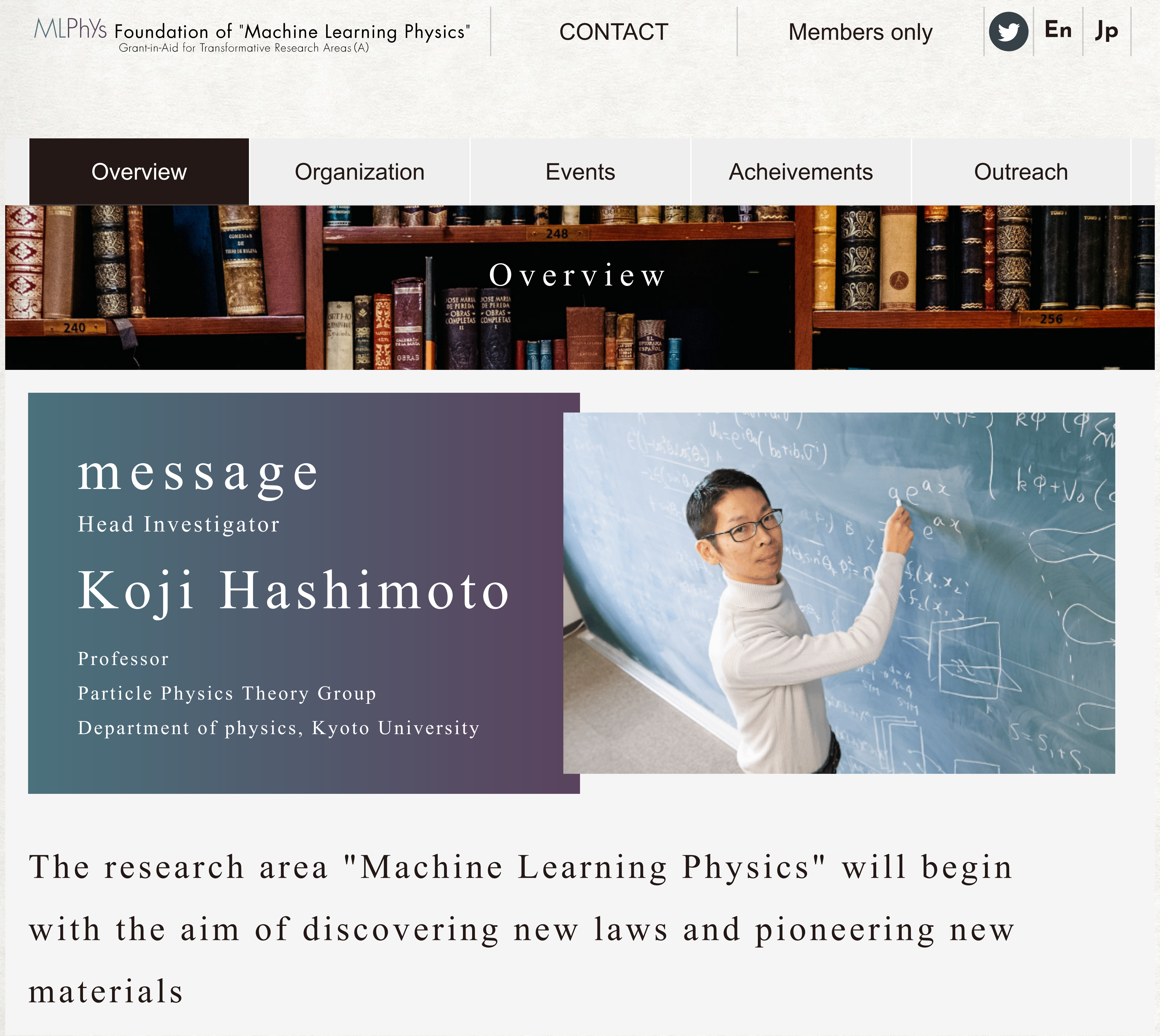
LANGUAGE OF JETS WITH TRANSFORMERS

MIHOKO NOJIRI
(KEK)

with Waleed Esmail(Munster) , Ahmed Hammad(KEK)

with Amon Furuihchi, Sung Hak Lim(IBS)

MLPHYSICS GRANT IN JAPAN“MACHINE LEARNING PHYSICS “



B01 Math and application of DL

B02 Statistical data and ML

B03 Topology and Geometry of ML

A01 Lattice

A02 Mihoko Nojiri HEP

Junichi Tanaka (ICEPP Tokyo, ATLAS)

Masako Iawasaki (Osaka Metropolitan Belle II)

Noriko Takemura and Hajime Nagahara (Data Science)

