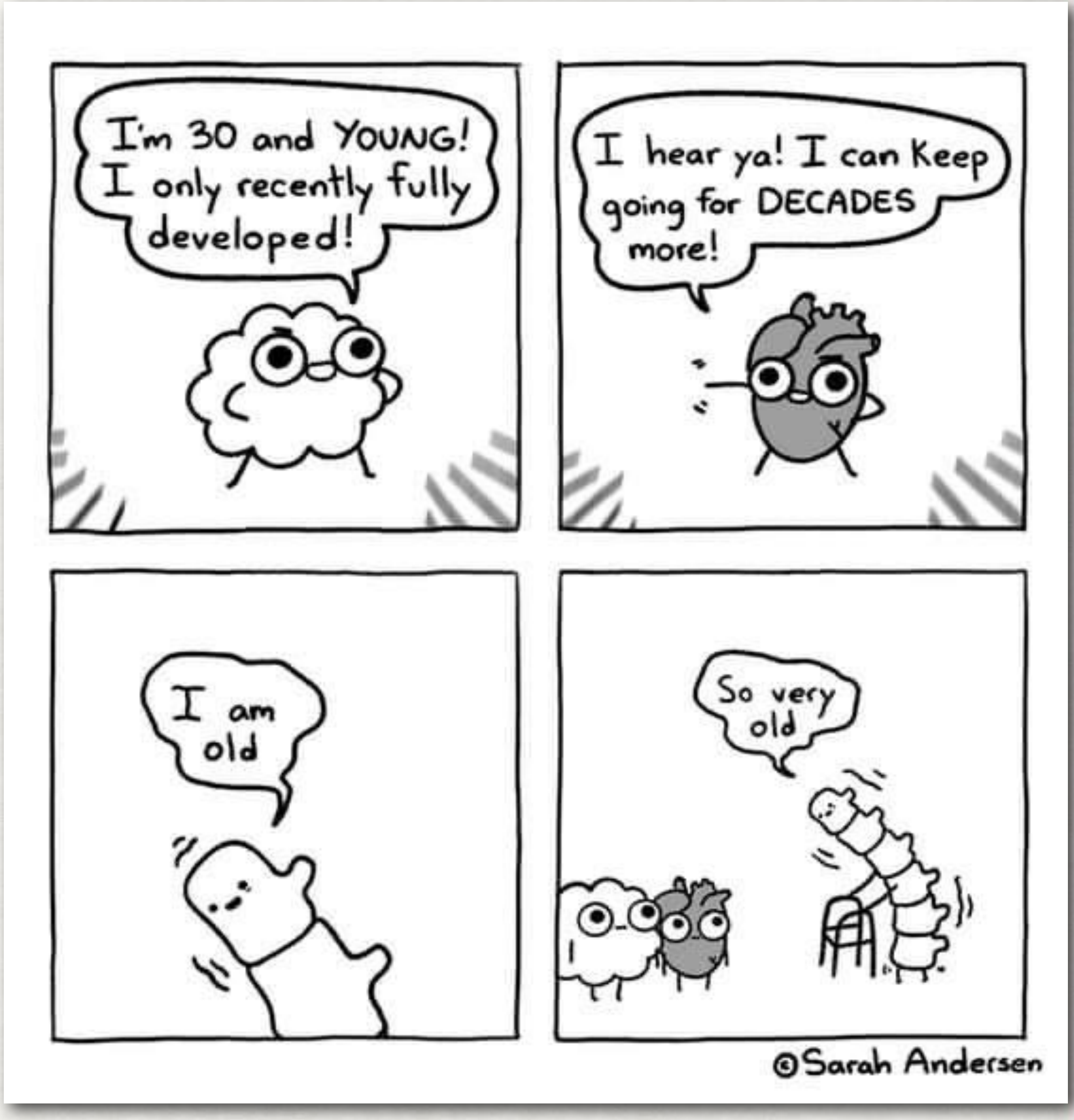
Multi-scale Cross-Attention Transformer encoder for event classification

Mihoko Nojiri(IPNS, KEK), with Ahmed Hammad Stefano Moretti
arXiv 2401.00452

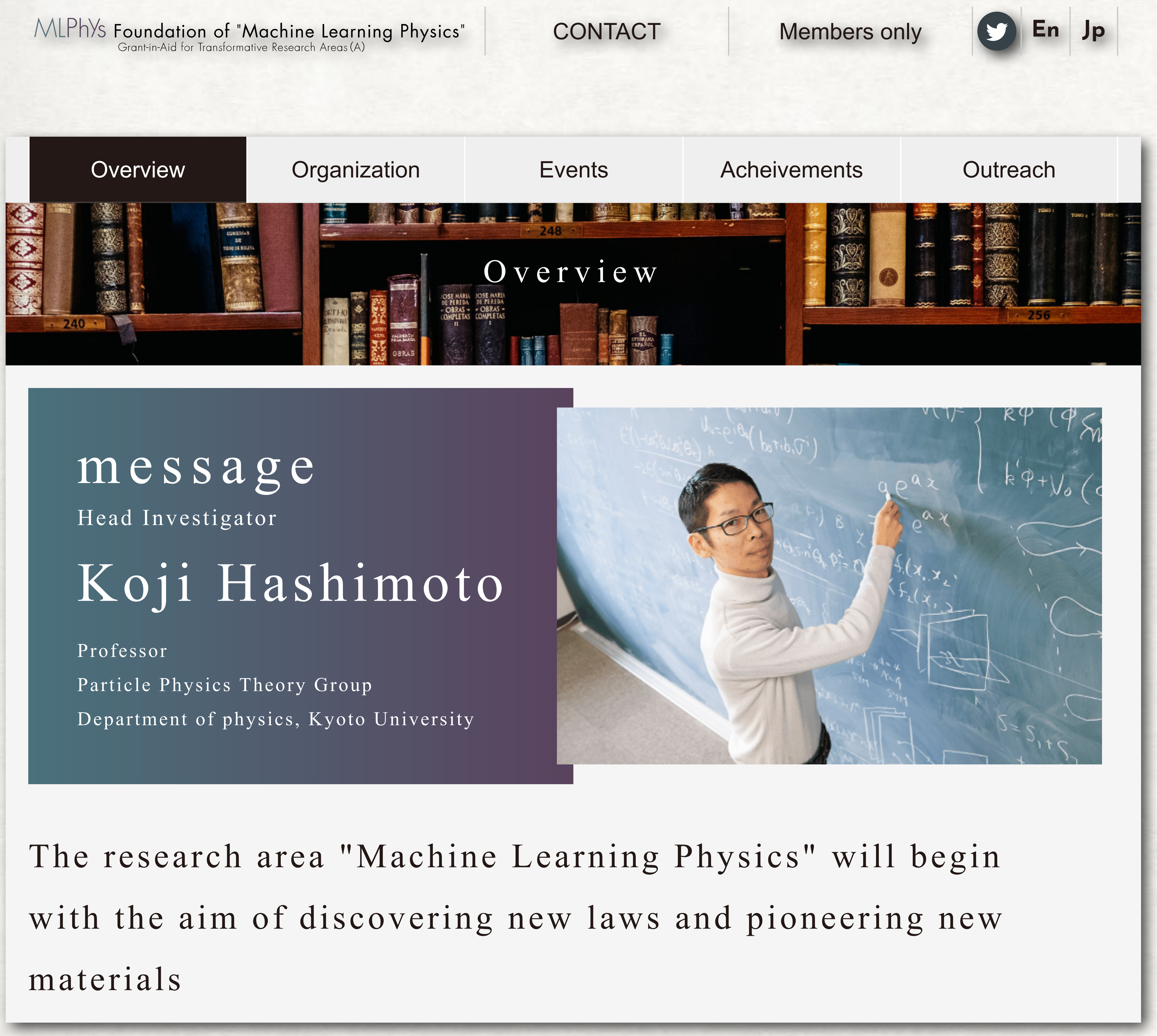
ABOUT MYSELF

- PhD Kyoto (1990) a bit old.
- PD: Supergravitiy study in heavy top era →   SUSY dark matter. One of the author of first Sommerfeld effect in dark matter annihilation. (2003)
- Collider:
  - 1996: JLC study and Snowmass
  - 2002-2008 LHC BSM study in ATLAS SUSY group. BSM Convener of Les Houches TeV collider workshop twice → Jet substructure study → **Deep Learning**
- **Service**: JPS executive board member →member of Science Council of Japan(SCJ) working on Diversity Issues .
  - In KEK, we just had DEI workshop last Dec, and trying establish more DEI activities. (https://www2.kek.jp/ipns/en/news/5320/)

"a young mind",
(according to Tilman Plehn)
but this makes me cry



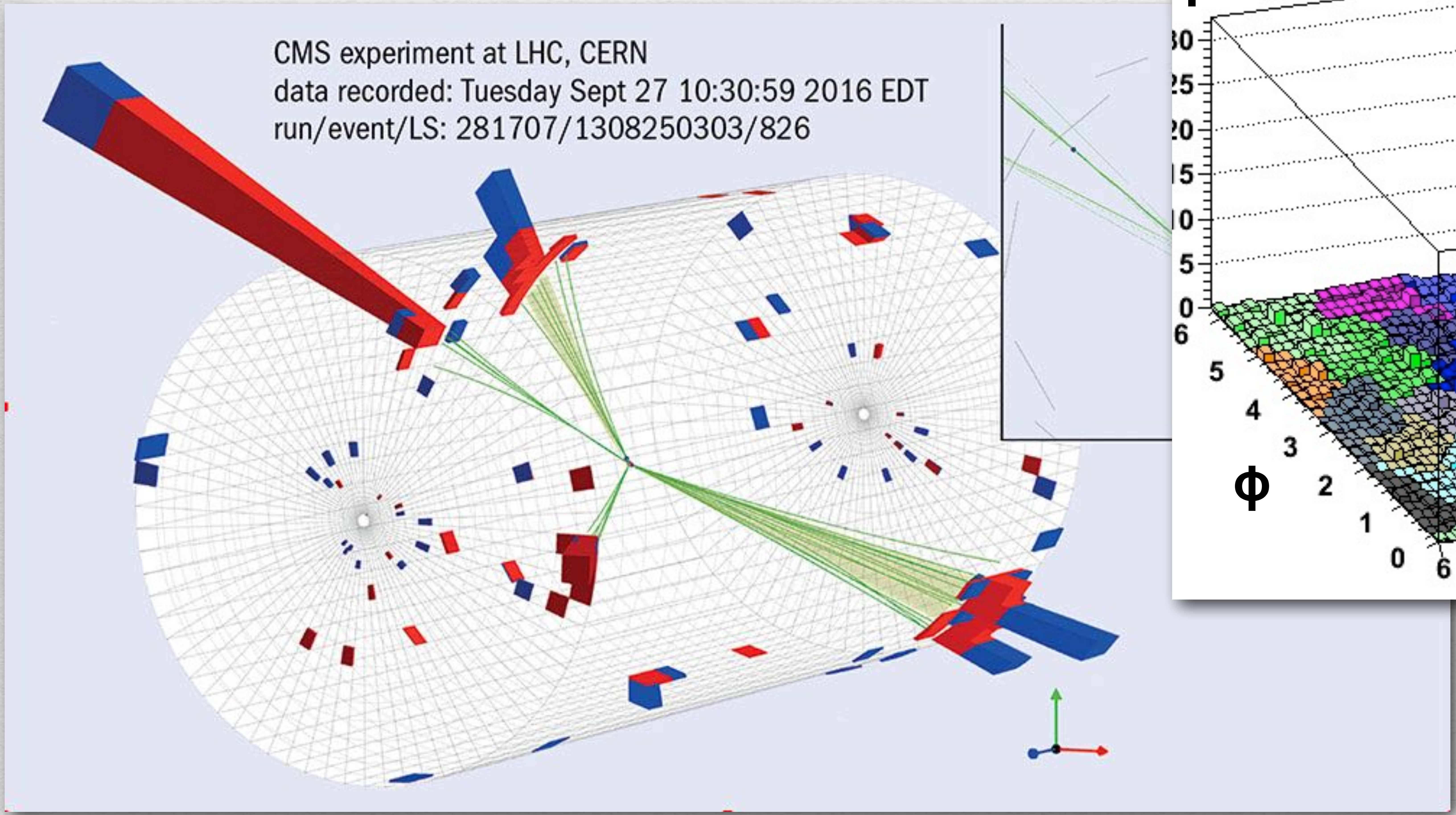
ML(THEORY) IN JAPAN: GRANT "MACHINE LEARNING PHYSICS "

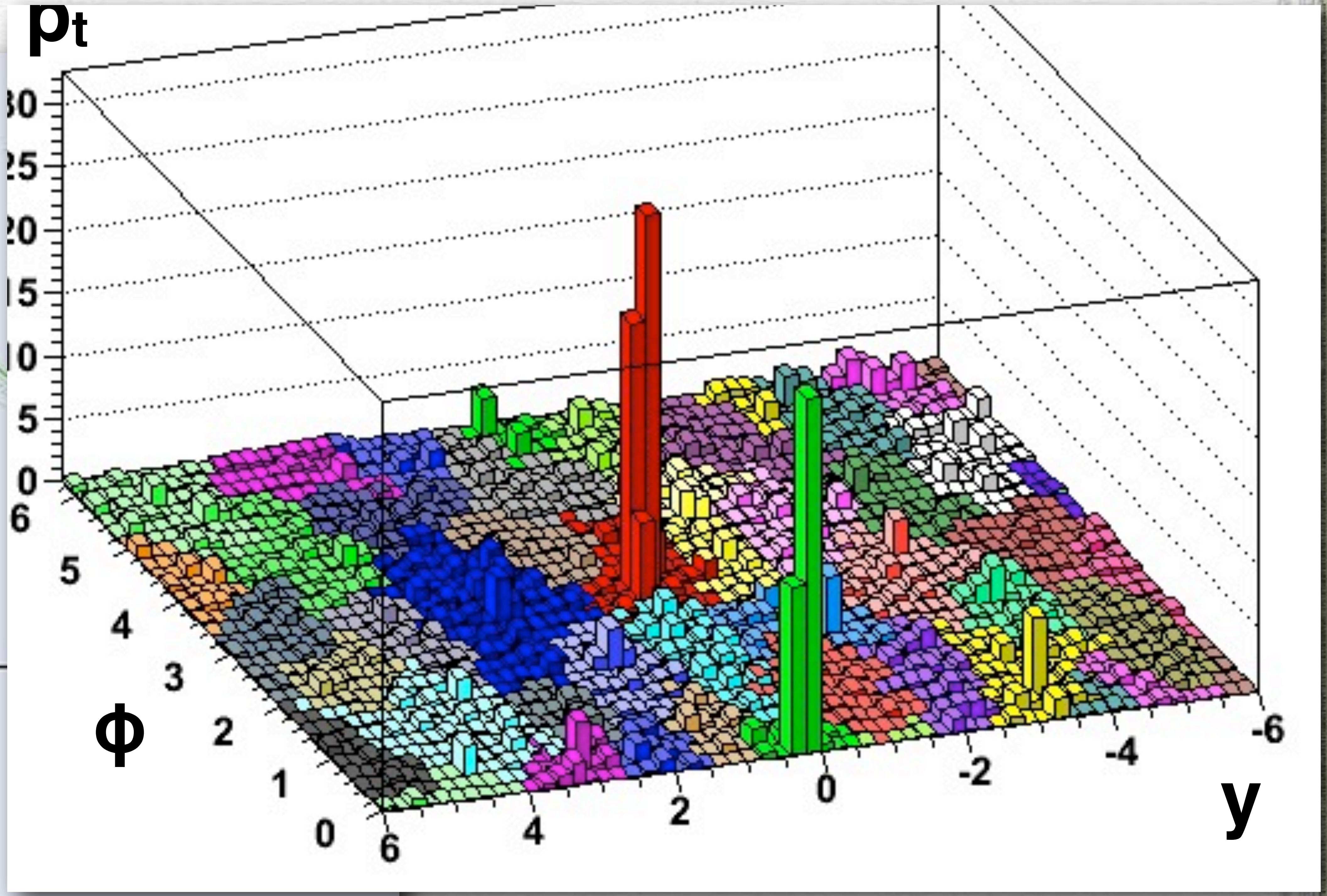


B01 Akinori Tanaka (Riken AIP) Math and application of DL
B02 Yoshiyuki Kabashima (Tokyo) Statistical data and ML
B03 Kenji Fukushima (Tokyo) Topology and Geometry of ML
A01 Akio tomiya (IPUT Osaka) Lattice
A02 Mihoko Nojiri HEP
Junichi Tanaka (ICEPP Tokyo, ATLAS)
Masako Iawasaki (Osaka Metropolitan Belle II)
Noriko Takemura and Hajime Nagahara (Data Science)
A03 Tomi Ohtsuki (Sophia U) Condensed Matter
A04 Koji Hashimoto Quantum and Gravity

Ahmed Hammad
2017-2020: Ph.D Basel University, Basel Switzerland
2020-2023: SeoulTech, Korea
2023- KEK

