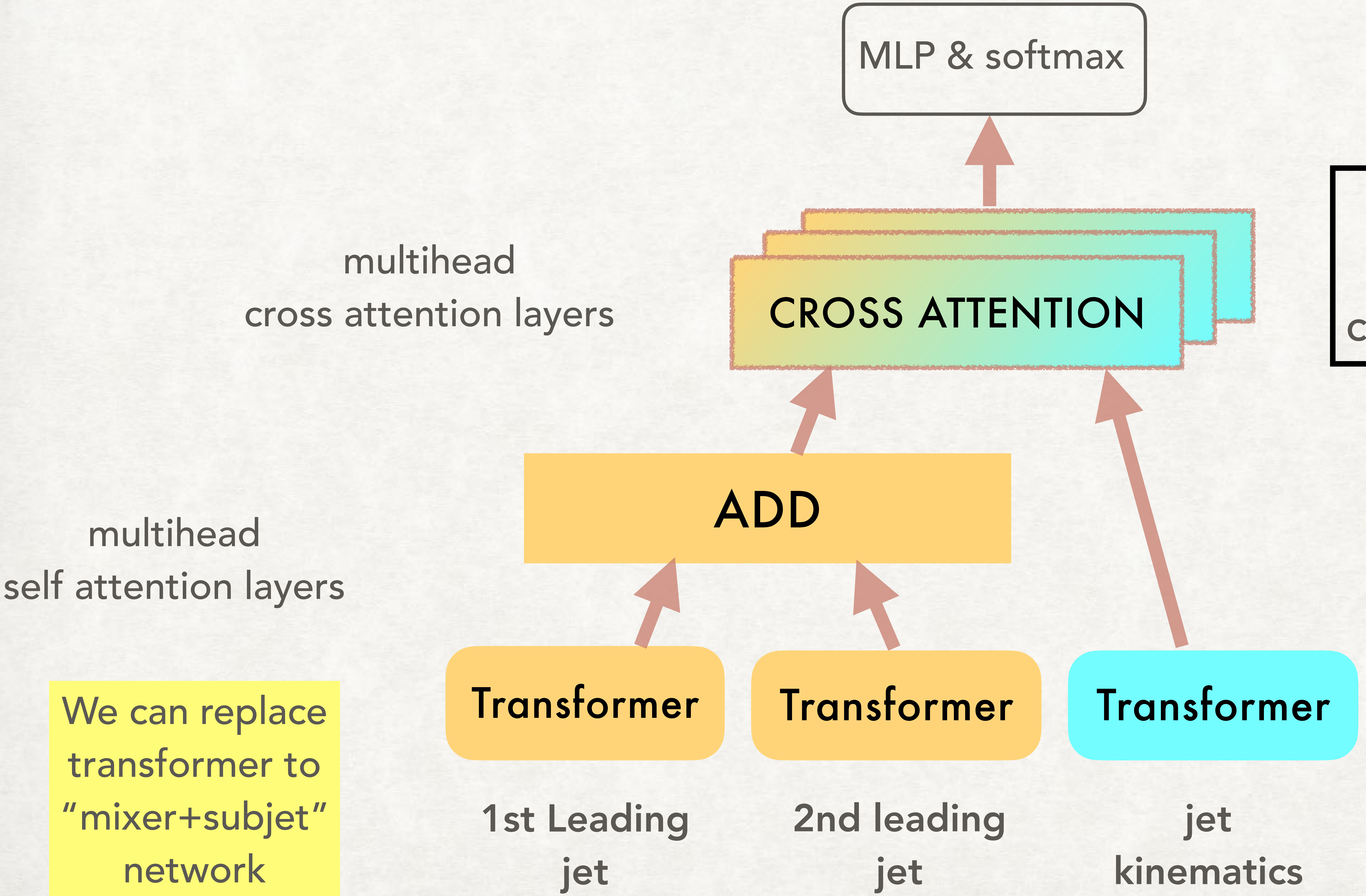


Figure 2: Feynman diagram for the signal process.