A03 Condensed Matter

A04 Quantum and Gravity

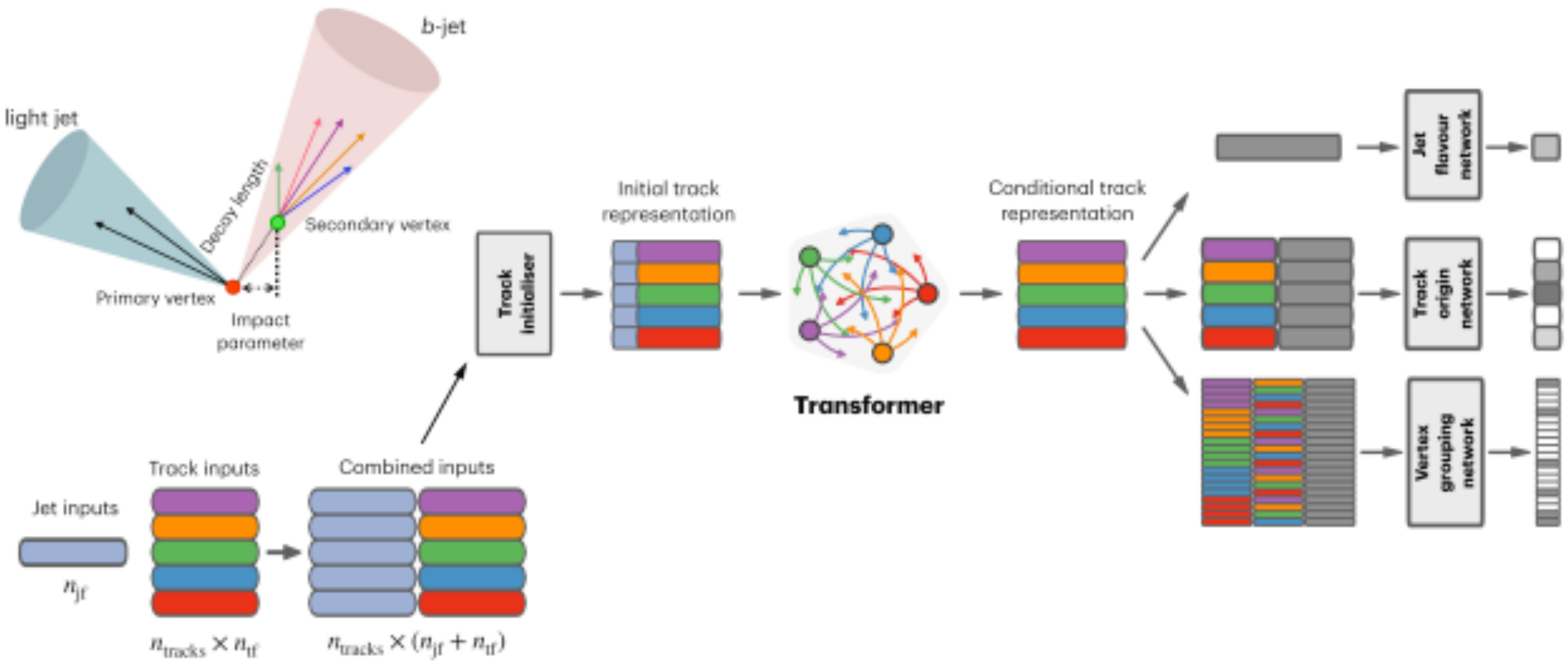
PD. Ahmed Hammad

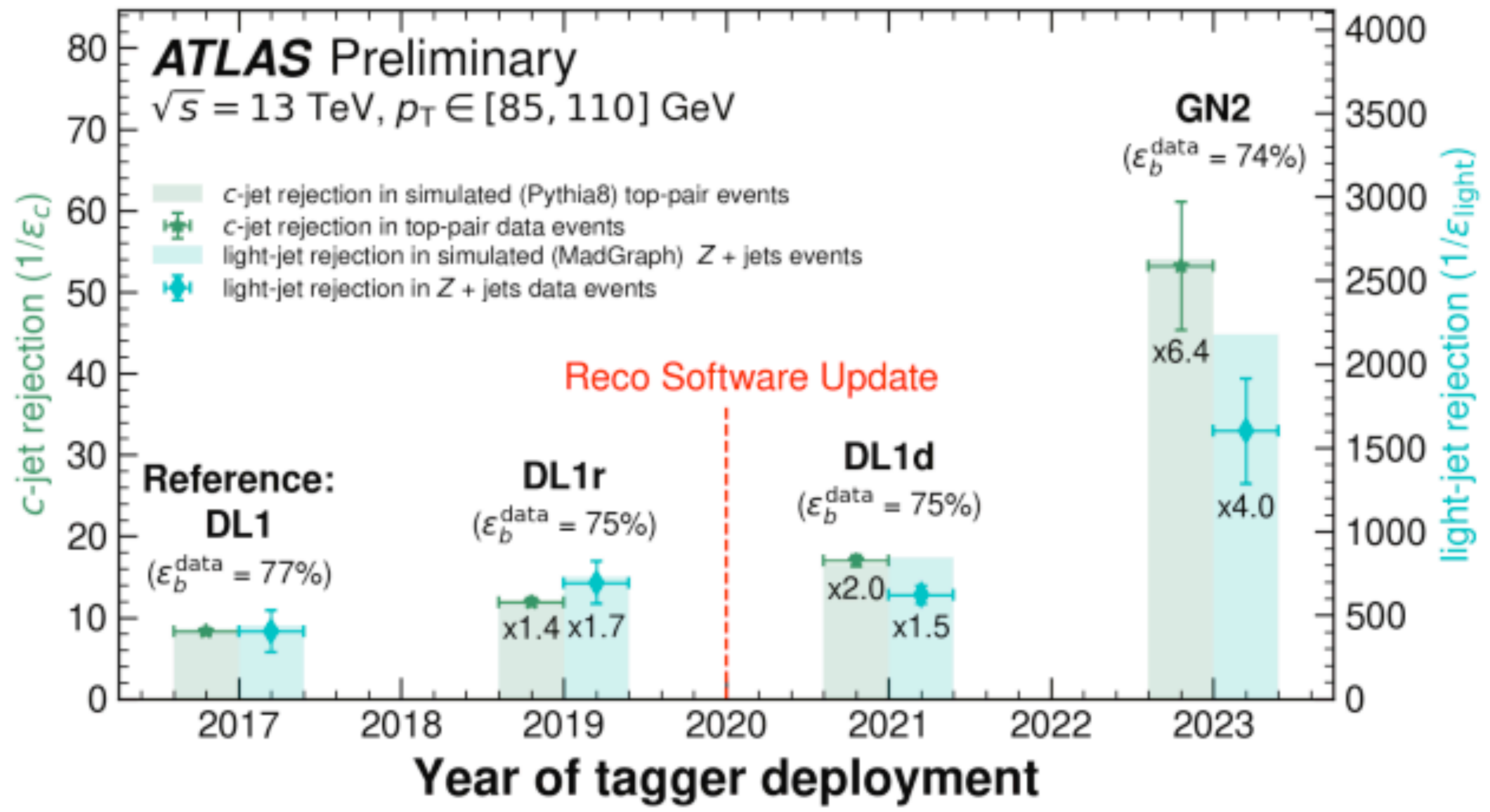
2017-2020: Ph.D Basel University, Basel Switzerland