How machine Leaning help Collider Analysis

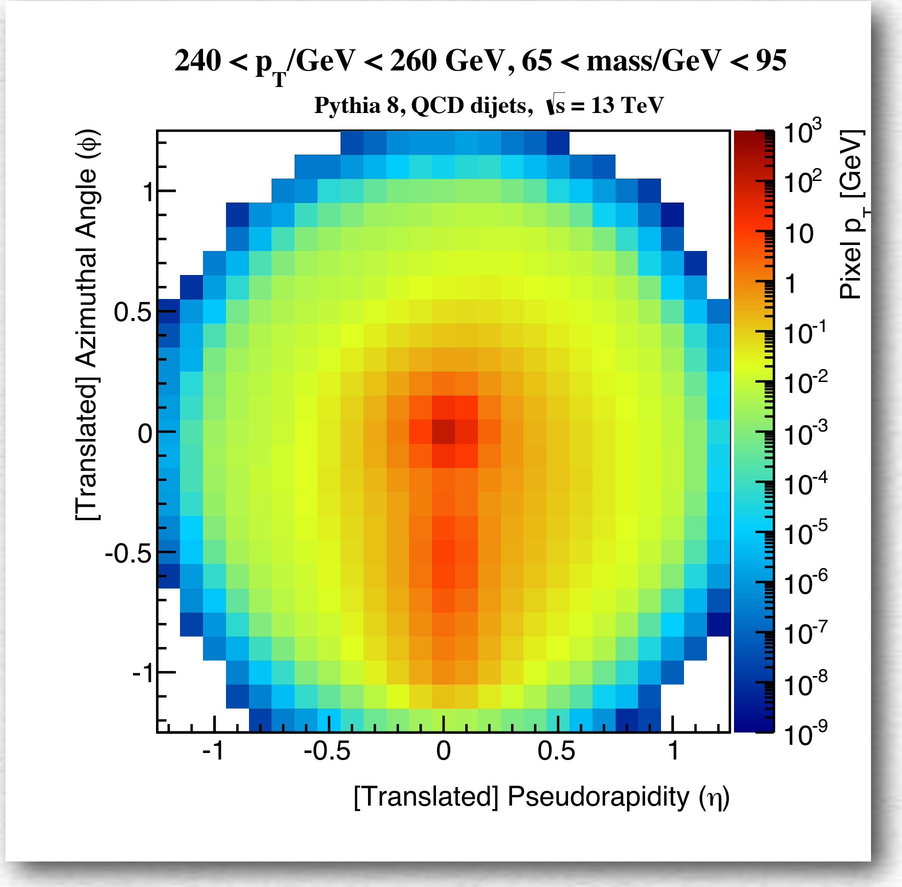




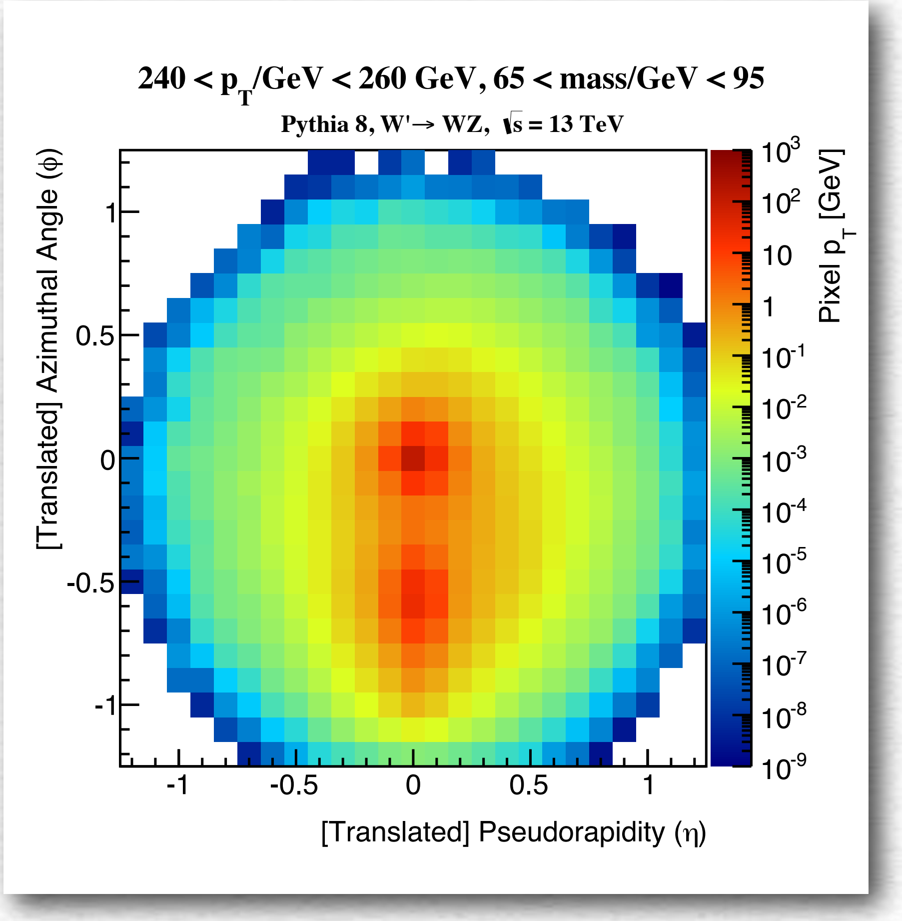
Jet clustering

Jet classification using ML

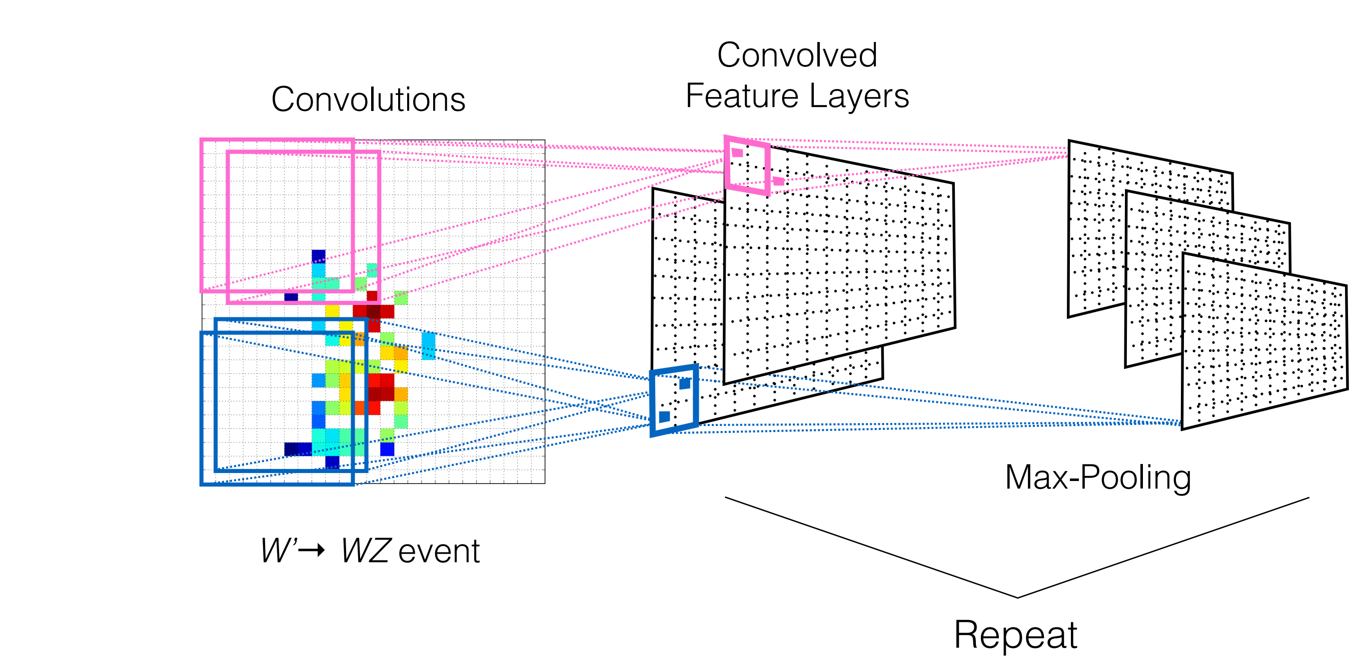
QCD jet



W jet

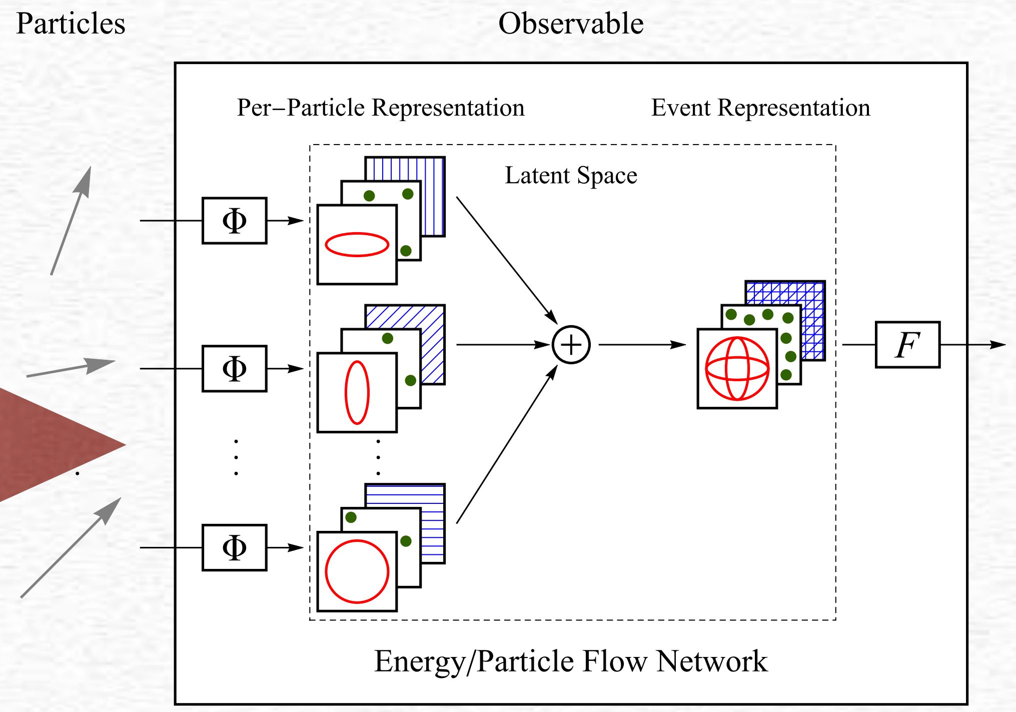


from Schwartzman et all
https://iopscience.iop.org/article/10.1088/1742-6596/762/1/012035



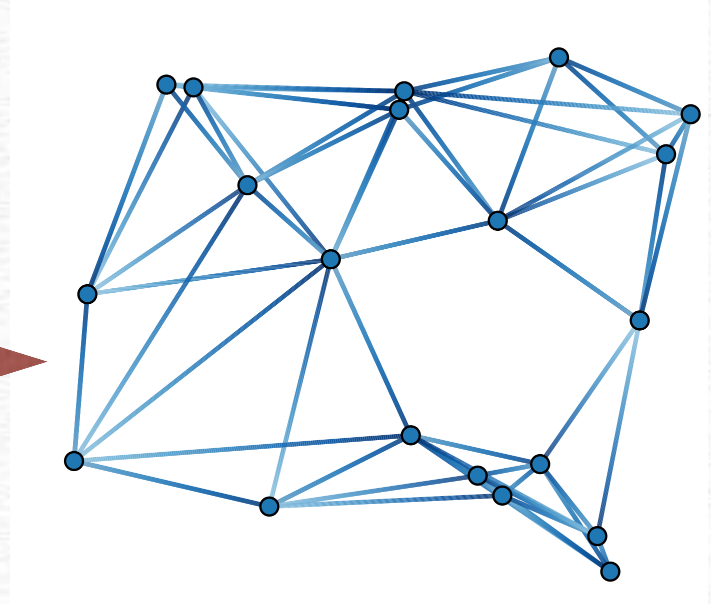
CNN Oliverira et al
(1511.05190)

as sets



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

as graphs

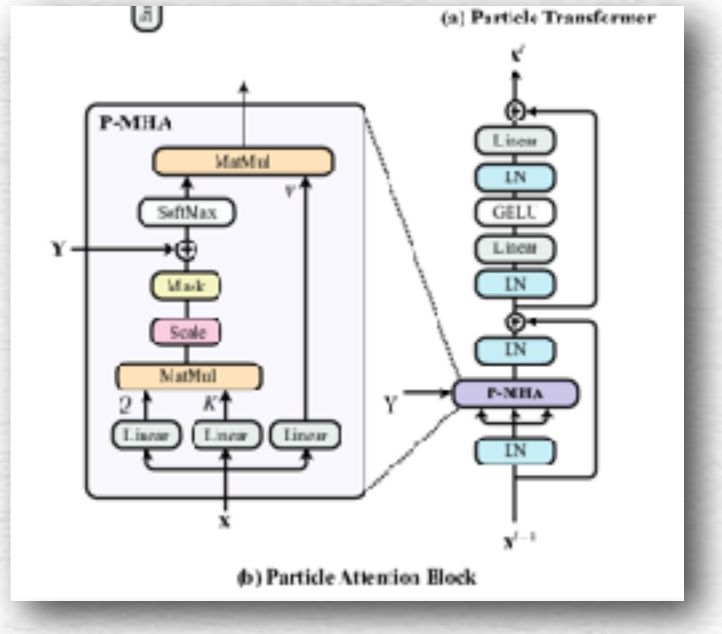


sparse data

1902.08570 Particle Net

Dreyer et al LundNet (1807.04758)
Gong et al LorentzNet(2201.08187)

transformer



building key
and query
2202.03772

Bogatskiy et al PELICAN (2211.00454)

CONNECTING JET STRUCTURE INFORMATION TO EVENT KINEMATICS

- Non SM Higgs boson (Two Higgs doublet model)
  - pp → H (Heavy Higgs boson) → hh → 4 bjet
  - mH=600-2000 GeV, mh=125.11GeV
- Meta stable vaccum of SM → extension of Higgs sector
- why doing Deep Leaning?
  - Sensitivity under S/BG~1 scale by $1/\sqrt{N}$
    with background rejection $1/N$

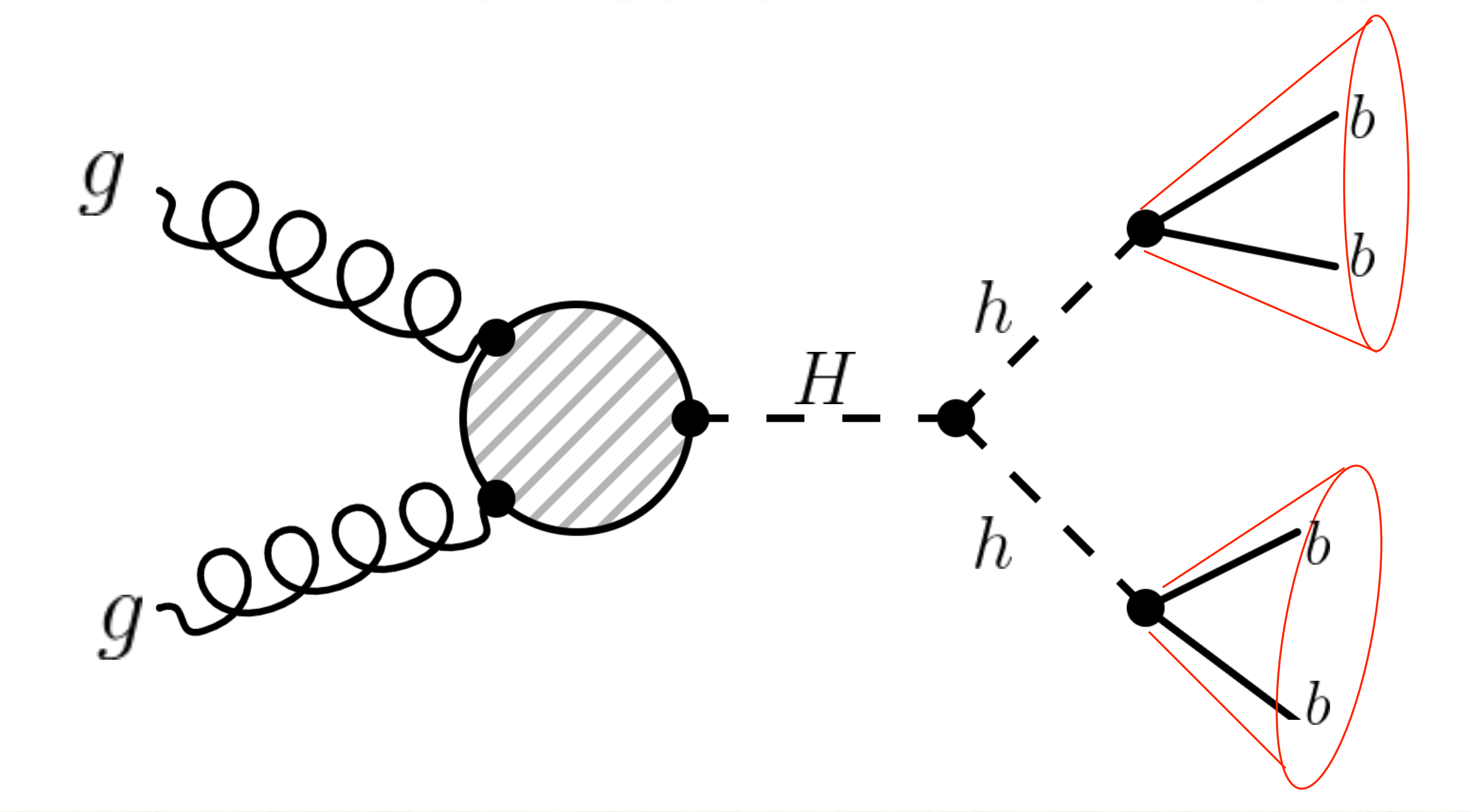
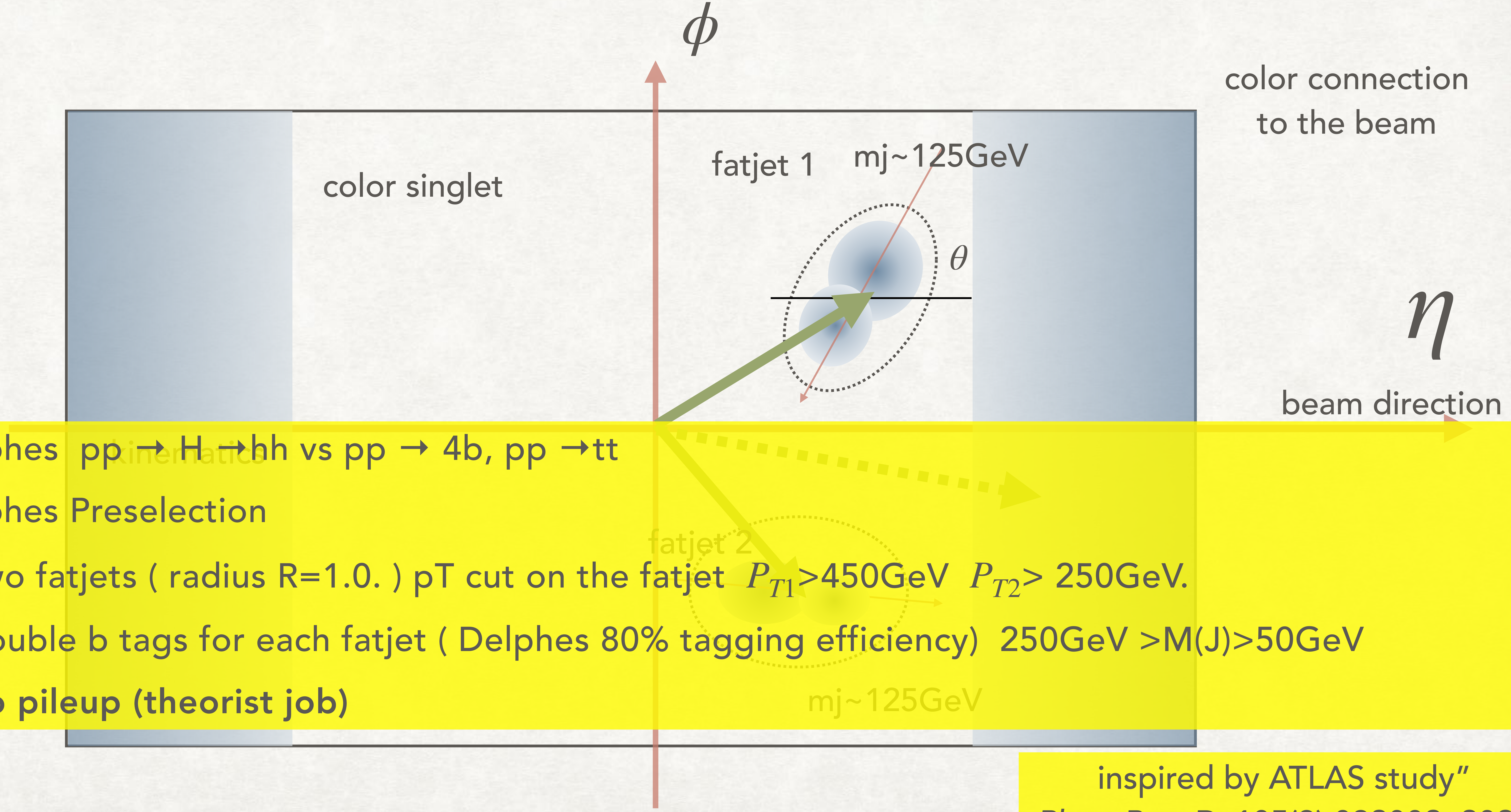


Figure 2: Feynman diagram for the signal process.