cross attention motention for 2 fatjet events



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

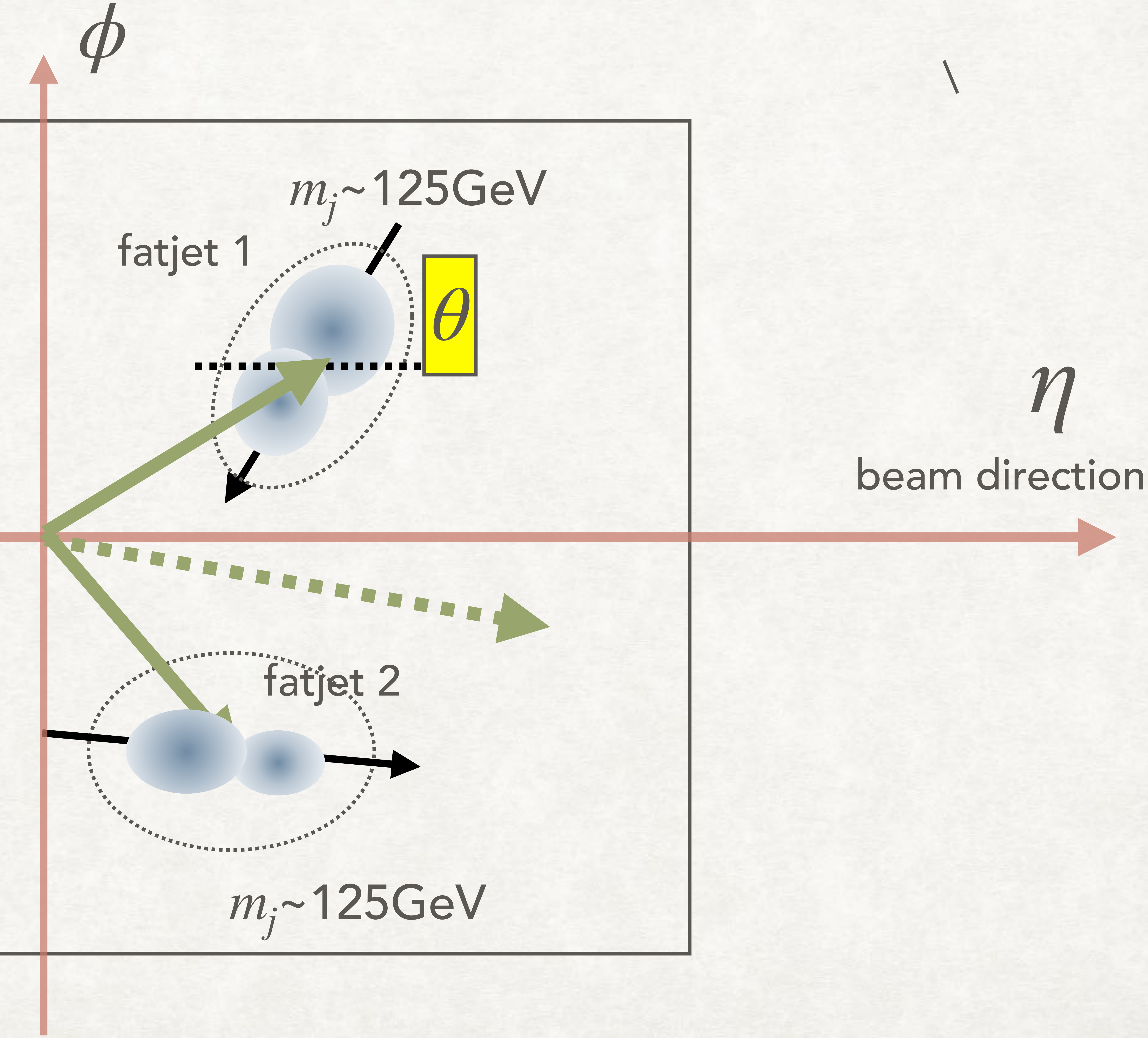
INPUT TO NETWORK : EVENT KINEMATICS