2020-2023: SeoulTech, Korea

2023- KEK

ML already helping particle physics significantly: Jet tagging using transforrmers





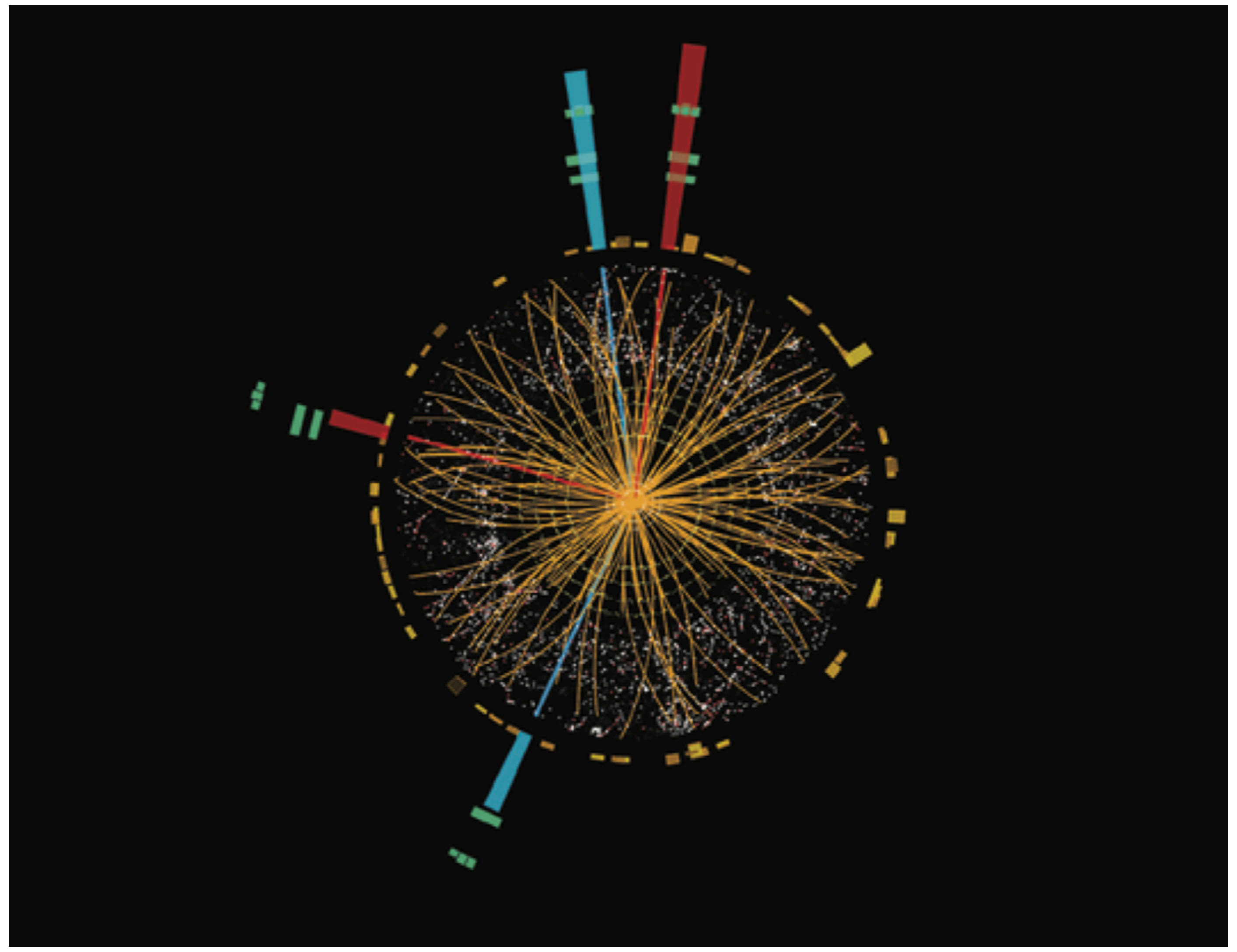
Quantum coherence vs Machine learning

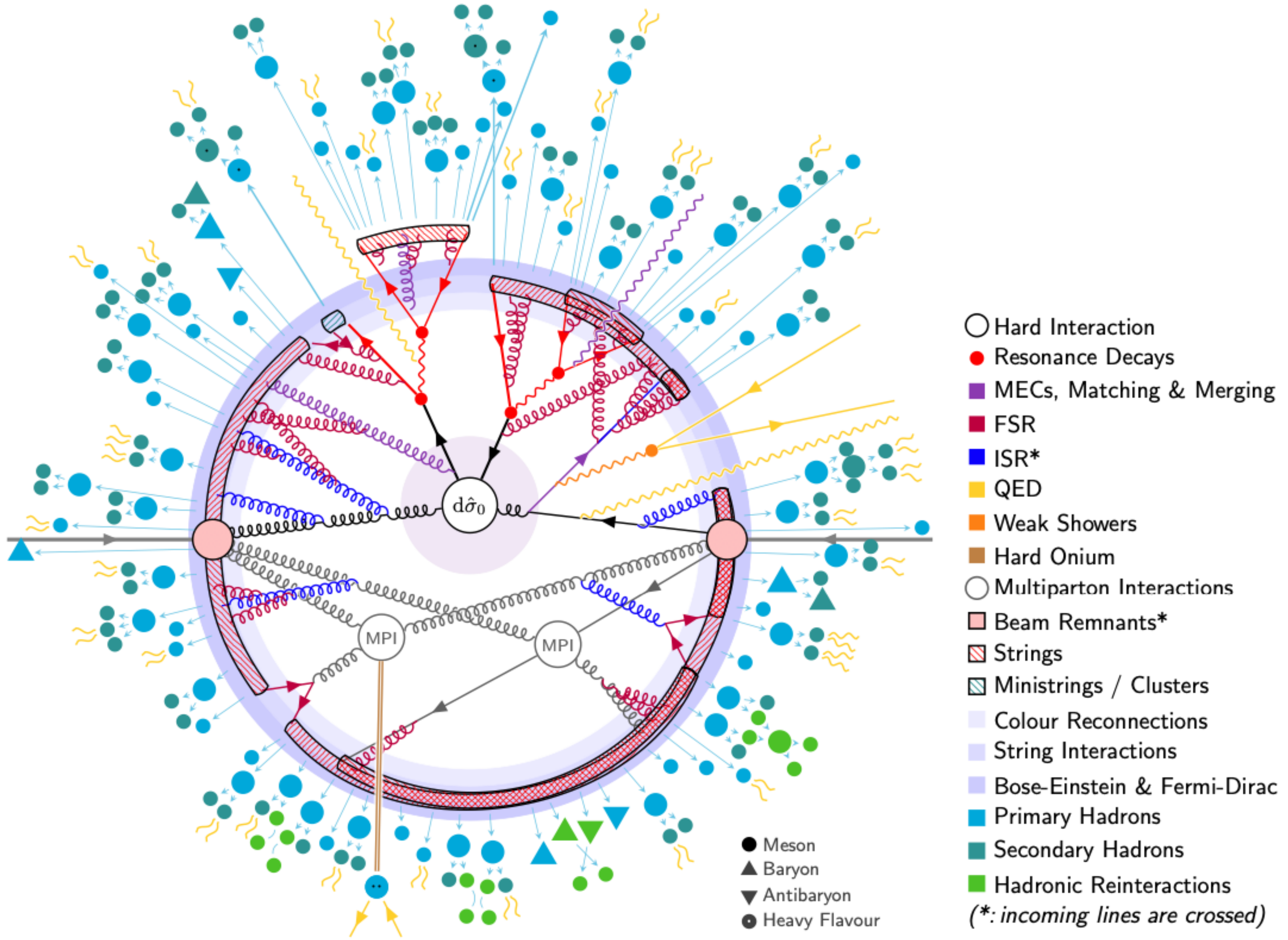
Hadolon collider events are the result multiple particles interaction with

Color coherence, factorization, spin correlation, and entanglements.

How they affect training results?

Event generators





In one way or another, this is how events are modelled in all event generators.