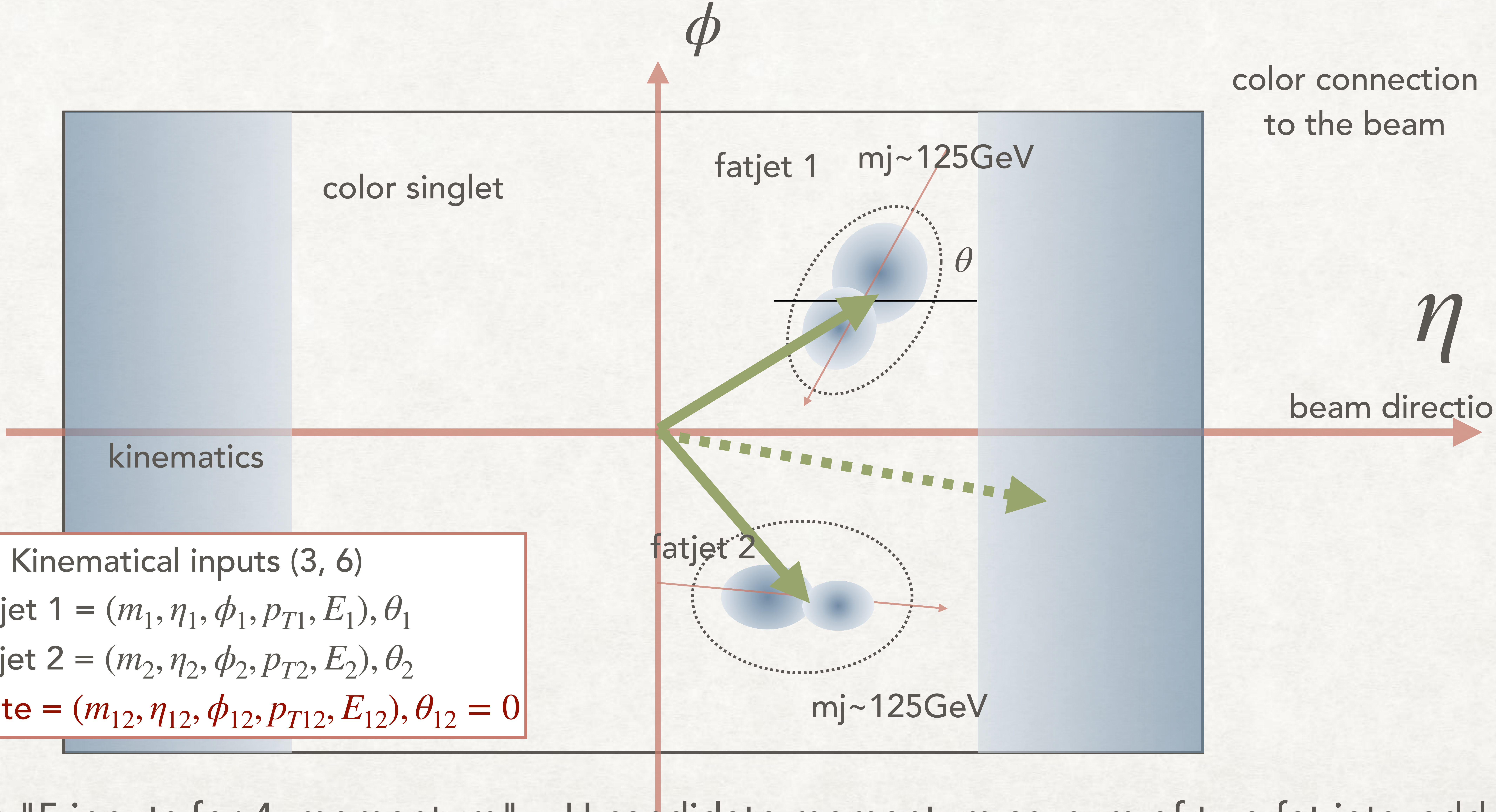
HOW SIGNAL LOOKS LIKE: KINEMATICAL INPUTS



- Delphes pp → H →hh vs pp → 4b, pp →tt
- Delphes Preselection
  - two fatjets (radius R=1.0.) pT cut on the fatjet P_{T1}>450GeV P_{T2}> 250GeV.
  - double b tags for each fatjet (Delphes 80% tagging efficiency) 250GeV >M(J)>50GeV
  - **no pileup (theorist job)**

inspired by ATLAS study"
Phys. Rev. D, 105(9):092002, 2022.

INPUT TO NETWORK : EVENT KINEMATICS



Kinematical inputs (3, 6)

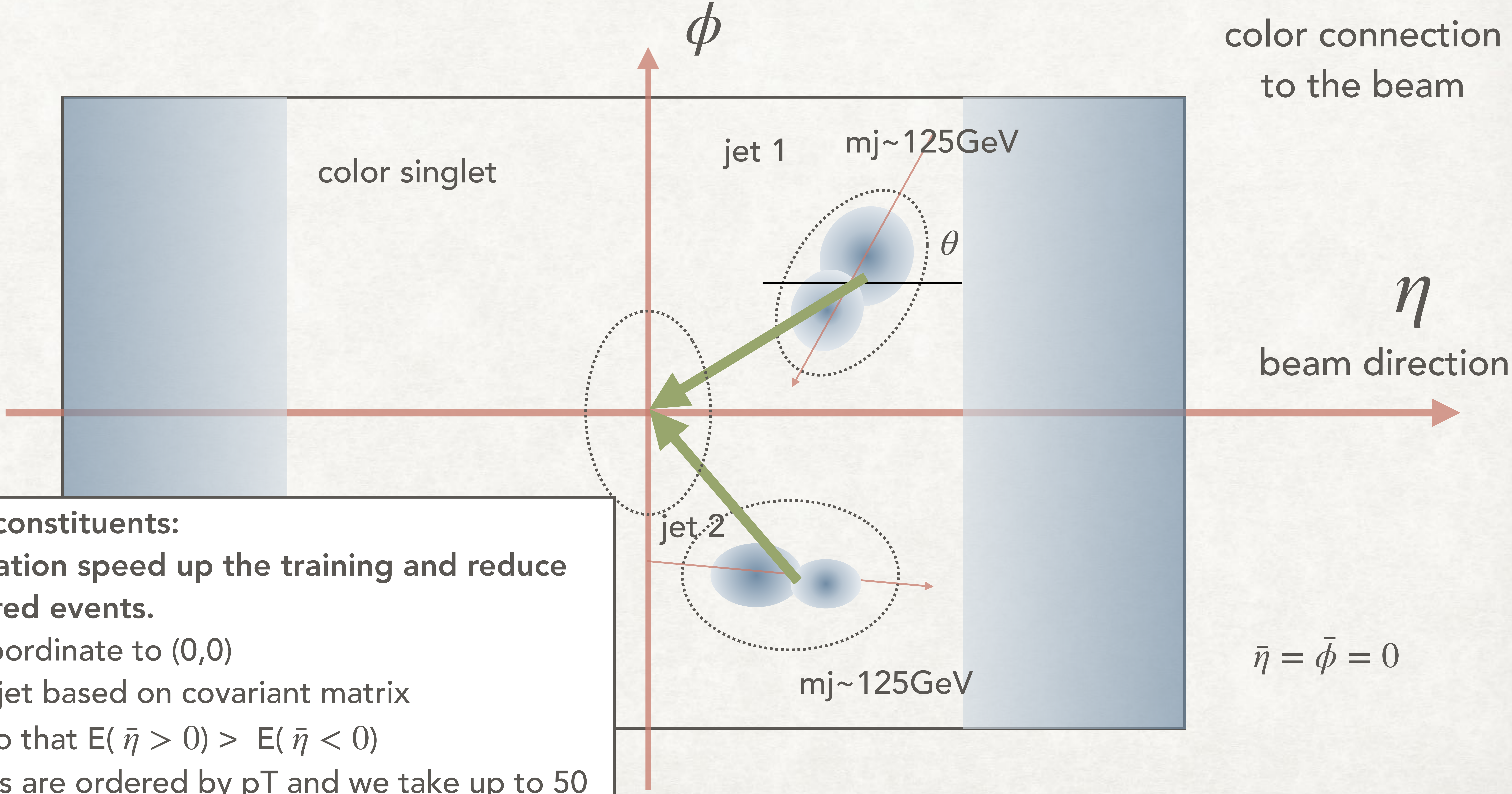
fatjet 1 = $(m_1, \eta_1, \phi_1, p_{T1}, E_1), \theta_1$

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 : "5 inputs for 4 momentum" , H candidate momentum as sum of two fat jets, add **θ**,

INPUT TO NETWORK: JET SUBSTRUCTURE INFO AS PARTICLE CLOUD



up to 50 constituents:
Regularization speed up the training and reduce the required events.

1. shift coordinate to (0,0)
2. rotate jet based on covariant matrix
3. flip η so that E($\bar{\eta} > 0$) > E($\bar{\eta} < 0$)
4. particles are ordered by pT and we take up to 50

$p_i = (\bar{\eta}_i, \bar{\phi}_i, p_{Ti}, \log p_{Ti})$ → (50, 4) data

HOW TO COMBINE JET STRUCTURE AND EVENT KINEMATICS

Naive approach "simple concatenation"

2311.16674[hep-ph] K. Ban, KC Kong, M Park,S.C. Park

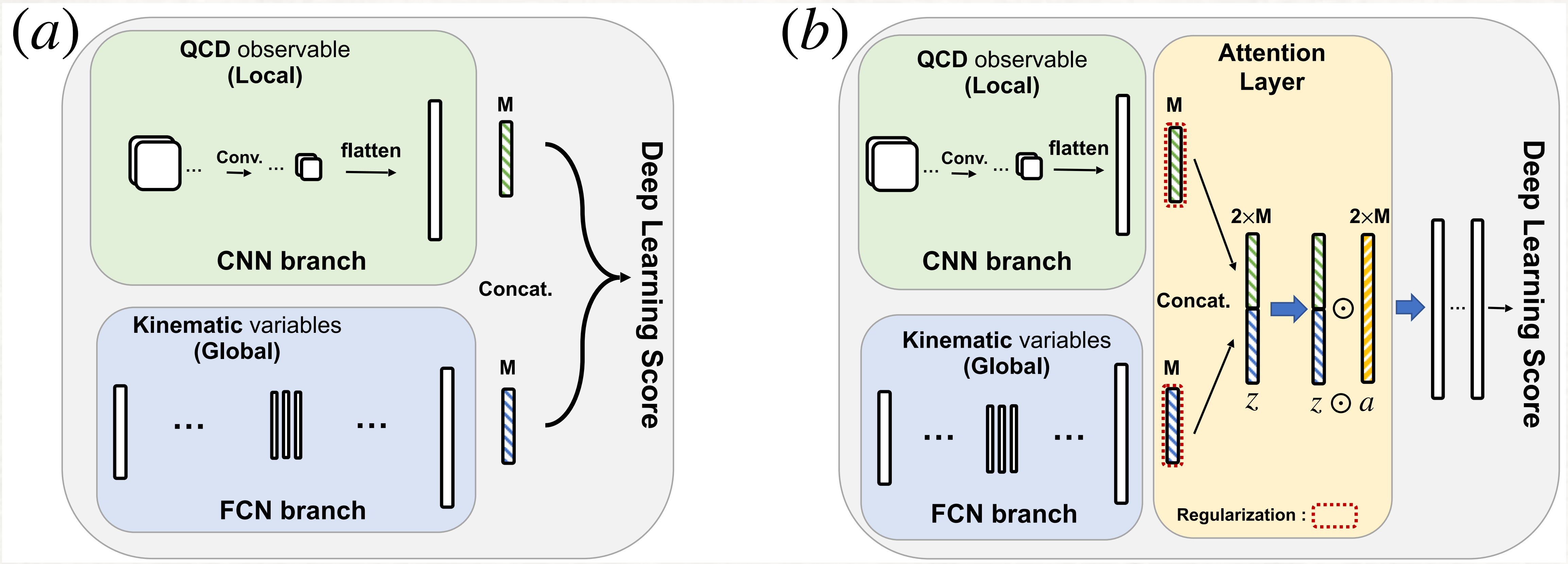
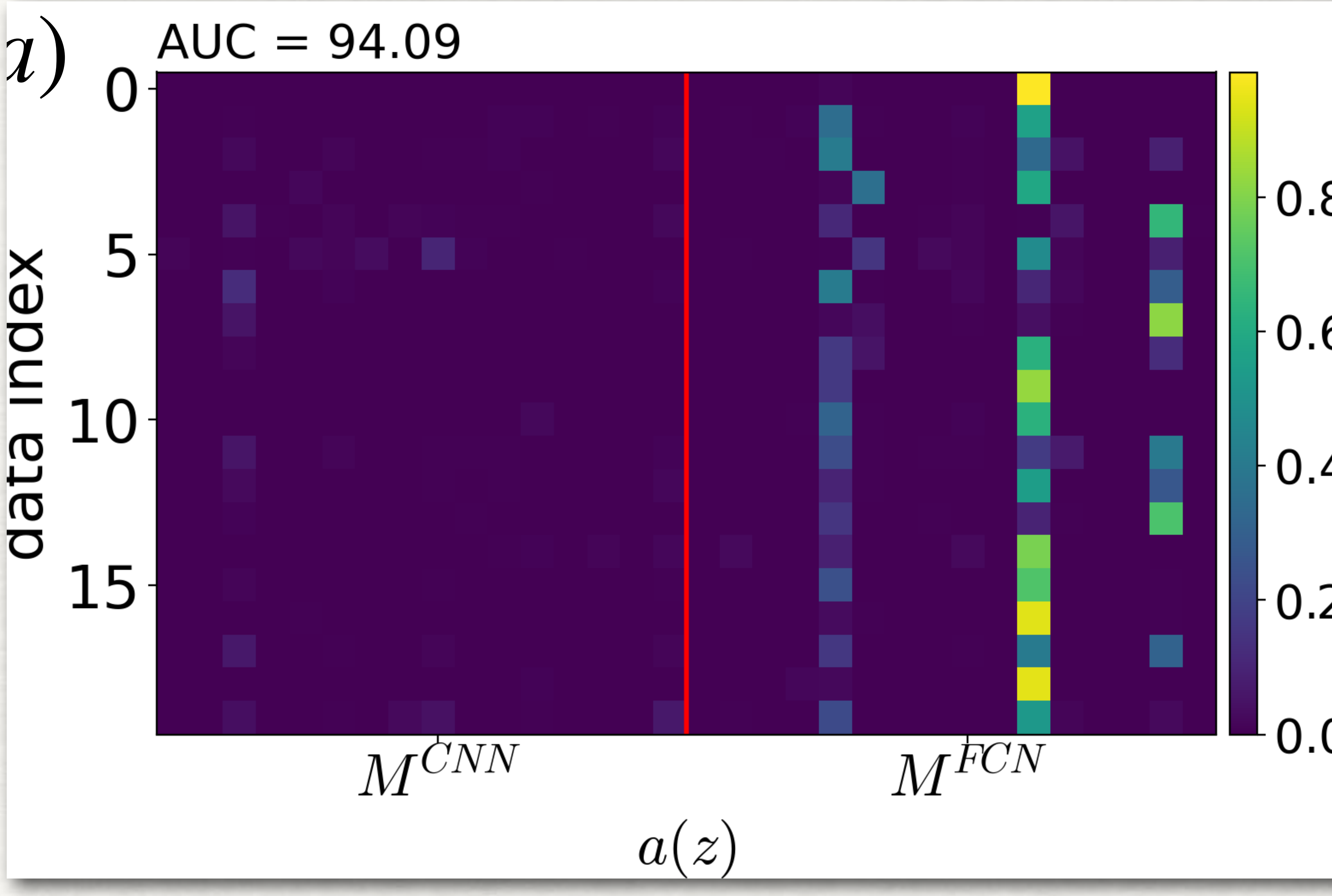


FIG. 2. The schematic plots for neural network structures: (a) conventionally used one in previous studies only with concatenation and (b) our proposed one with a regularized attention mechanism.



a) [Jet momentum (parton momentum)]+[jet concatenation] does not work.

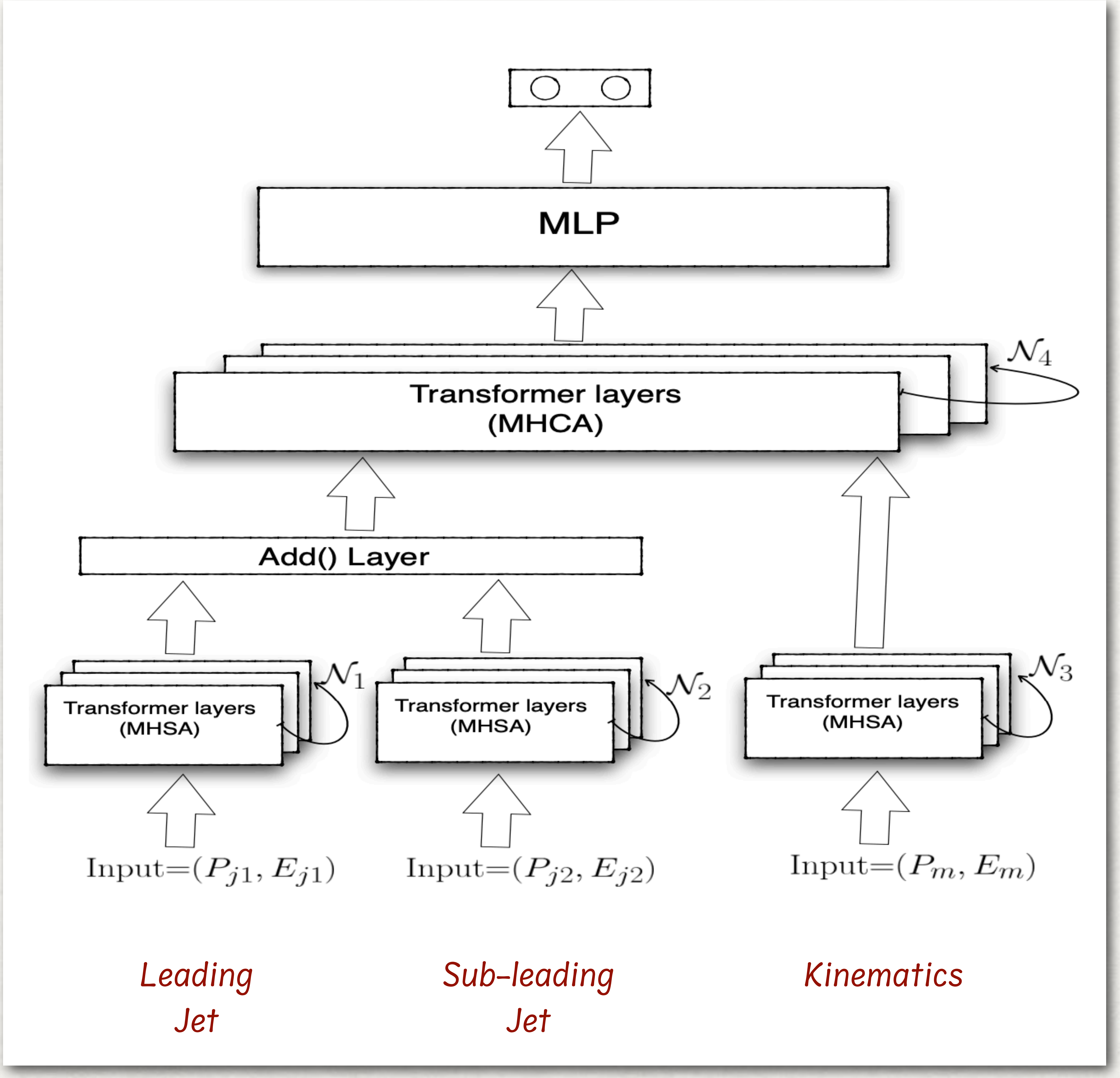
because of imbalance of "importance" of two information → the minor one can be ignored in the training.