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" ,

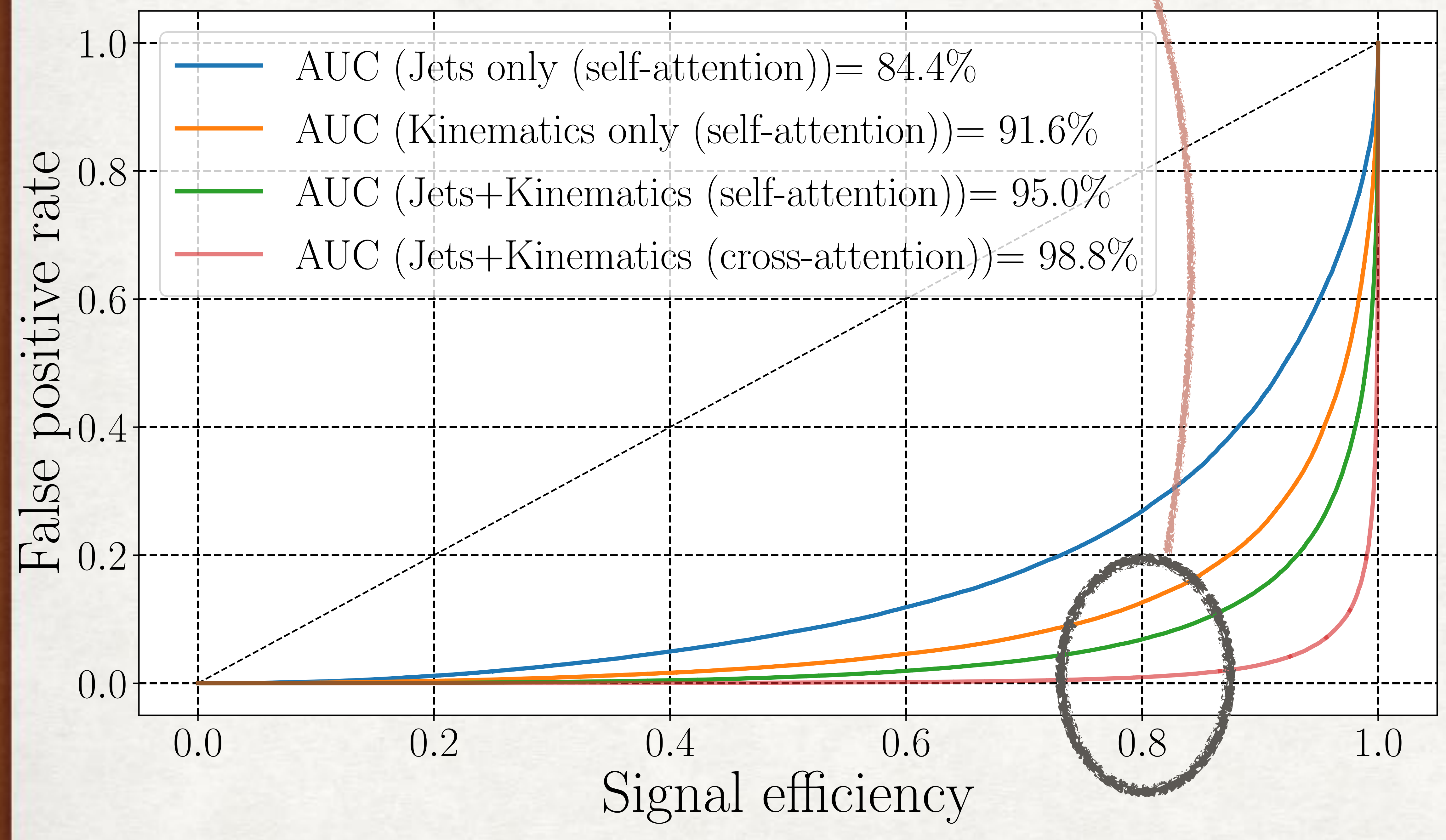
2. H candidate momentum as sum of the fat jet momentum.