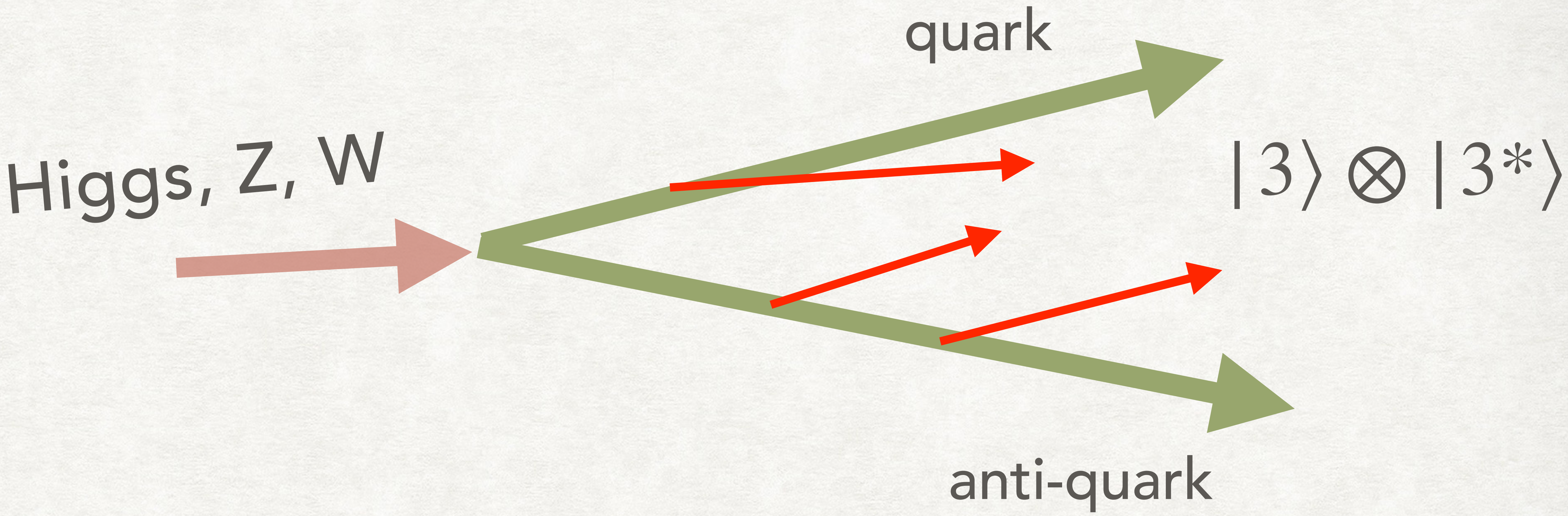
Jet structures

quark and gluon jet

Parton shower

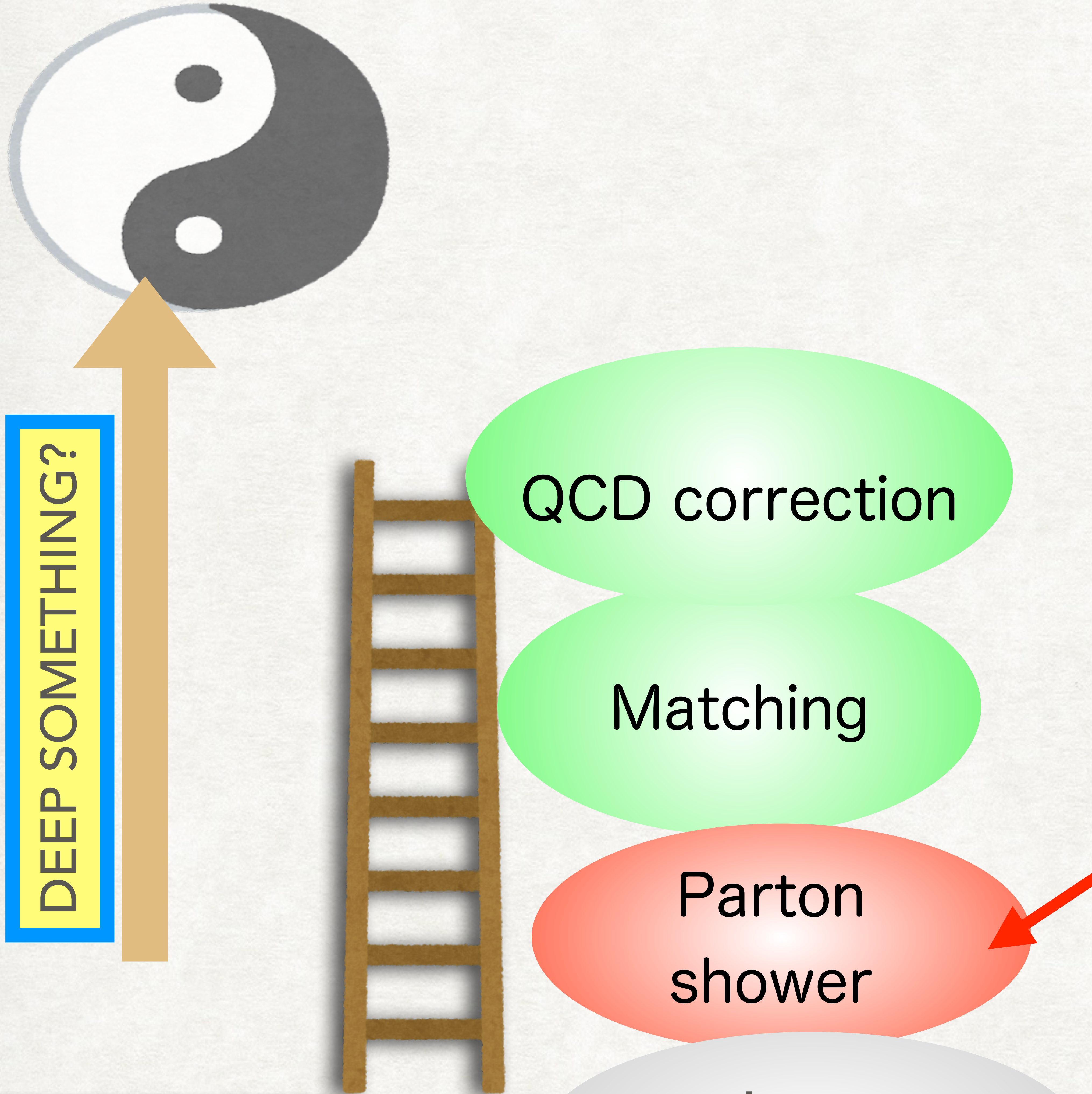
(angular ordering and pT ordering)

Heavy particle decay



High pT H, Z, top
is important for BSM study
and they maybe highly boosted

Particle Theory in DeepLearning Era



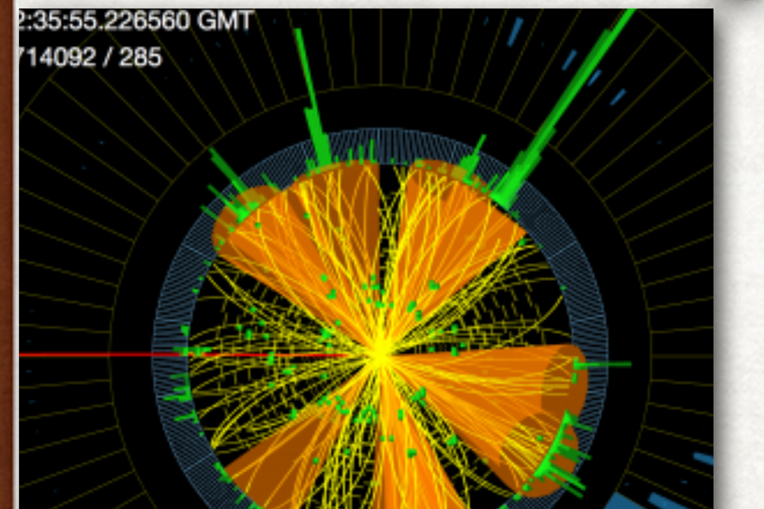
QCD multiple interactions
connecting BSM to events

Madgraph: Automatic Amplitude calculation in NLO level

QCD aware definition of jets(fastjet)