Pre-training and freeze substructure analysis? We would loose the correlation to global kinematics.

OUR CROSS ATTENTION MODEL

multihead
cross attention layers

multihead
self attention layers



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]

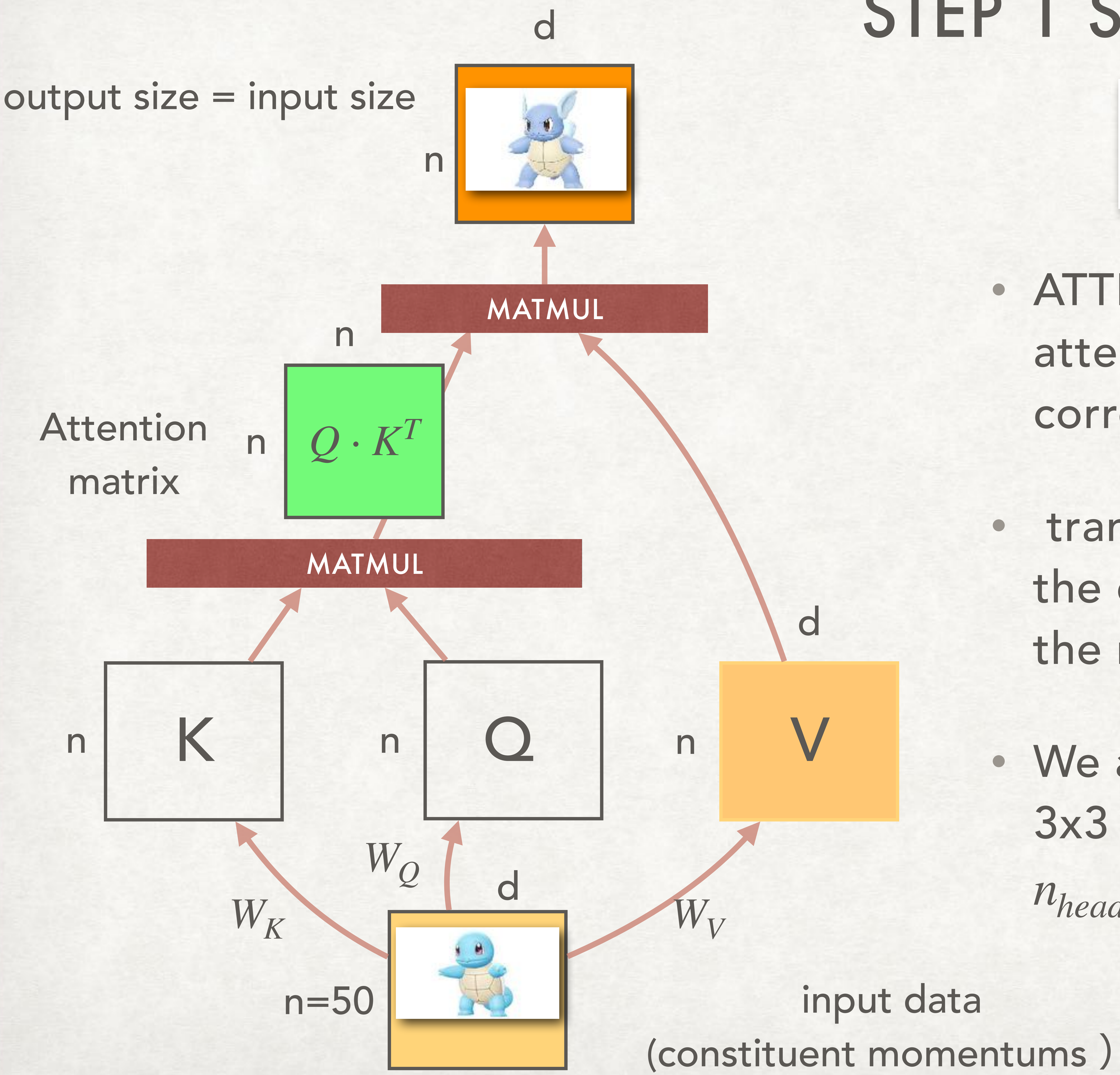
TAKEAWAYS

- **use "cross attention"** when you combine the "high scale information" to the "low energy scale", because cross attention layer gives extra emphasis to the information linked to the high energy kinematics.

- **skip connection and Interpretation** : Skip connection helps to maintain some connection to the inputs

- **More Physics:** Heavy particles decay into colored particles (discovery, spin, color structure?) Cross attention network probably more useful to resolve correlation of jet structures.

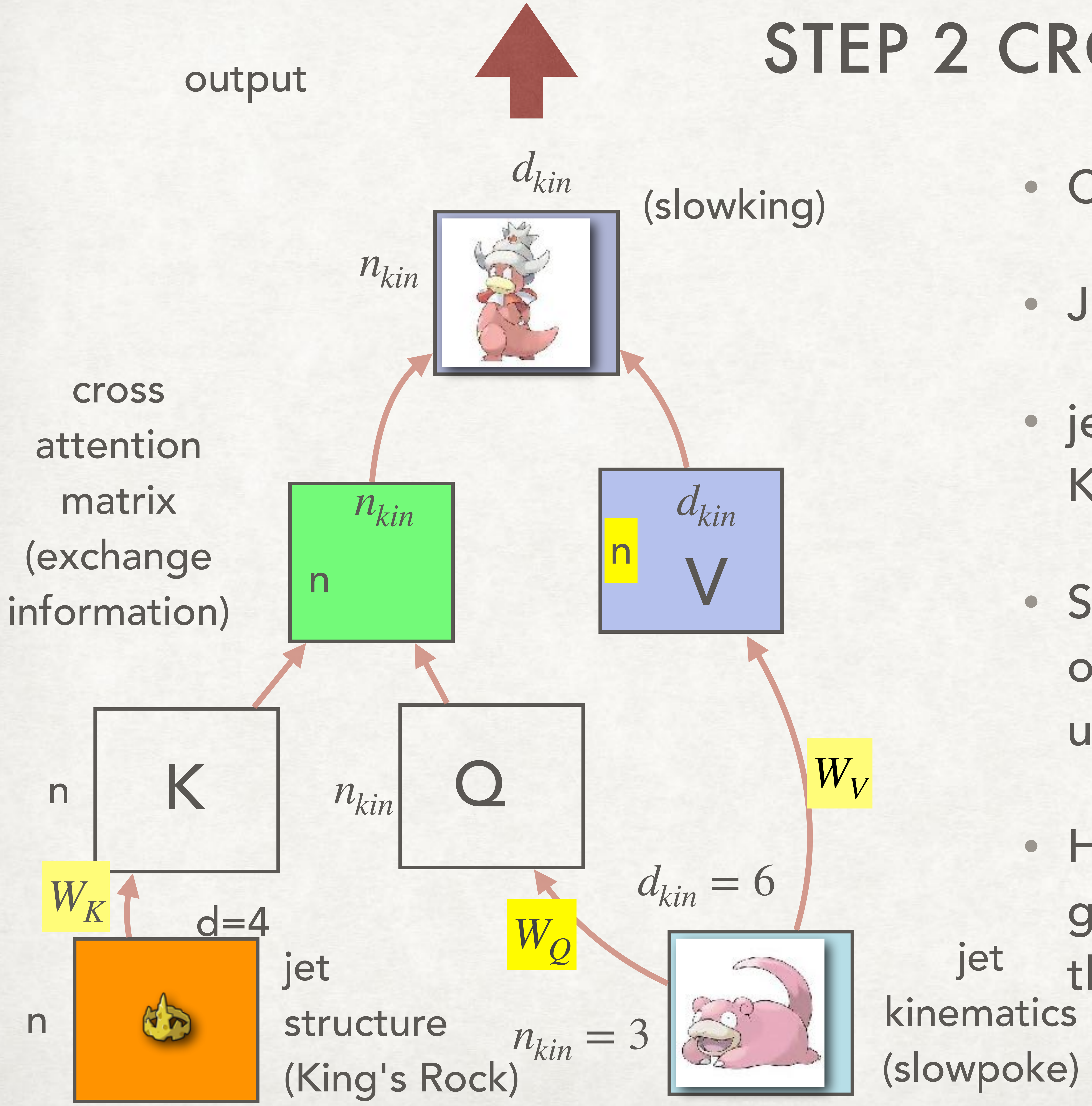
STEP 1 SELF ATTENTION LAYERS



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

- ATTENTION Matrix mix all features. Higher attention elements indicates important correlations
- transformation V → V' does not change the dimension. Structure of V retained for the next transformation.
- We adopt 50x50 self attention for jet and 3x3 self attention for kinematics, with $n_{head} = 5$

STEP 2 CROSS ATTENTION LAYERS

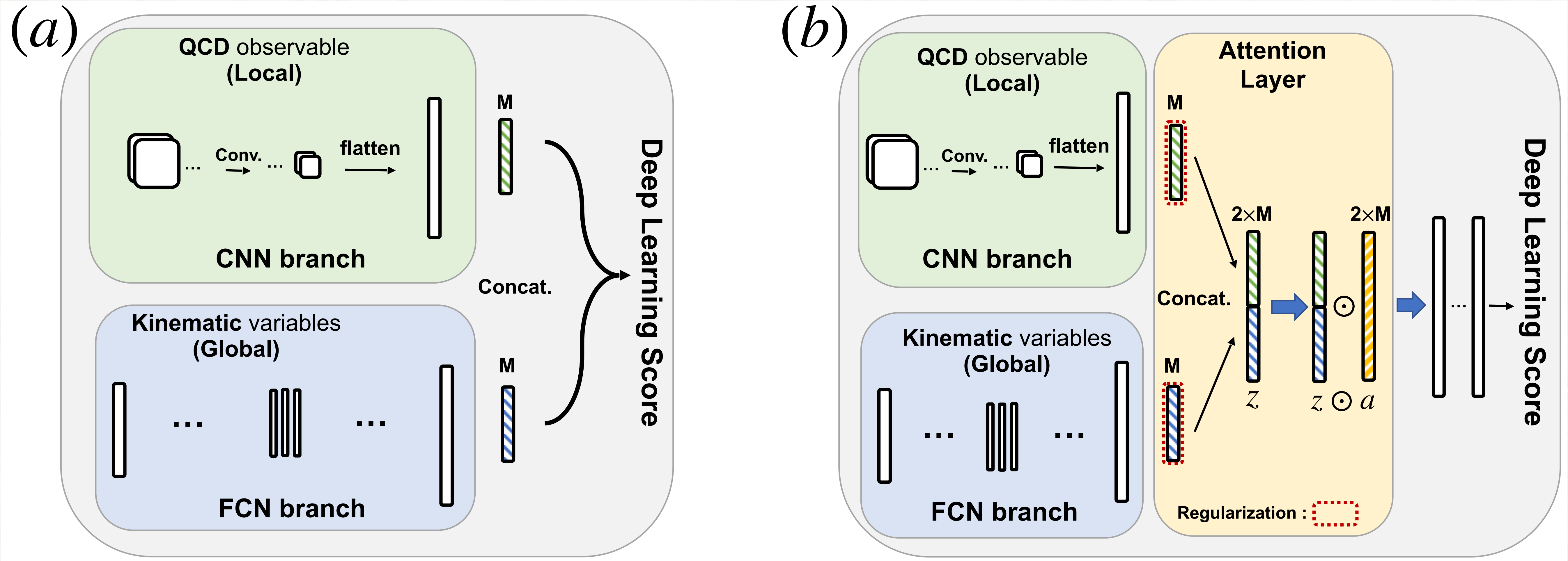


- Choose cross attention (jet kin) x (jet str.)
- Jet momentum : hard physics of partons Q , V
- jet substructure: parton shower, hadronization K
- Substructure output K and Jet kinematics output Q make attention matrix. The pairs update V (jet Kin)
- High scale feature relevant for classification gives extra weight to the corresponding jets though backward propagation

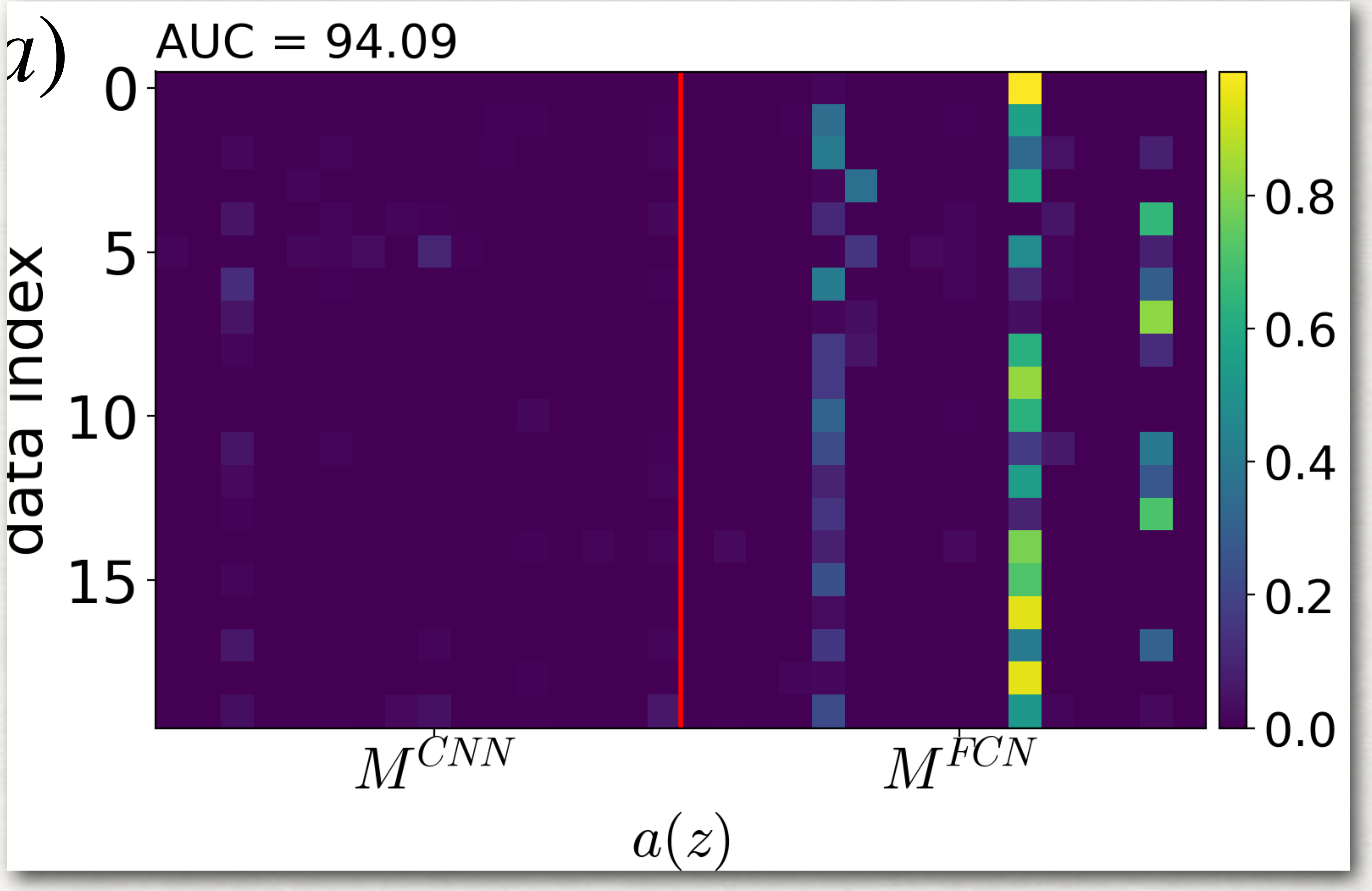
COMPARISON WITH OTHER APPROACH

Naive approach "simple concatenation"

2311.16674[hep-ph] K. Ban, KC Kong, M Park,S.C. Park



G. 2. The schematic plots for neural network structures: (a) conventionally used one in previous studies only with concatenation and (b) our posed one with a regularized attention mechanism.



our network kill this term
and keep off diagonal part only

(b) self attention matrix of combined information

$$AV = \begin{pmatrix} Q(Sub) \times K(Sub) & Q(Kin \times K(Sub) \\ Q(Sub) \times K(Kin) & Q(Kin)\,K(Kin) \end{pmatrix} V = \boxed{Q(kin)\,K(kin)\,V(kin)} + \ldots$$

PHYSICS

- a jet:

P(hadrons in jets | parton or jet) = $P(\{x_i\}\,|\,y)$

- a fatjet or a jet with substructure

$$P(\{x_i\}\,|\,\{y_\alpha\})$$

- two fatjets in an event

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

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

cross attention

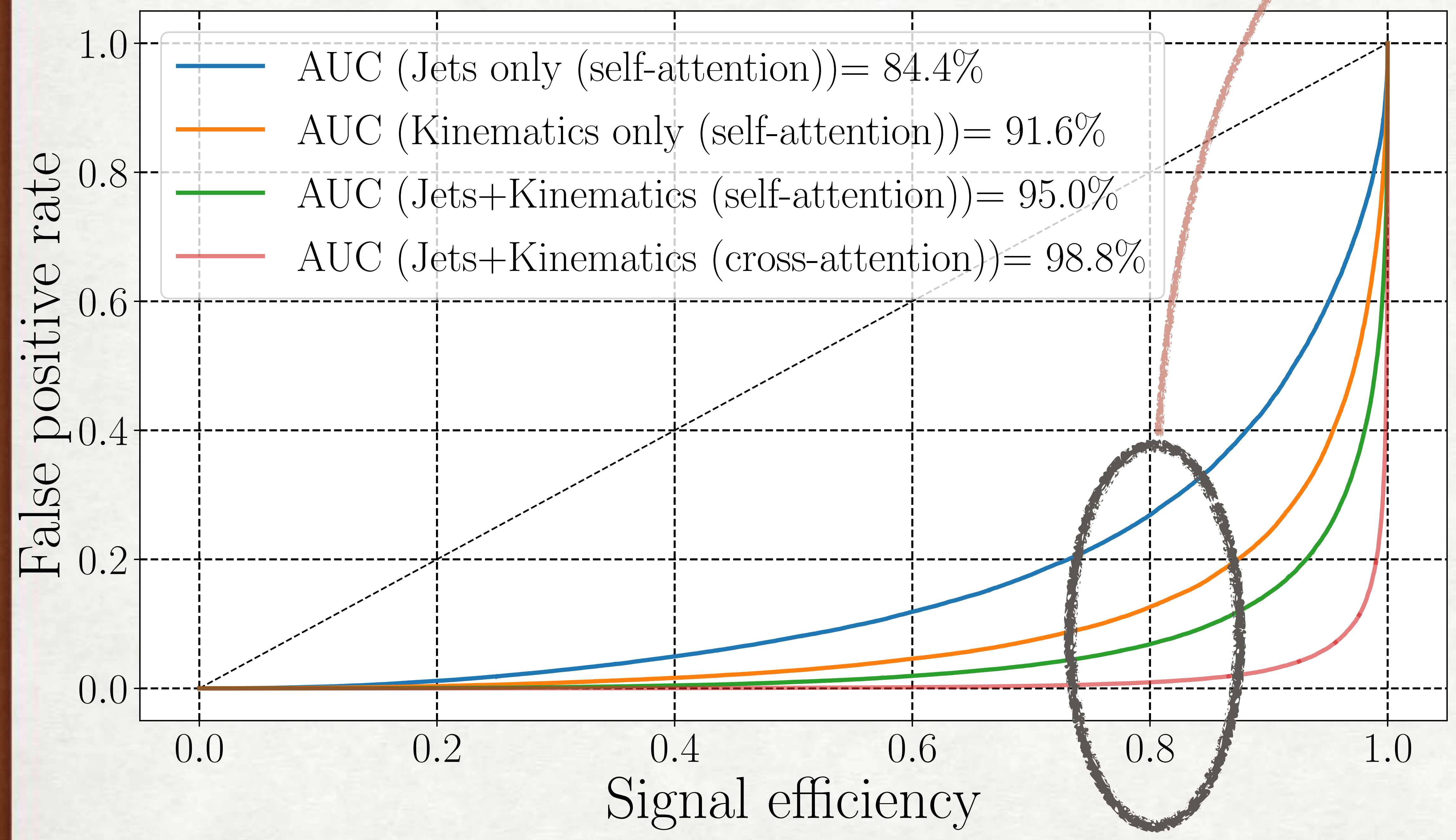
jet kinematics

IMPROVEMENT USING CROSS ATTENTION

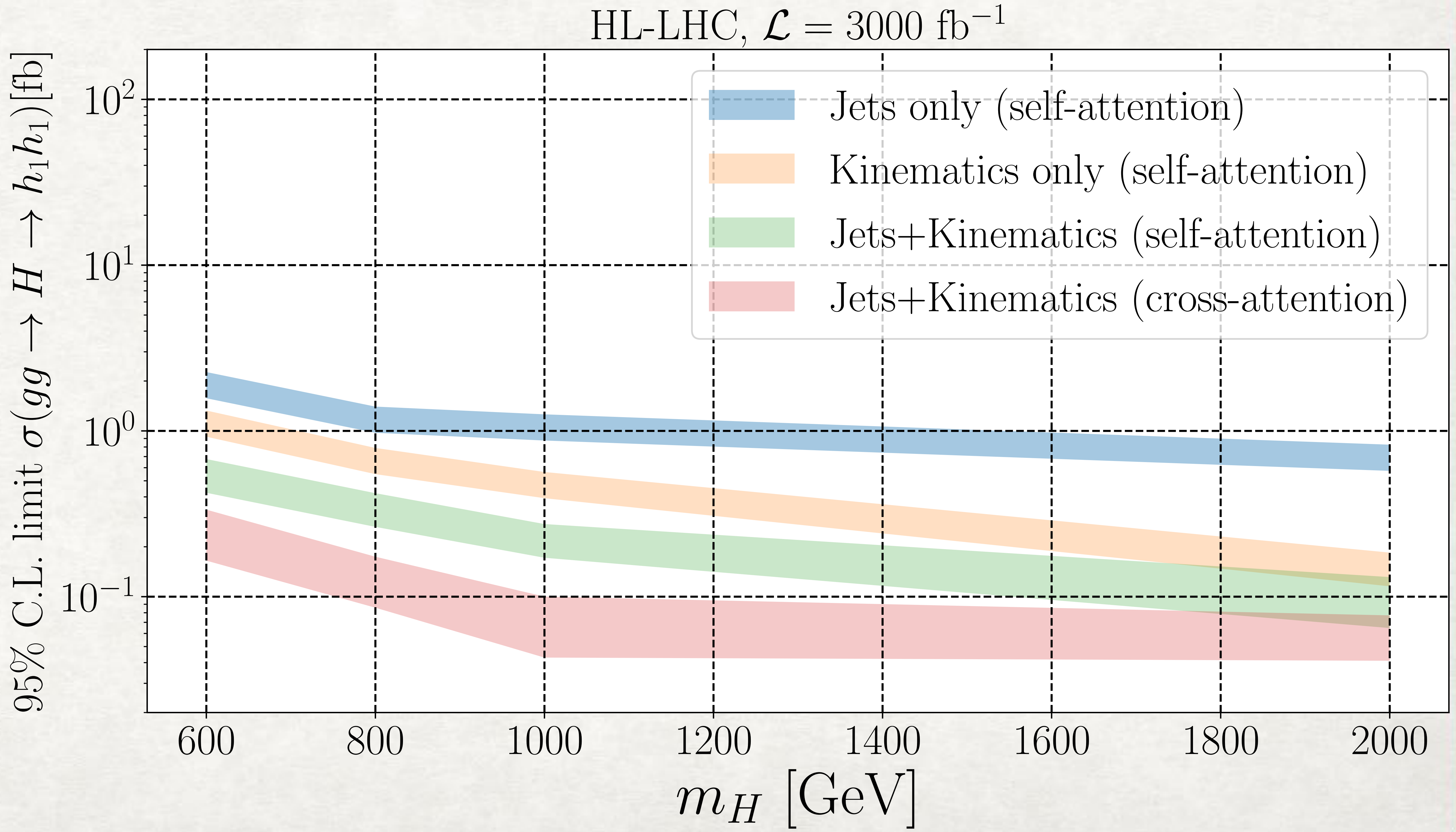
green: self attention of Jet str. and Kin
→ concatenate and MLP
red line : cross attention

factor 5 improvement at the same acceptance.

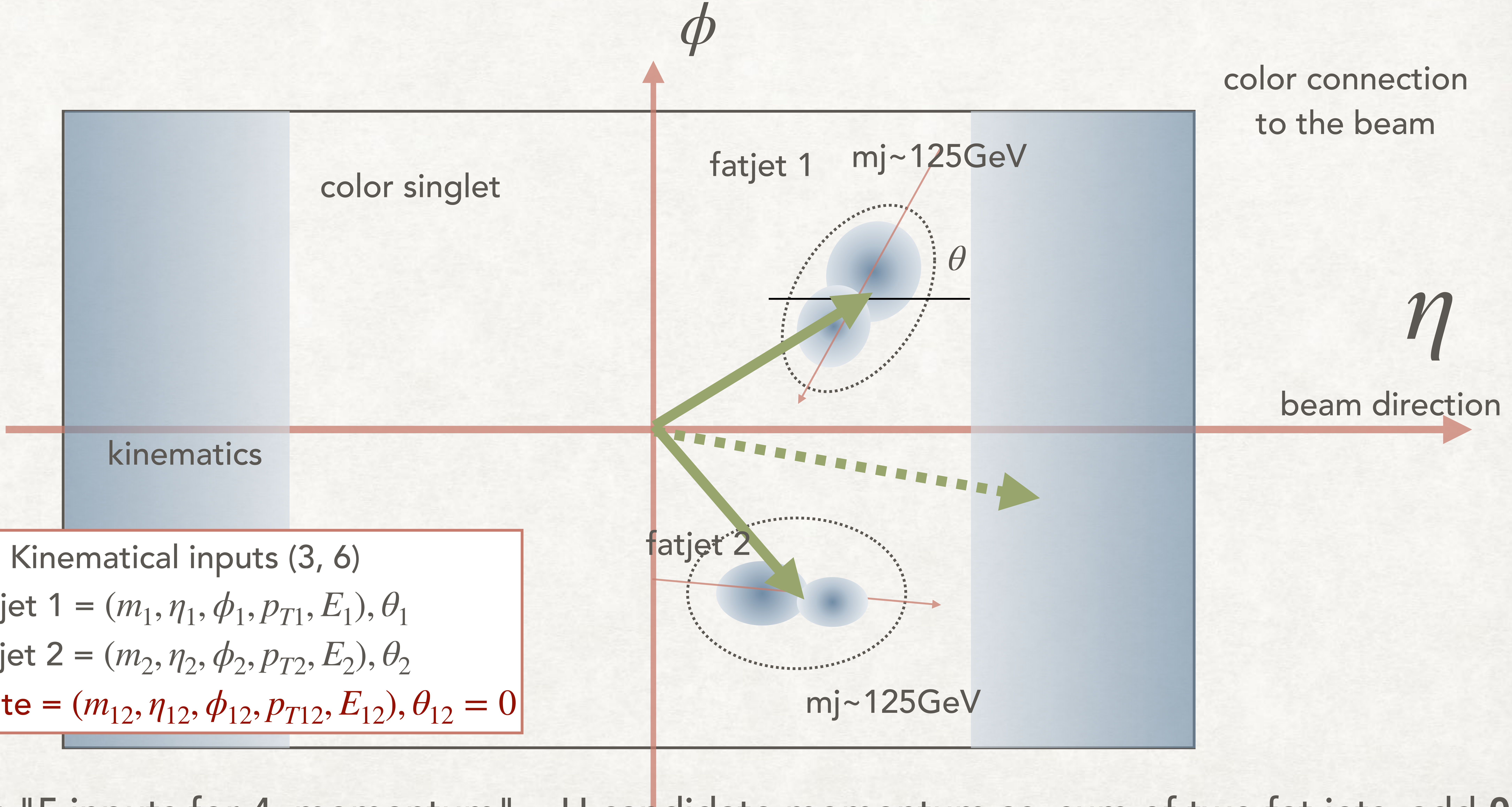
Simple estimation of the upper limits



Cross attention improve the rejection efficiency significantly



INPUT TO DL : 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 = $(m_{12}, \eta_{12}, \phi_{12}, p_{T12}, E_{12}), \theta_{12} = 0$

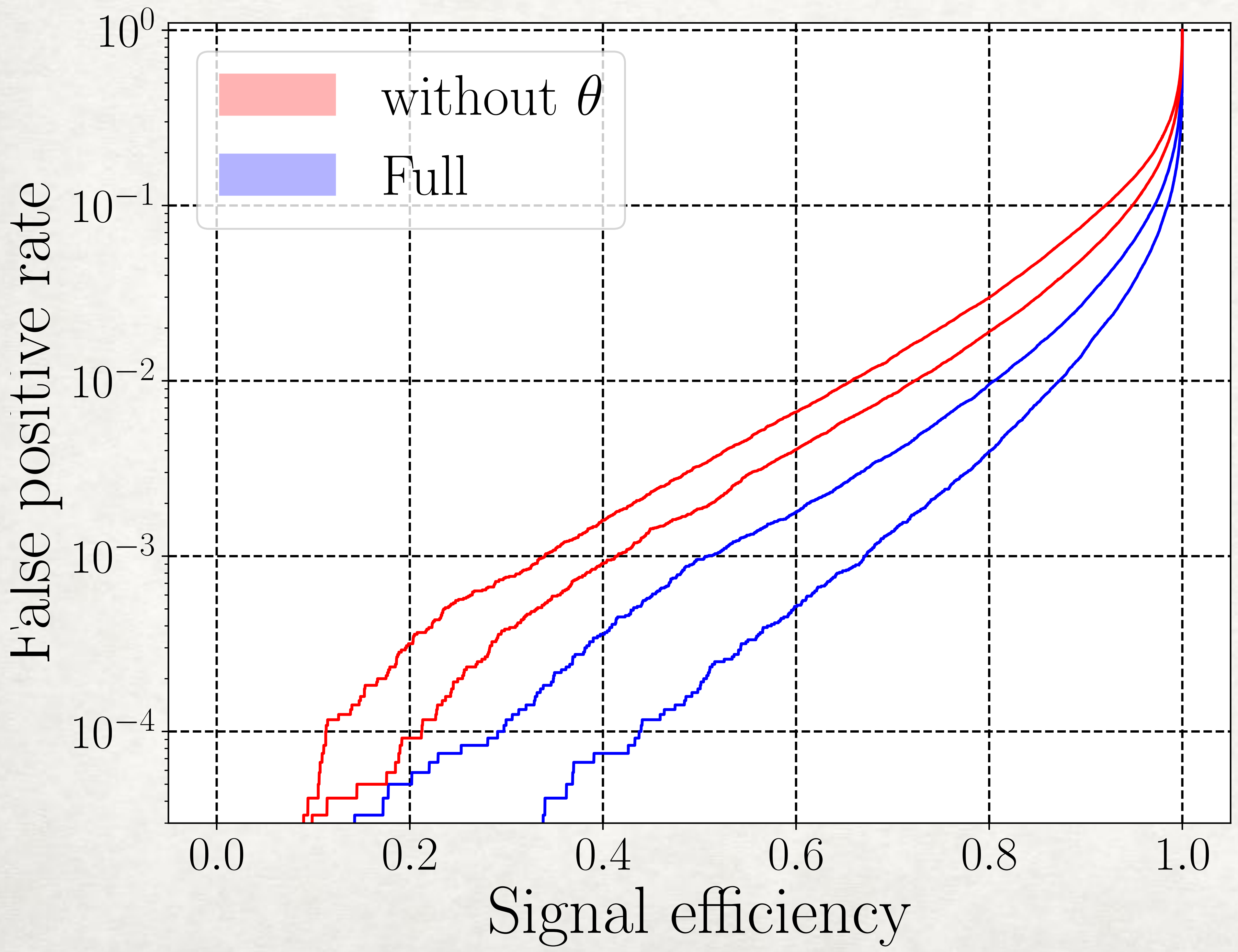
NOTE : "5 inputs for 4 momentum" , H candidate momentum as sum of two fat jets, add **θ**,

ROLE OF θ

minor

large improvement

	Kinematics	Kin +θ	jet str.+kin	jet str +Kin + θ
ROC	91.01%	91.6	97.23-98.16	98.68-99.28

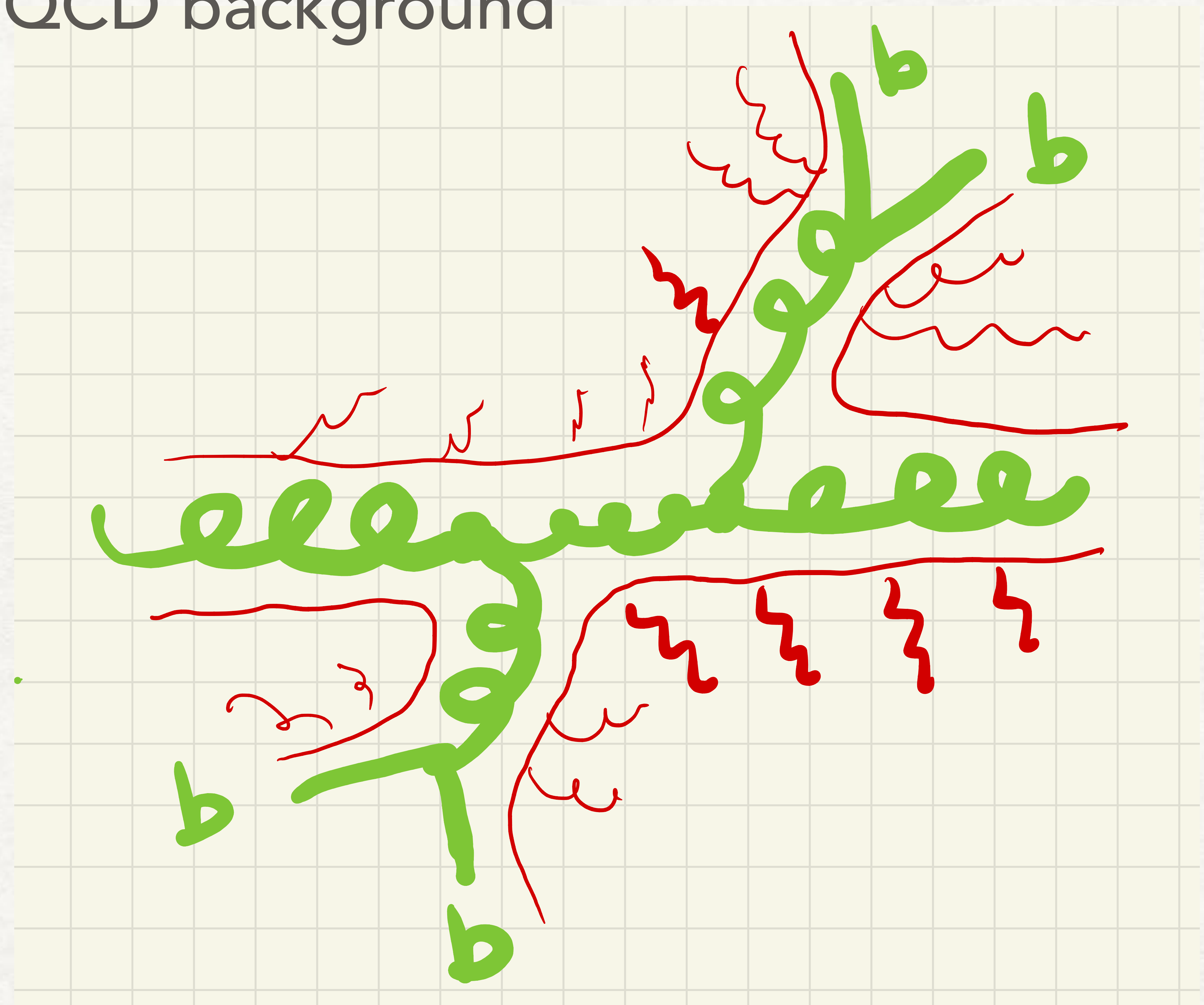


adding rotation angle θ improve classification when both jet str. and kinematical information available.

We are working on to idententify the origin. (color connection? momentum resolution?)