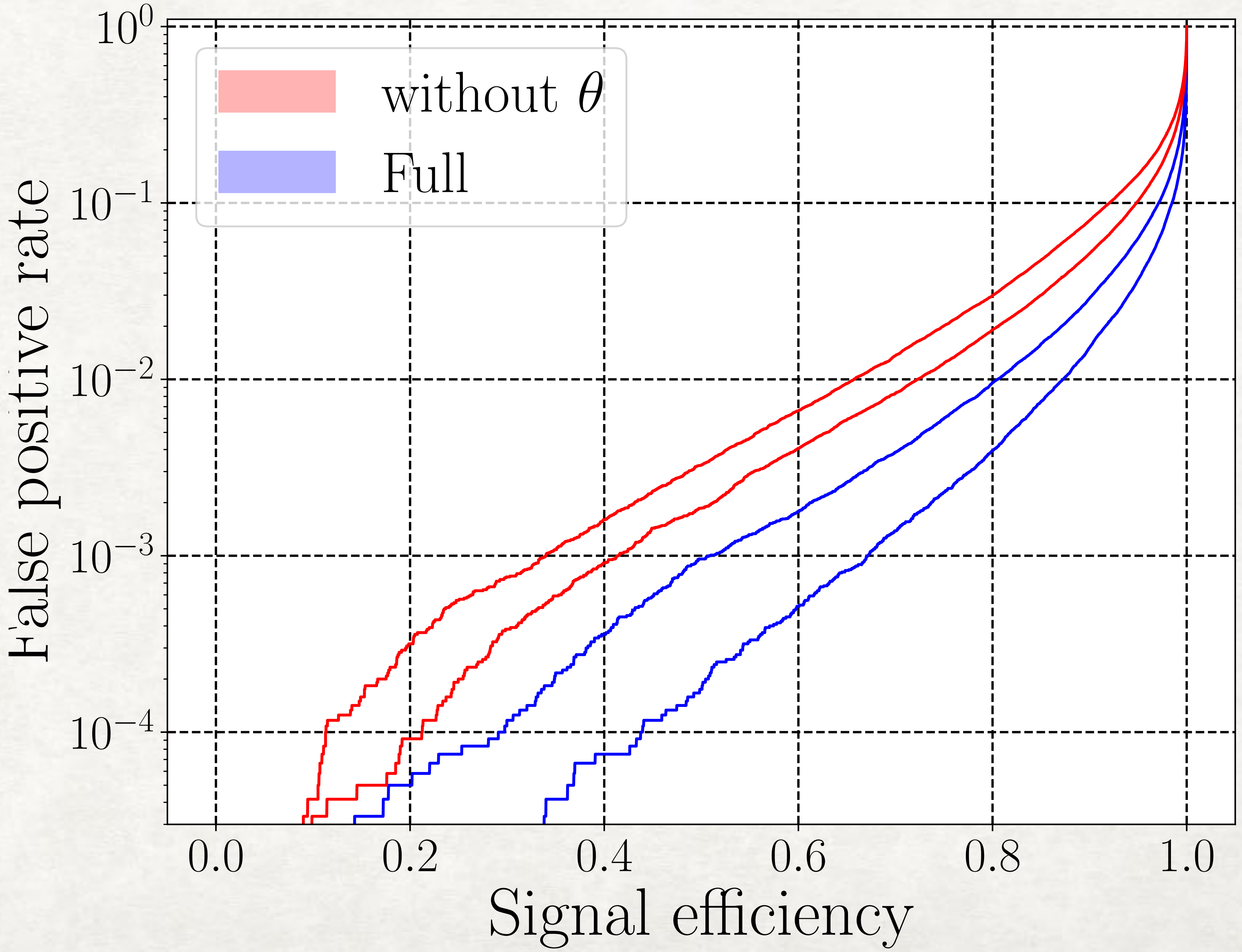
3. add "θ" :the correlation beyond a subjet

IMPROVEMENT USING CROSS ATTENTION

factor 5 improvement at the same acceptance.

Decay correlation is important
(because QCD background is correlated)





Cross attention improves the rejection efficiency significantly

SUMMARY

LHC process

Hard scattering

Jet function Parton shower

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

Something soft

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

constituent information

$A = QK^T$

Local information via (sub)jets K

SUMMERY

Mixer+ Subjet network

- Small, first, and high perfomance (you can test it on your computer!)
- Can apply repeatedly without losing information.
- you can stack all information (vertex, track, etc)