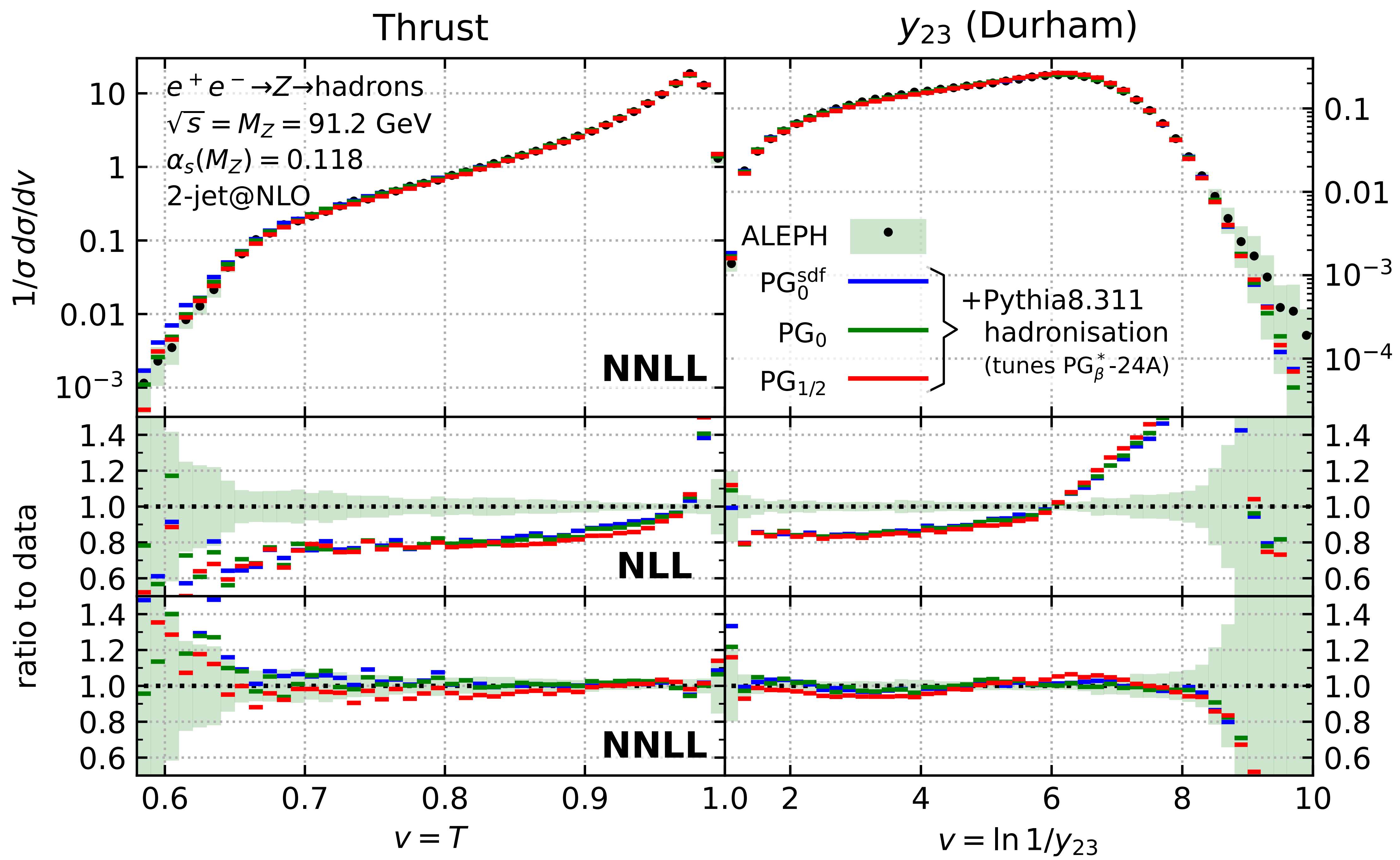
Matrix element and Parton shower matching
MLM, CKKW → 2007 Madgraph Sherpa

angular order, pT order →Dipole shower
with NNLL correction. (Panscale…)



Deep Learning require the theory applicable to
soft particles in the events.

Comparison to LEP data



$$\alpha_s(m_Z) = 0.118$$

Colour is handled using the NODS scheme which gives full colour accuracy at NLL for global observables (includes those shown)