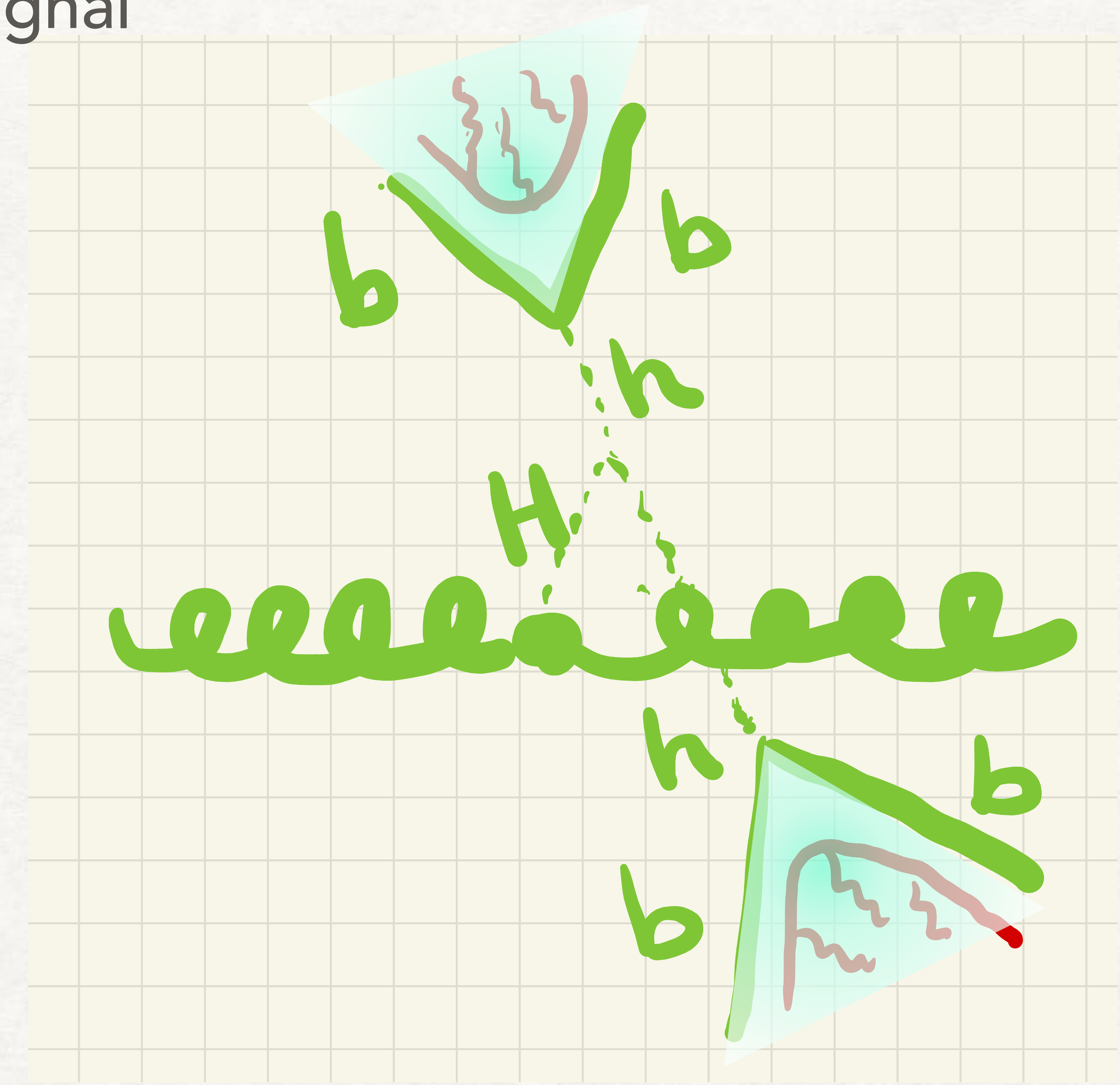
event color structure

QCD background



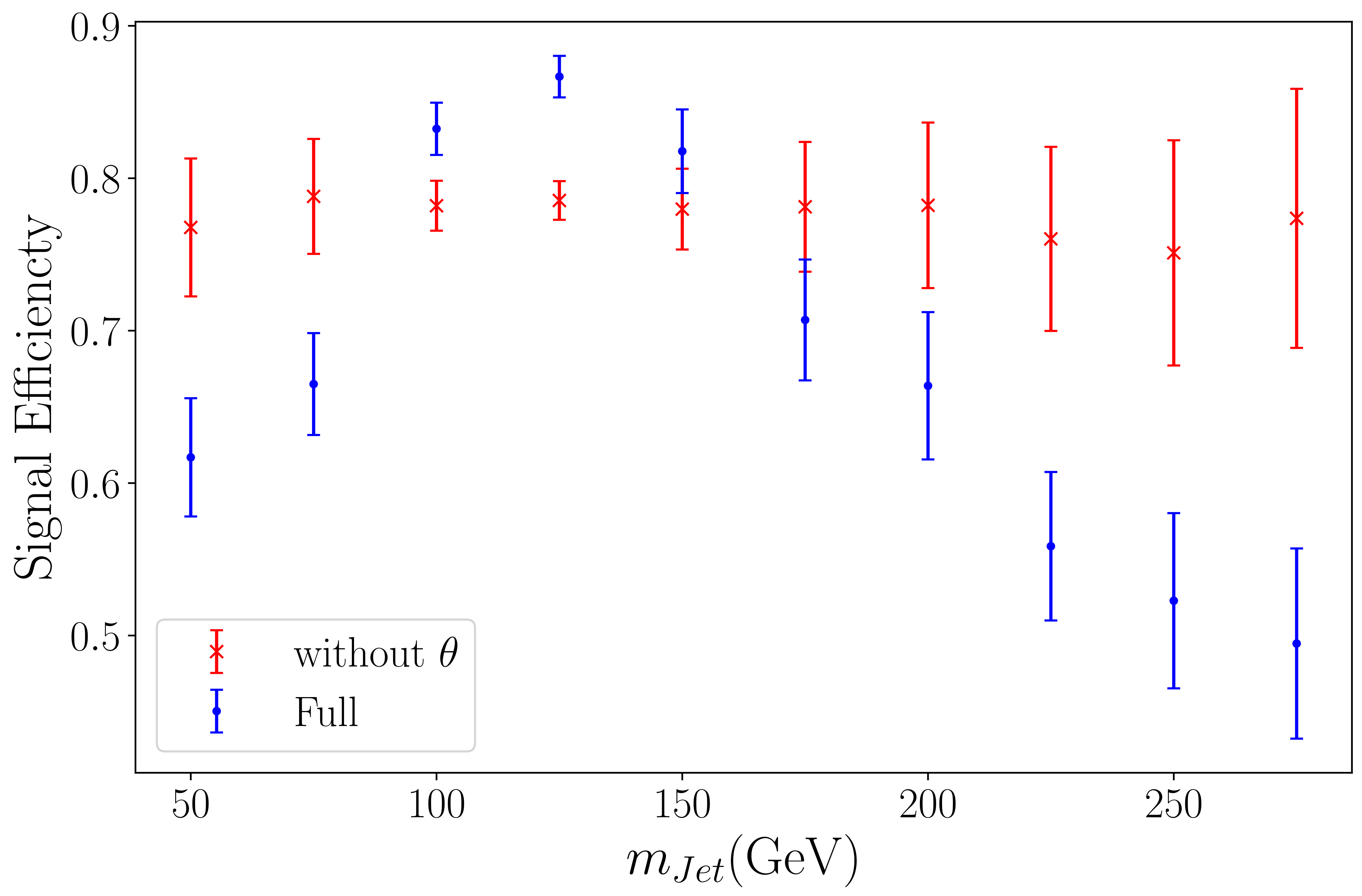
For QCD and top event, fatjets are likely color connected to the other activities of the event

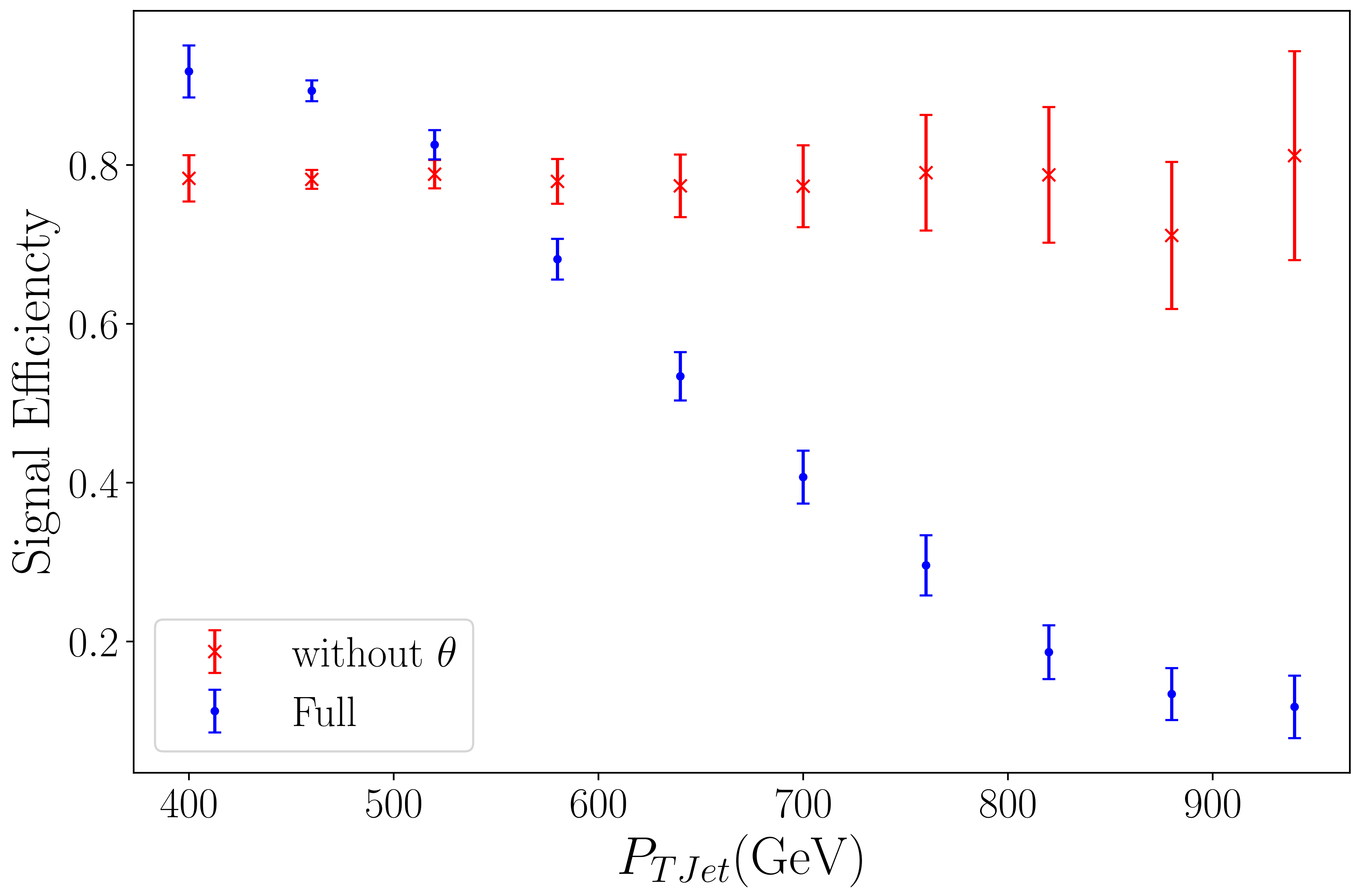
signal



Higgs bosons are color isolated.

SOME SIGNAL SELECTION EFFICIENCY

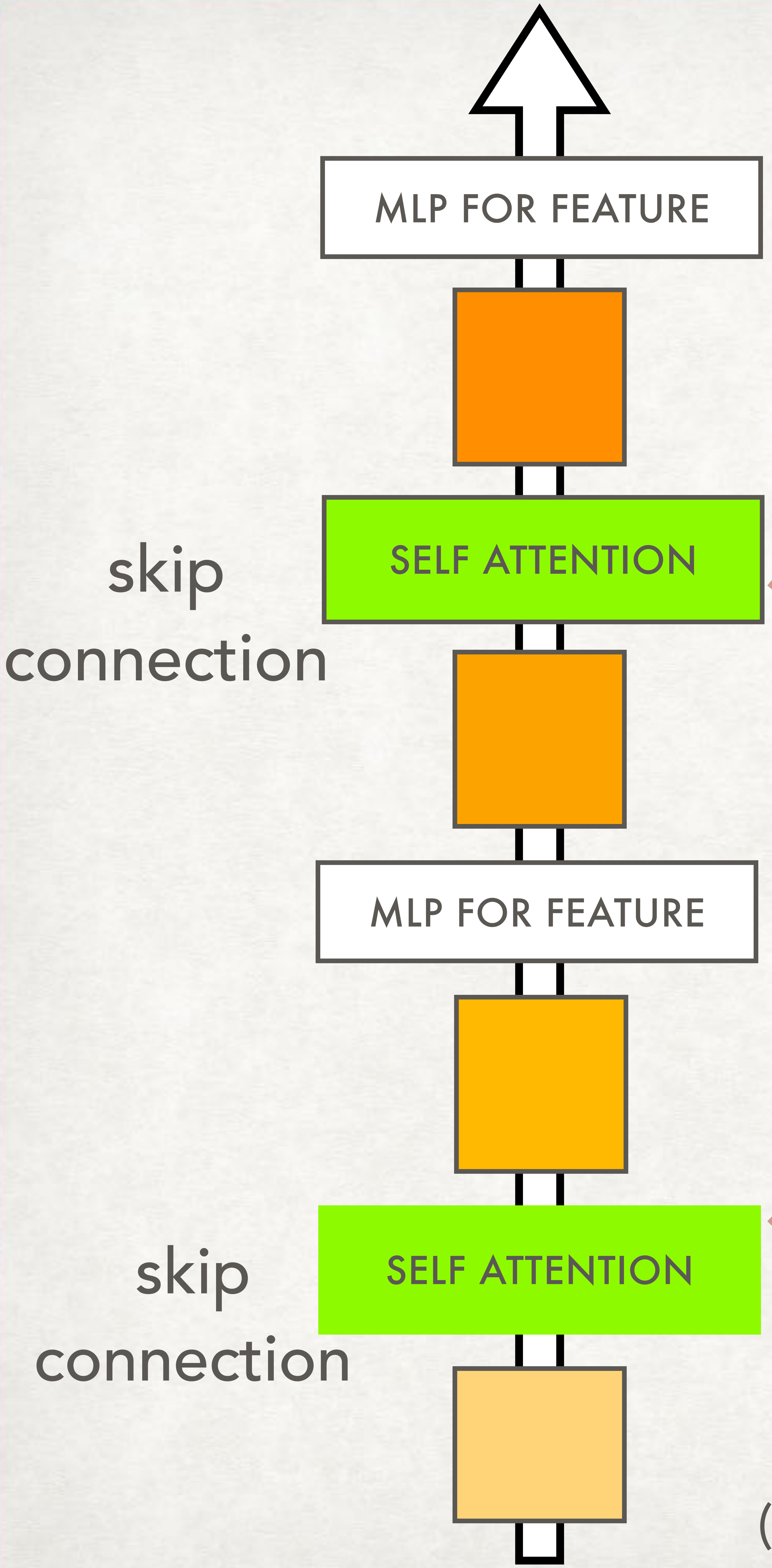


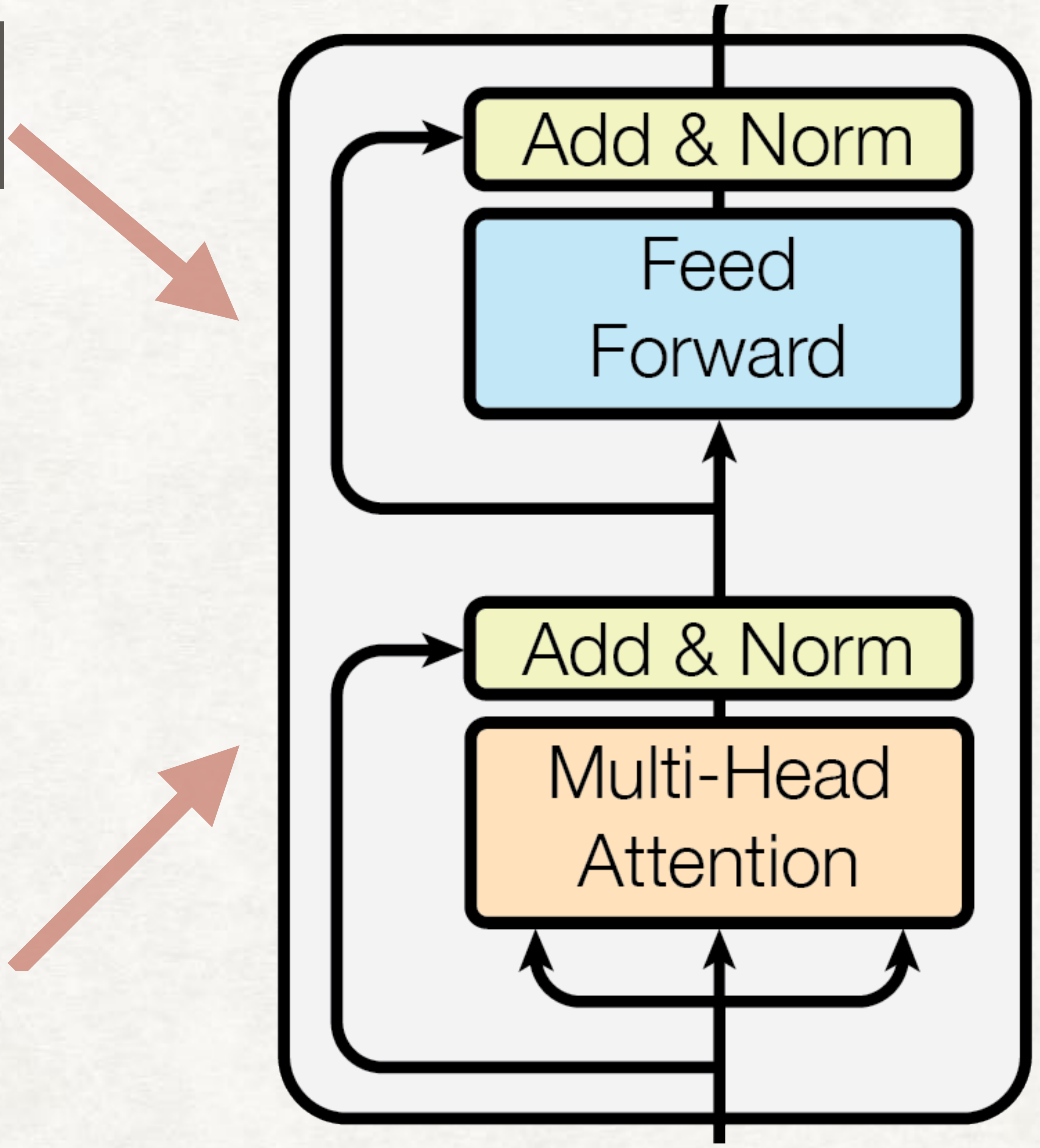


better selection of Higgs mass

rejecting high PT events

Interpretation and Skip Connection





input data
(particle x feature)