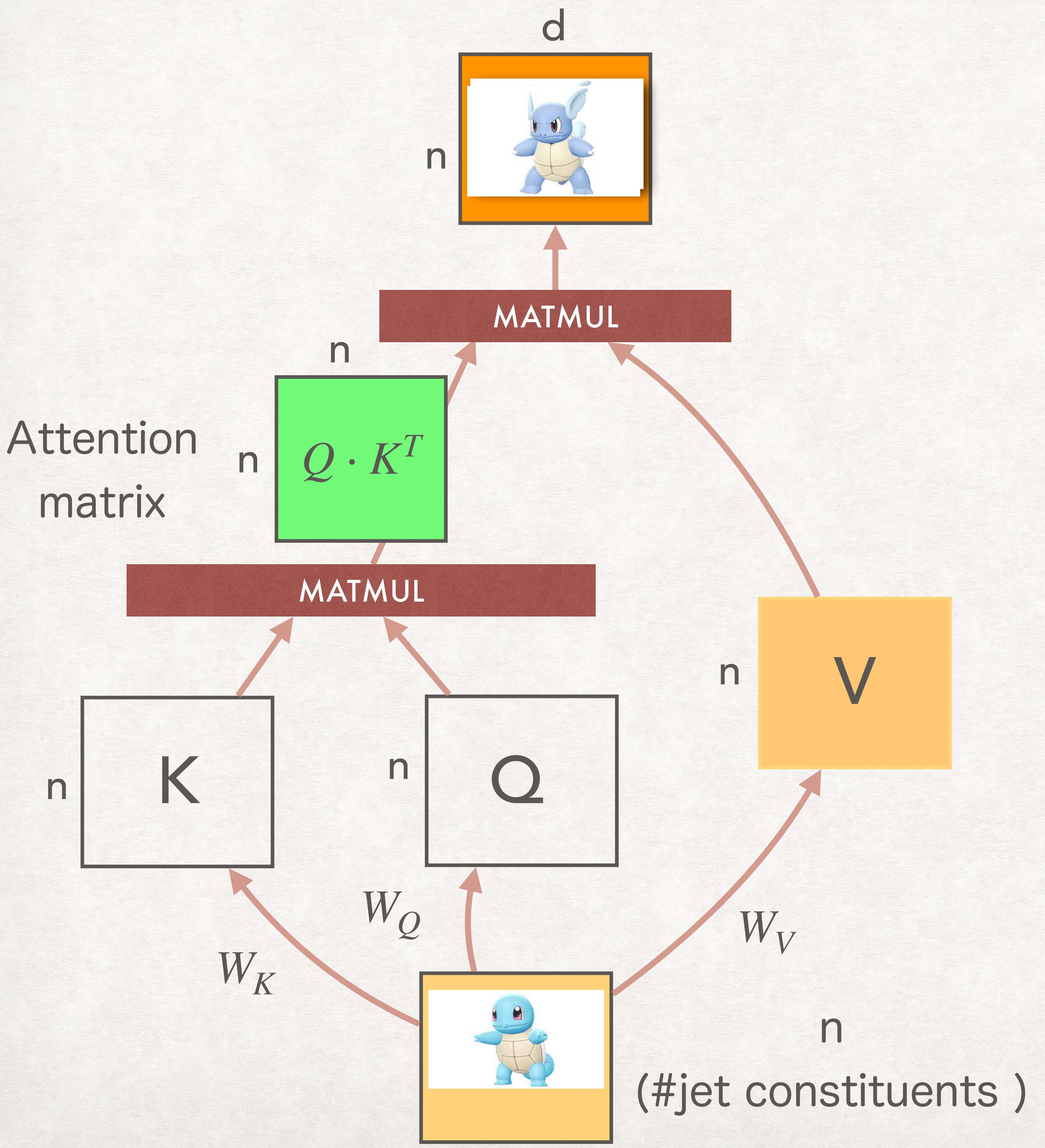
- Inclusion of NNLL potentially resolves the issue of needing an anomalously large value of $\alpha_s(m_Z)$ to achieve good agreement with LEP data. ($\alpha_s(m_Z) = 0.137$ in Pythia's Monash 13 tune * arxiv:1404.5630, Skands, Carrazza, Rojo)
- Some caution needed as no 3-jet NLO matching, which is known to be releva from the 2-jet regi
- A comprehensive study of shower uncertainties is still to be done.

PanScale shower reproducing $\alpha_s(m_Z)$ at last!

https://gsalam.web.cern.ch/panscales/

*This should be taken as an average α_s^{eff} not an $\alpha_s^{\overline{MS}}$

"PARTICLE TRANSFORMER" FOR JET IDENTIFICATION



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

$$Q = XW_Q, K = XW_K, V = XW_V$$

SETUP

- n(particles in the jet) x d(features)
- n x n Attention matrix from Key K and Query Q
- Multiply Value V to get n x d output
- stack attention layers for X → X'→X''… with skip connection
- $X' = X + \delta X,\ \delta X = A \cdot V$

Particle Features for jet classification

Particle momentum

charge,particle ID

displaced vertex

Category	Variable	Definition
Kinematics	$\Delta\eta$	difference in pseudorapidity η between the particle and the jet axis
	$\Delta\phi$	difference in azimuthal angle ϕ between the particle and the jet axis
	$\log p_{\mathrm{T}}$	logarithm of the particle's transverse momentum p_{T}
	$\log E$	logarithm of the particle's energy
	$\log \frac{p_{\mathrm{T}}}{p_{\mathrm{T}}(\text{jet})}$	logarithm of the particle's p_{T} relative to the jet p_{T}
	$\log \frac{E}{E(\text{jet})}$	logarithm of the particle's energy relative to the jet energy
	ΔR	angular separation between the particle and the jet axis ($\sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$)
Particle identification	`charge`	electric charge of the particle
	`Electron`	if the particle is an electron (`|pid|==11`)
	`Muon`	if the particle is an muon (`|pid|==13`)
	`Photon`	if the particle is an photon (`pid==22`)
	`CH`	if the particle is an charged hadron (`|pid|==211 or 321 or 2212`)
	`NH`	if the particle is an neutral hadron (`|pid|==130 or 2112 or 0`)
Trajectory displacement	$\tanh d_0$	hyperbolic tangent of the transverse impact parameter value
	$\tanh d_z$	hyperbolic tangent of the longitudinal impact parameter value
	σ_{d_0}	error of the measured transverse impact parameter
	σ_{d_z}	error of the measured longitudinal impact parameter

0. INTERPRETATION

What type of feature contributing to classification?

HL feature+MLP vs Transformer

Amon Furuichi, Sung Hak Lim, Mihoko M. Nojiri JHEP 07 (2024) 146 JHEP 07(2025) 111