- Deep Learning suffers low interpretability and it is always annoying.
- skip connection of attention blocks helps connecting input data to extracted feature(transformed quantity) in **some level.**

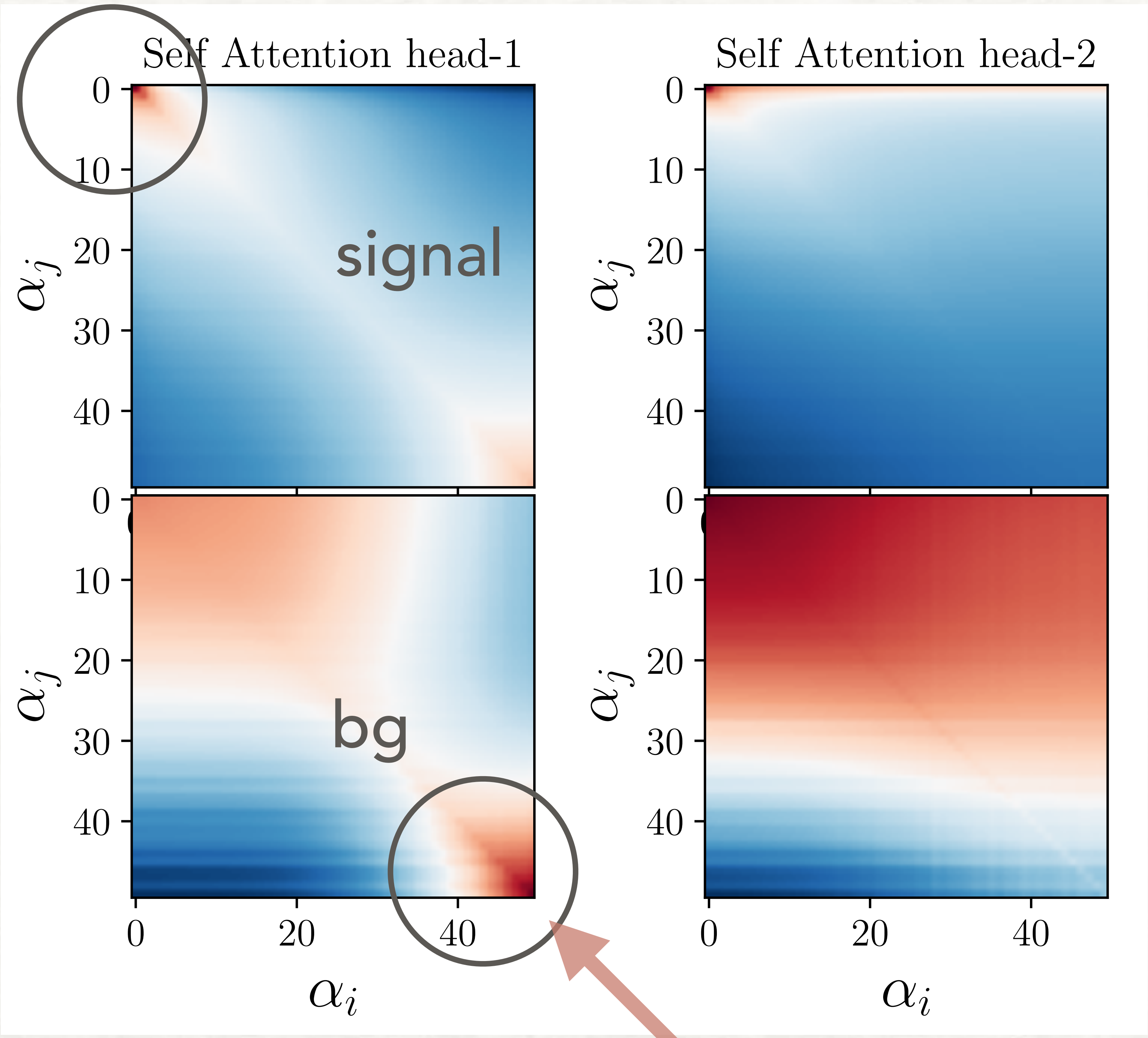
1706.03762 Vaswani et all “Attention is all you need

EX. SELF AND CROSS-ATTENTION MAP

self attention map

axis: odering of modified particles

First few "particle" token express
Higgs nature efficiently



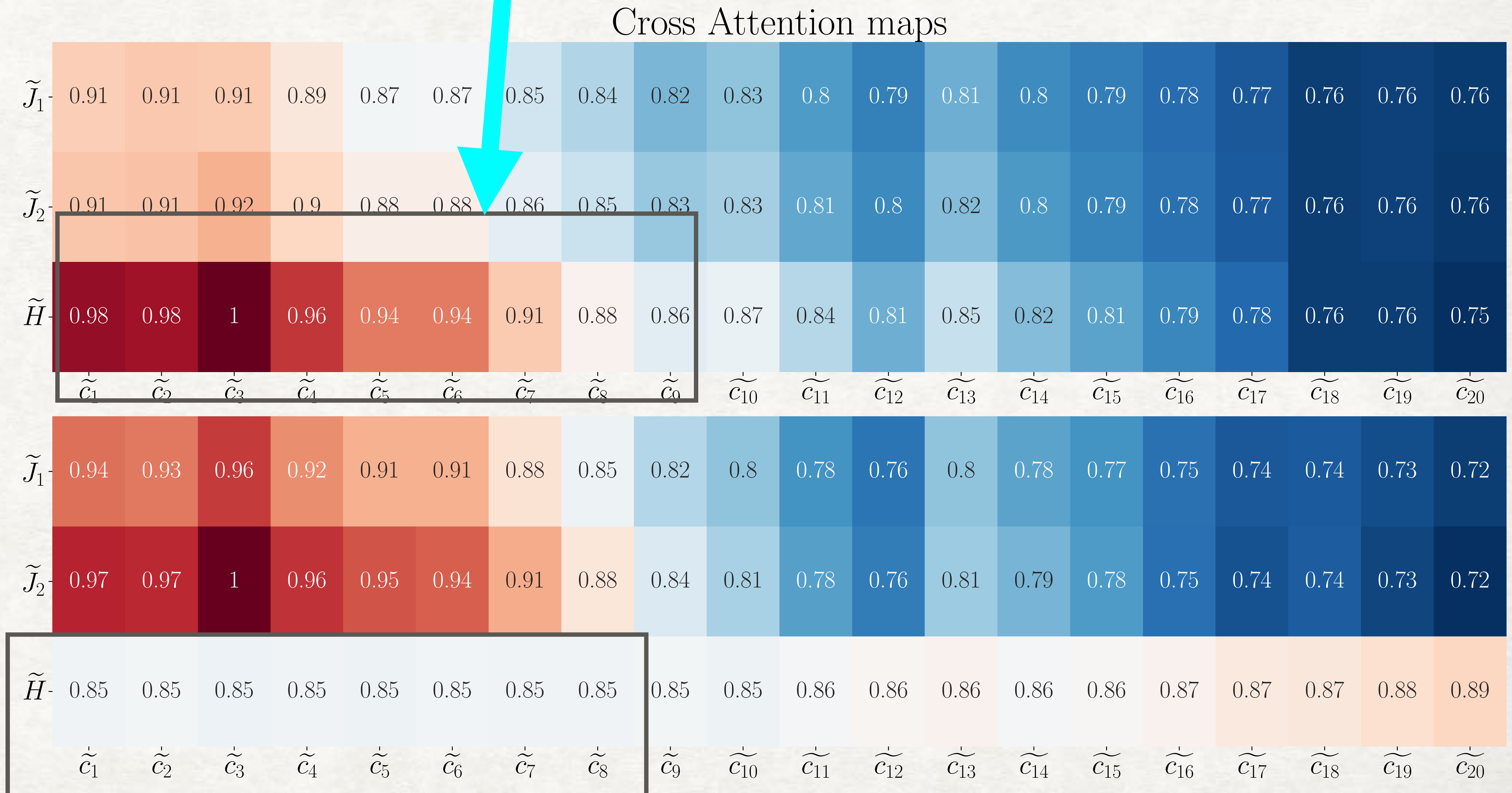
maybe number of averaged particles are different

Cross attention map:Particle in the jet (50)
and parent particle (3)

sum of fatjet momenta capture signal

Signal

BG

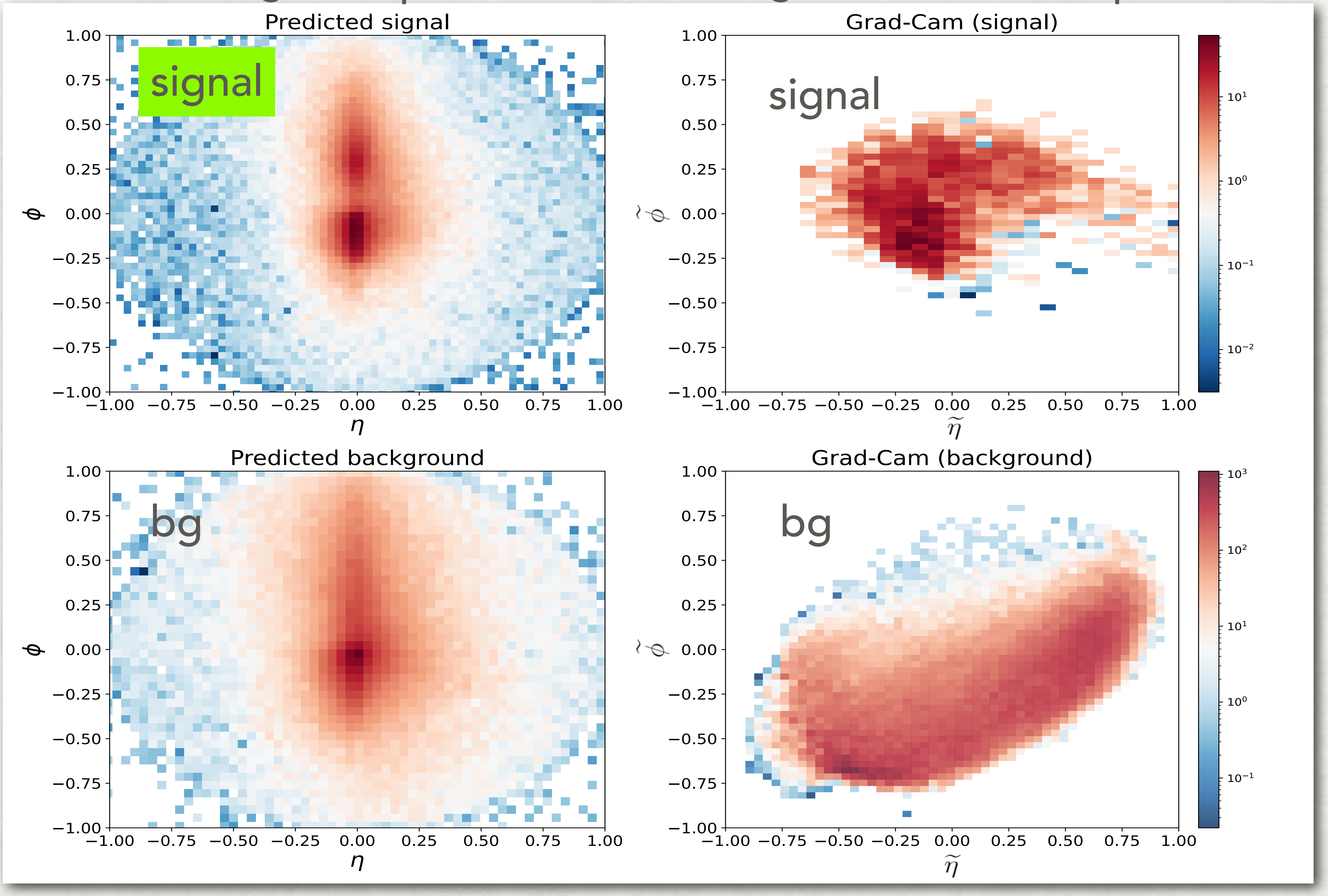


GRAD-CAM (1610.02391)

- Output of last attension layers (some correlation with original inputs)

5000 signal events

Original inputs grad-cam heatmap



Y : class scaore

F:output from last attention layer

$\tilde{\eta}, \tilde{\phi}$ transformed corrdinate

$$\alpha_k(\widetilde{\eta}, \widetilde{\phi}) = \frac{1}{Z} \sum \frac{\partial Y_c}{\partial F_k(\widetilde{\eta}, \widetilde{\phi}, \widetilde{p_T})}$$

$$\text{Grad-CAM}(\widetilde{\eta}, \widetilde{\phi}) = \frac{1}{k} \sum_k \alpha_k(\widetilde{\eta}, \widetilde{\phi}) F_k(\widetilde{\eta}, \widetilde{\phi}, \widetilde{p_T})$$

Still see some connection between particle location and transformed coordinates.

Inference vs Attention depth

TAKEAWAYS

- **use "cross attention"** when you combine the "high scale information" to the "low energy scale", because cross attention layer gives extra emphasis to the information linked to the high energy kinematics.

- **skip connection and Interpretation** : Skip connection helps to maintain some connection to the inputs

- **More Physics:** Heavy particles decay into colored particles (discovery, spin, color structure?) Cross attention network probably more useful to resolve correlation of jet structures.

- Result looks very good to me and I am still worrying about bugs...

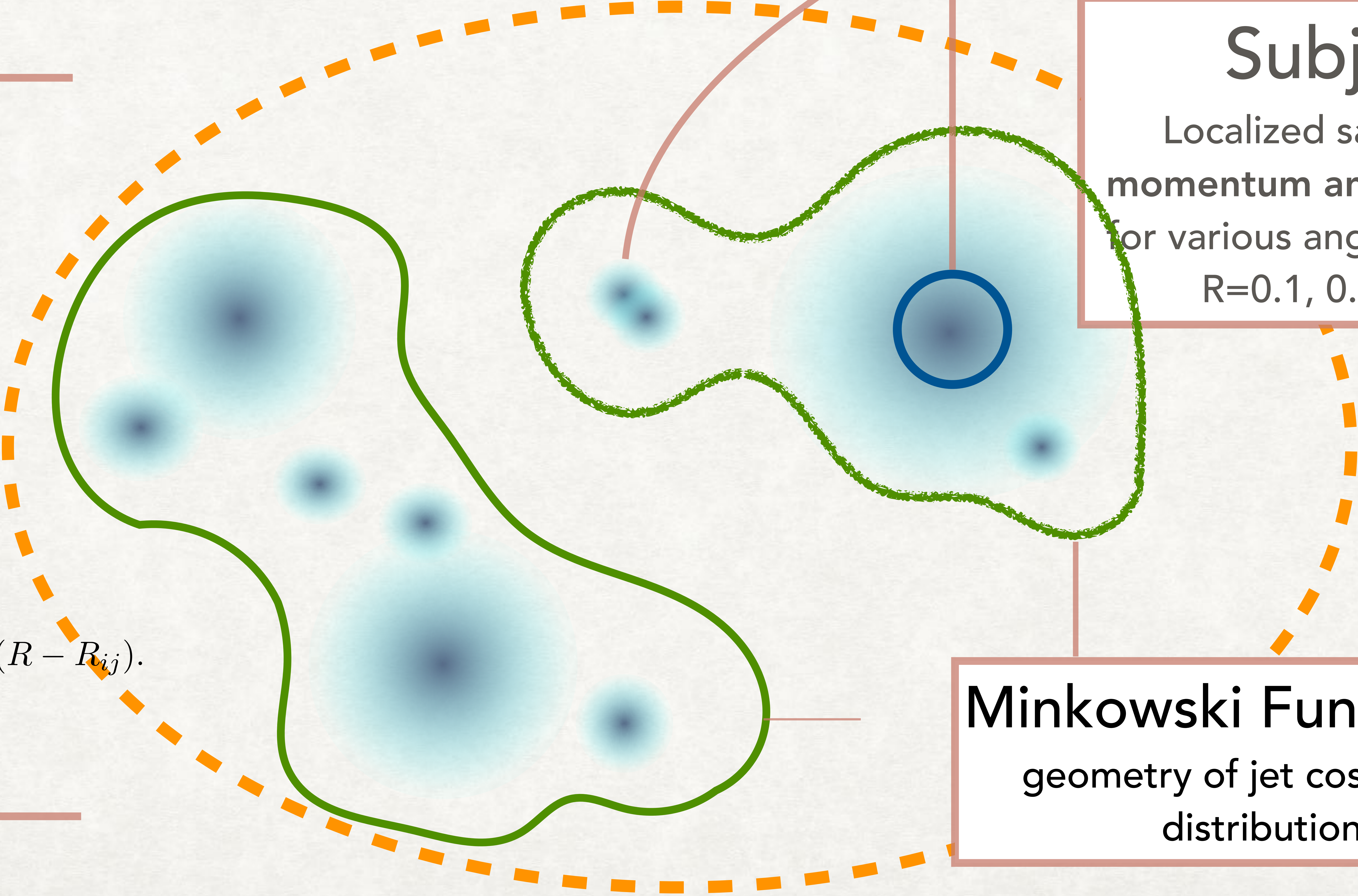
NEED TO BE IMPROVED

- Current GPU requirement: **2 x NVIDIA RTX A6000 (48GB) with 80% and 30% utilization in tensor flow mirror strategy.** 96% consumption /card 20min/ training.
- We definitely have to change "jet substructure part" to simpler one, keeping cross attention structure(this part is generic)
- Ex: "Modulated Network of HL variables"
  - QCD vs top, Amon Furuichi(Nagoya), Sung Hak Lim(Rutgers) and M. Nojiri arXiv 2312.11760[hep-ph] work as good as Particle Transformer.
- ...... but are they robust for color connection?

BACK UP SLIDES

JET High Level variables

pt distribution of constituents



Subjet

Localized sampling
momentum and counting
for various angular sccale
R=0.1, 0.2, 0.3

Jet spectrum

two point **Energy correlation**
(unlocalized sampling)

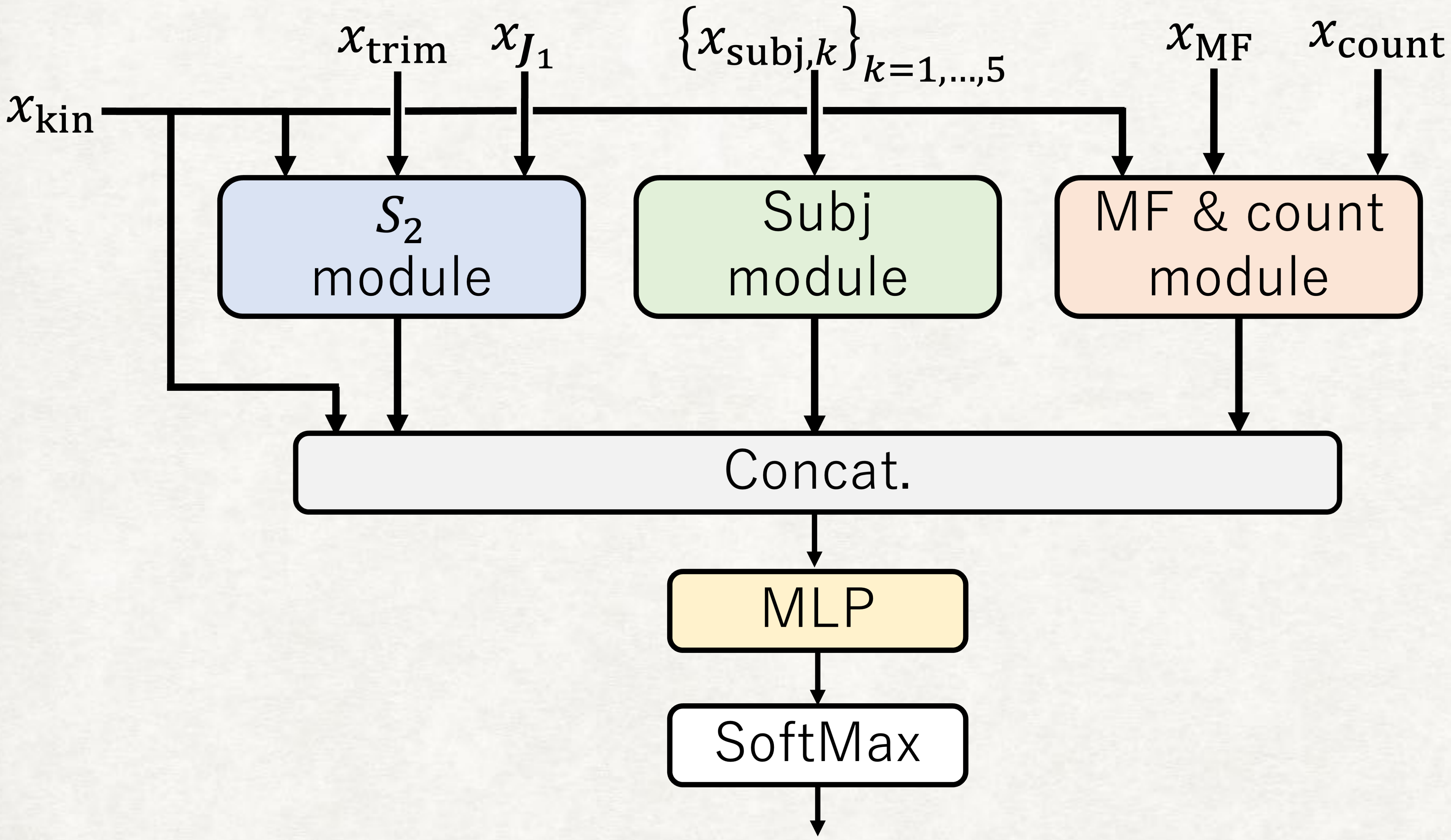
$$S_{2,ab}(R) \stackrel{\text{def}}{=} \sum_{i\in a}\sum_{j\in b} p_{T,i}p_{T,j}\delta(R-R_{ij}).$$

Minkowski Functionals

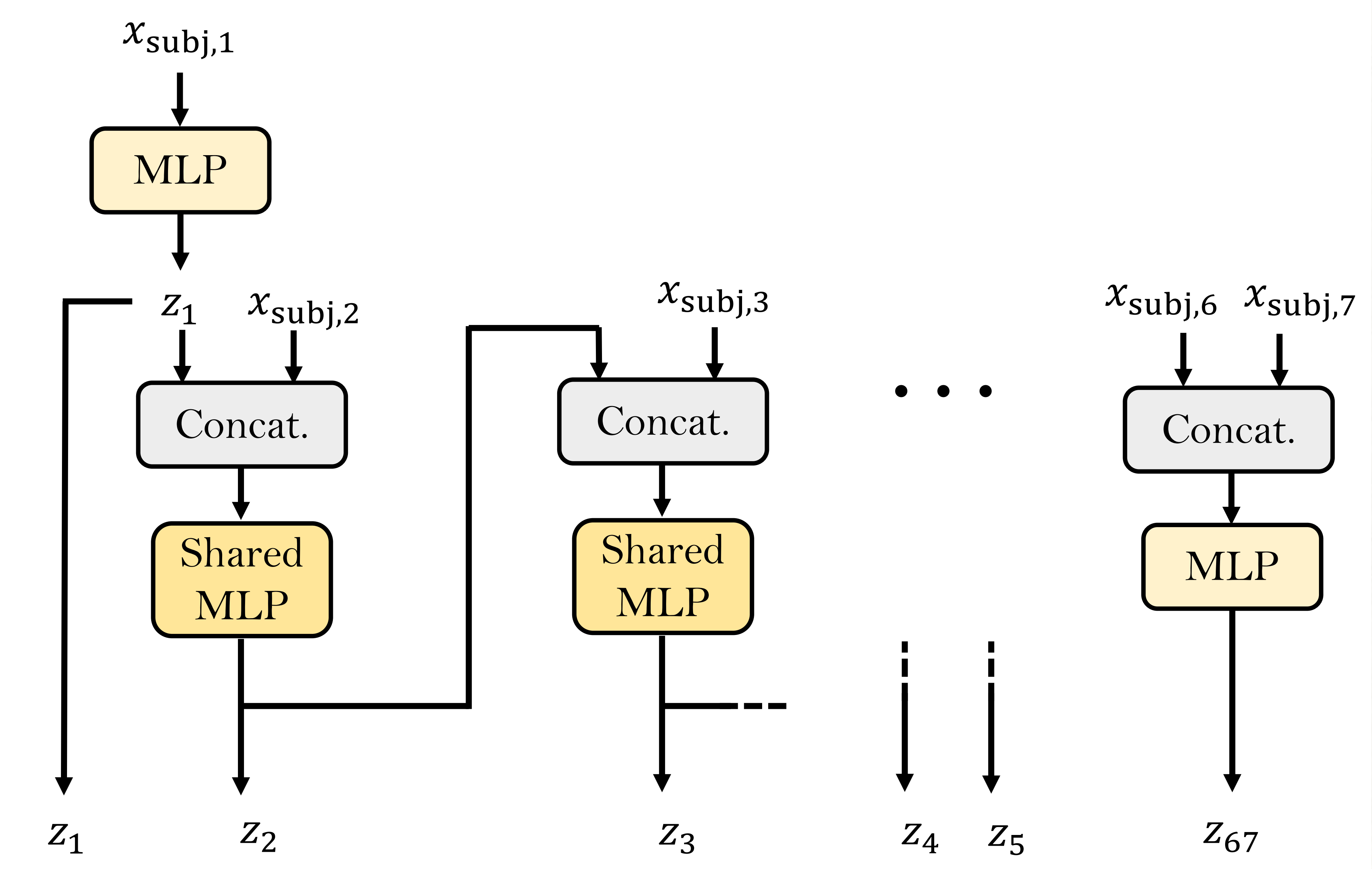
geometry of jet cosntituent distribution

NETWORK USING HL INPUTS (ANALYSIS MODEL=AM)

easy add and remove



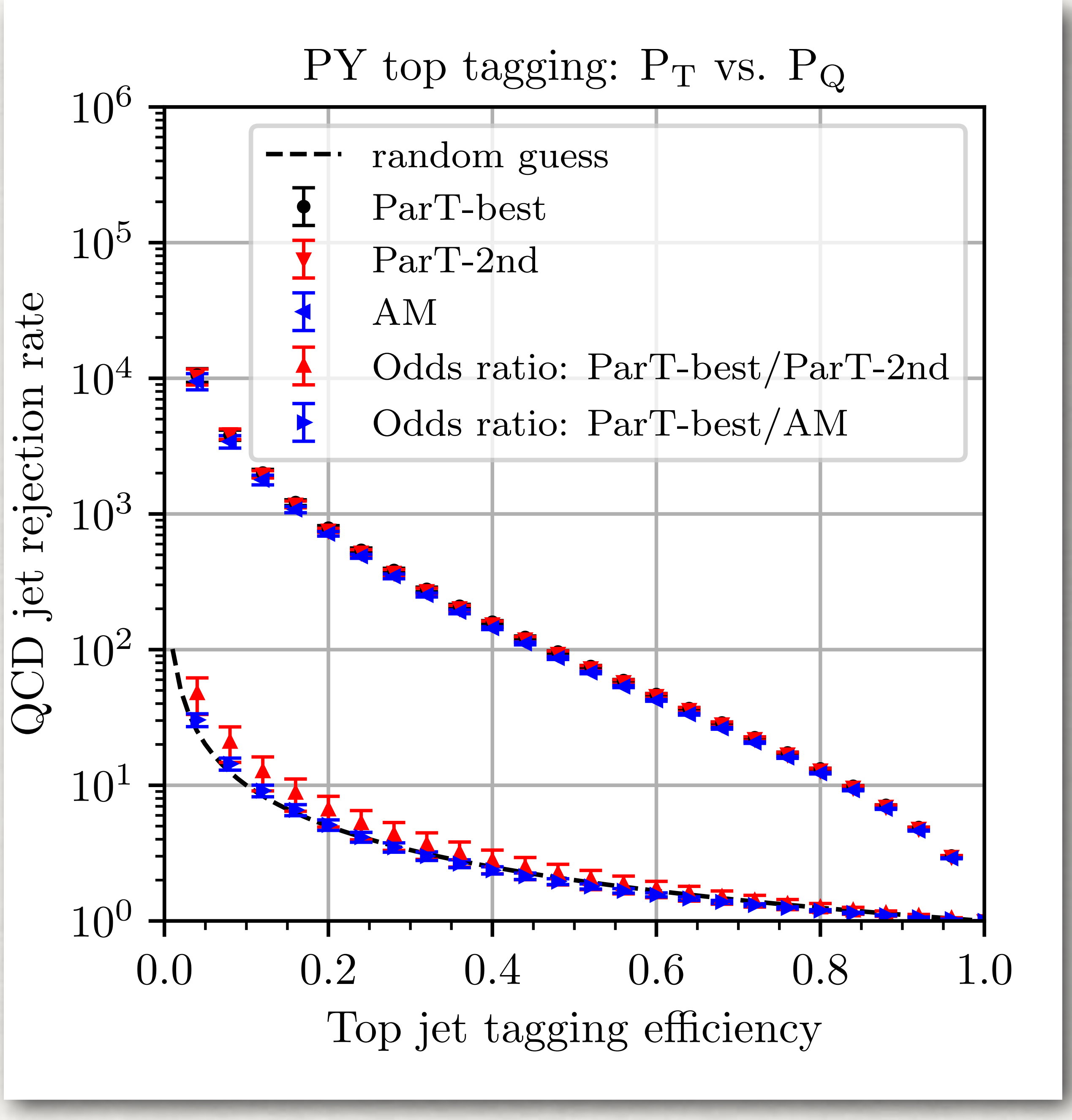
Recluster jet into
Recursive subjet network (pT ordered)

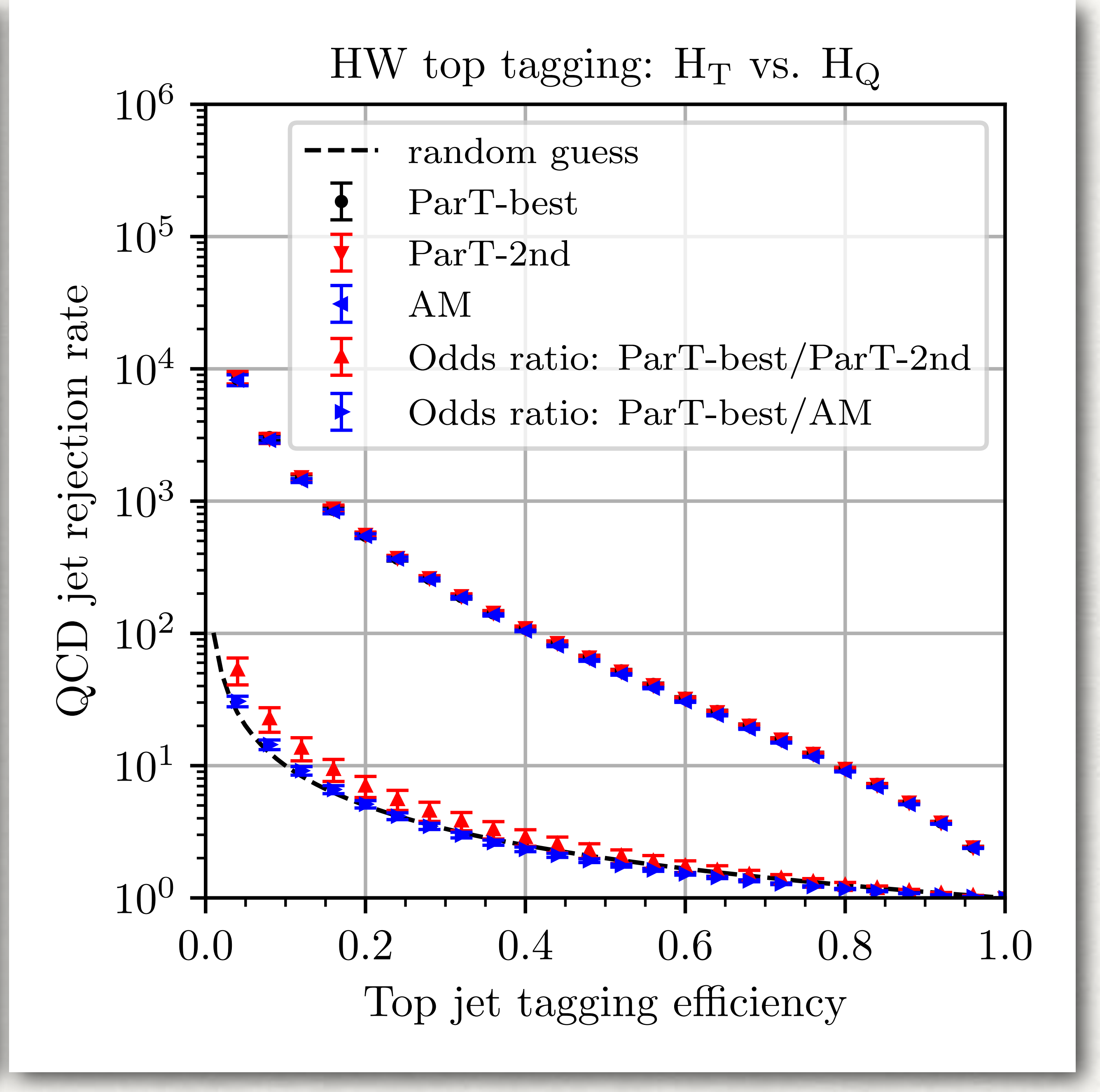


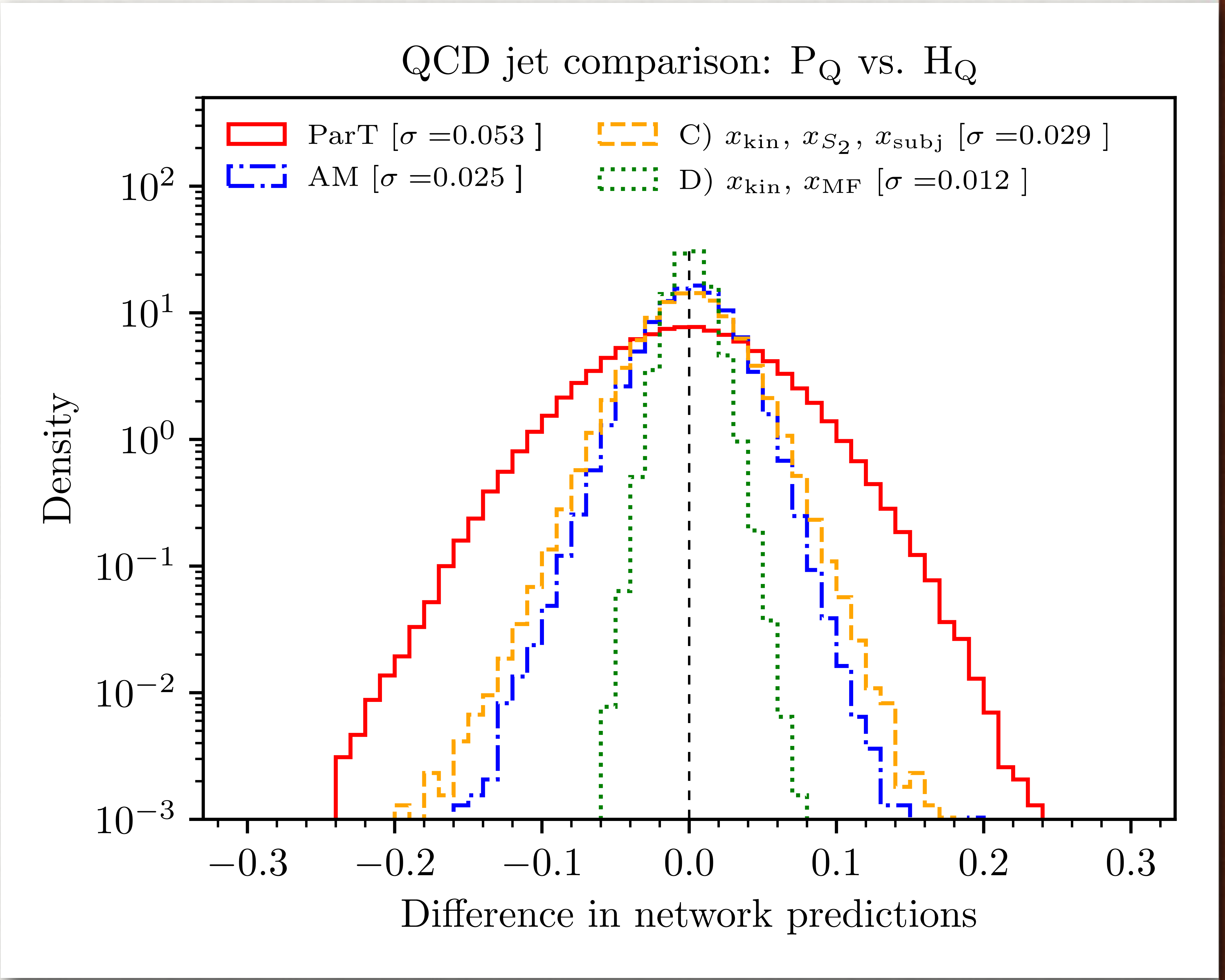
(b) A schematic diagram of subjet recursive module.

⭐input: subjet with multiple cone size
(R=0.1, 0.2, 0.3) =information of clustering
⭐Shared MLP for 2nd to 5th subjet to reduce paramters

PERFORMANCE AND STABILITY







AM model :**1GB GPU memory** on GeForce 1080Ti GPU(11.3TFLOPS) with 35% GPU utilization. need lots of preprocessing

ParT: **14GB GPU memory** RTX A6000 (38.7TFLOPS) GPU utilization 95%