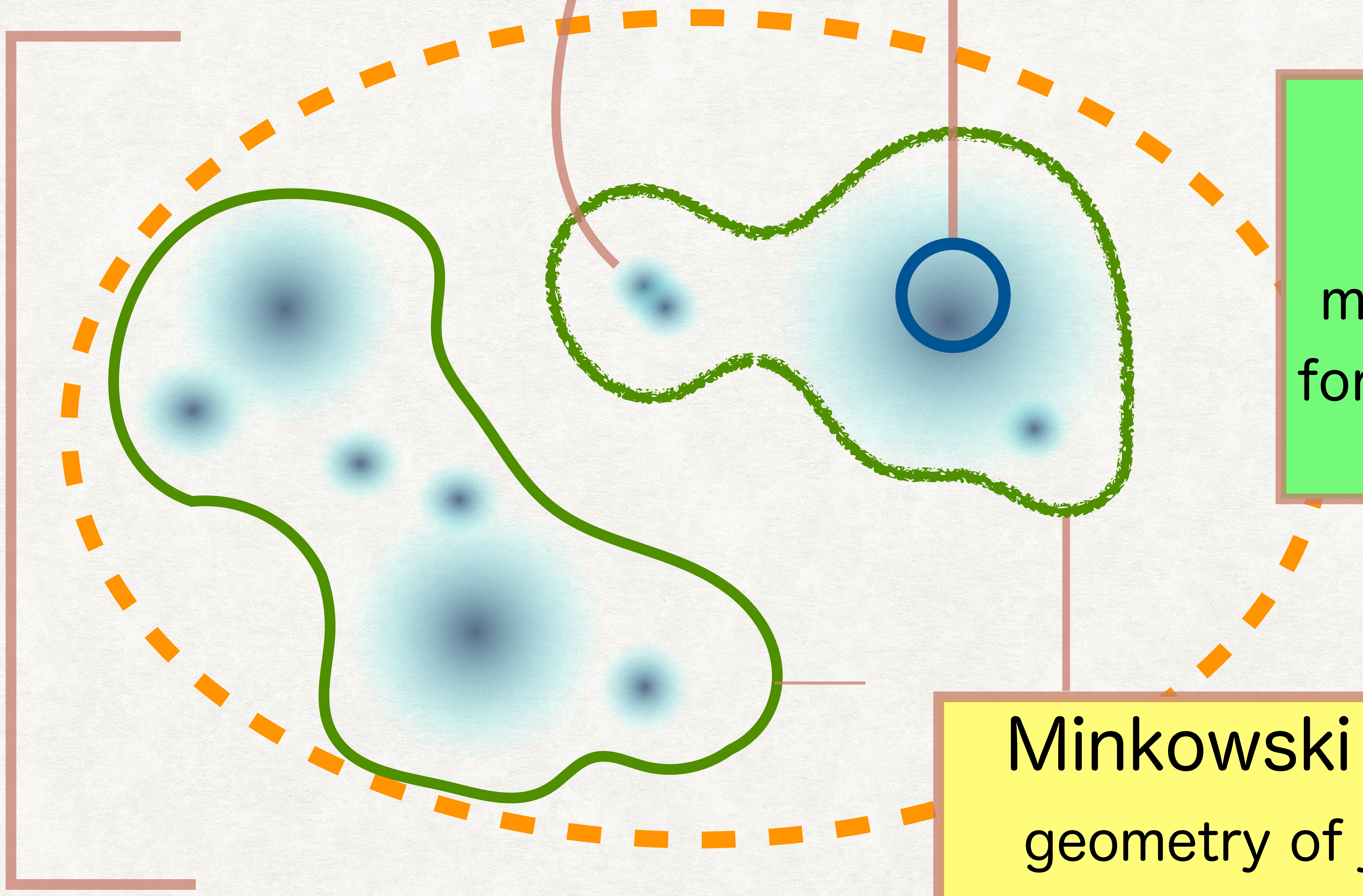
pt distribution of particles

Jet spectrum

two point Energy correlation (unlocalized sampling)

= EFP with N=2

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



Subjet

Localized sampling momentum and counting for various angular sccale

R=0.1, 0.2, 0.3

Minkowski Functionals

geometry of jet cosntituent distribution with various energy threshold

QCD jet (dijet) rejectrion factors for 50% top jet efficiency

	Pythia(8.308 simple shower)	Herwig (7.2) default
MLP by IRC safe 2point correlation and global	80.7	56.0
MLP using all HLF	85.7	61.3
ParT	90.5	62.6

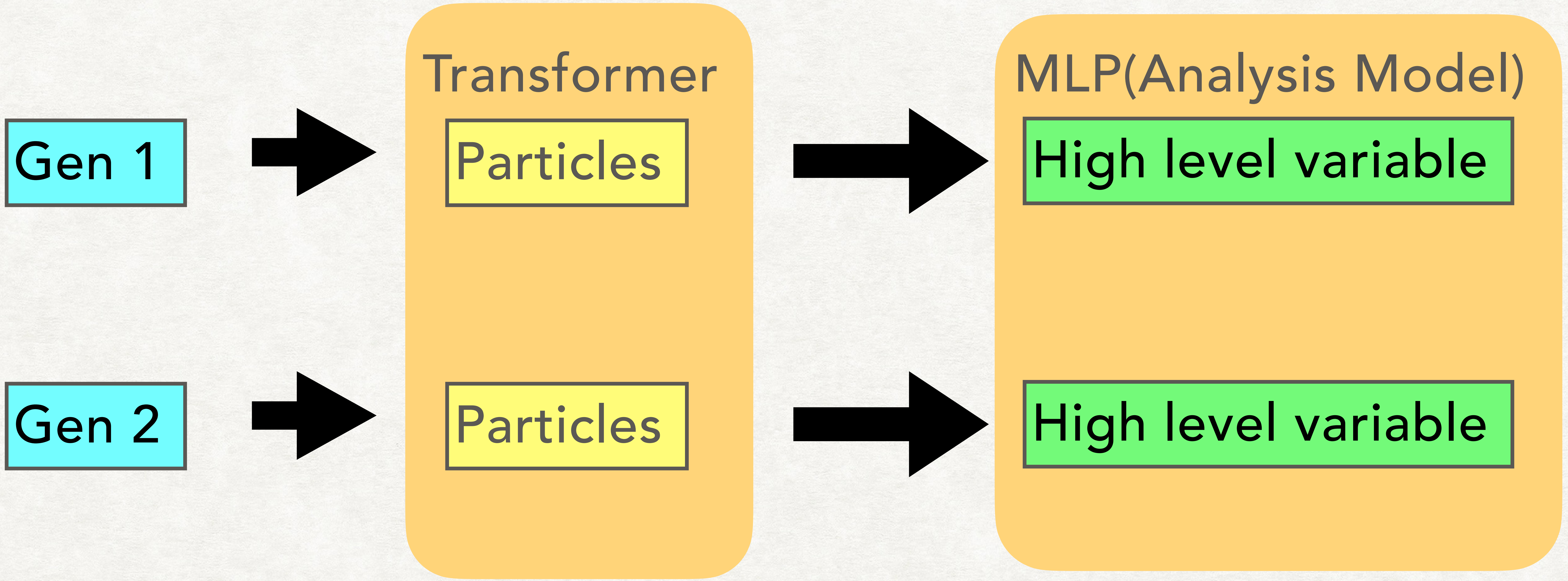
for 500GeV<PT<600GeV and 150 GeV<mj<200GeV

The simulation dependence coming from modeling difference(parton shower and hadronization)

Accurate parton shower(dipole +NNLL) modeling will help to settle the difference?

GENERATOR COMPARISON USING ML

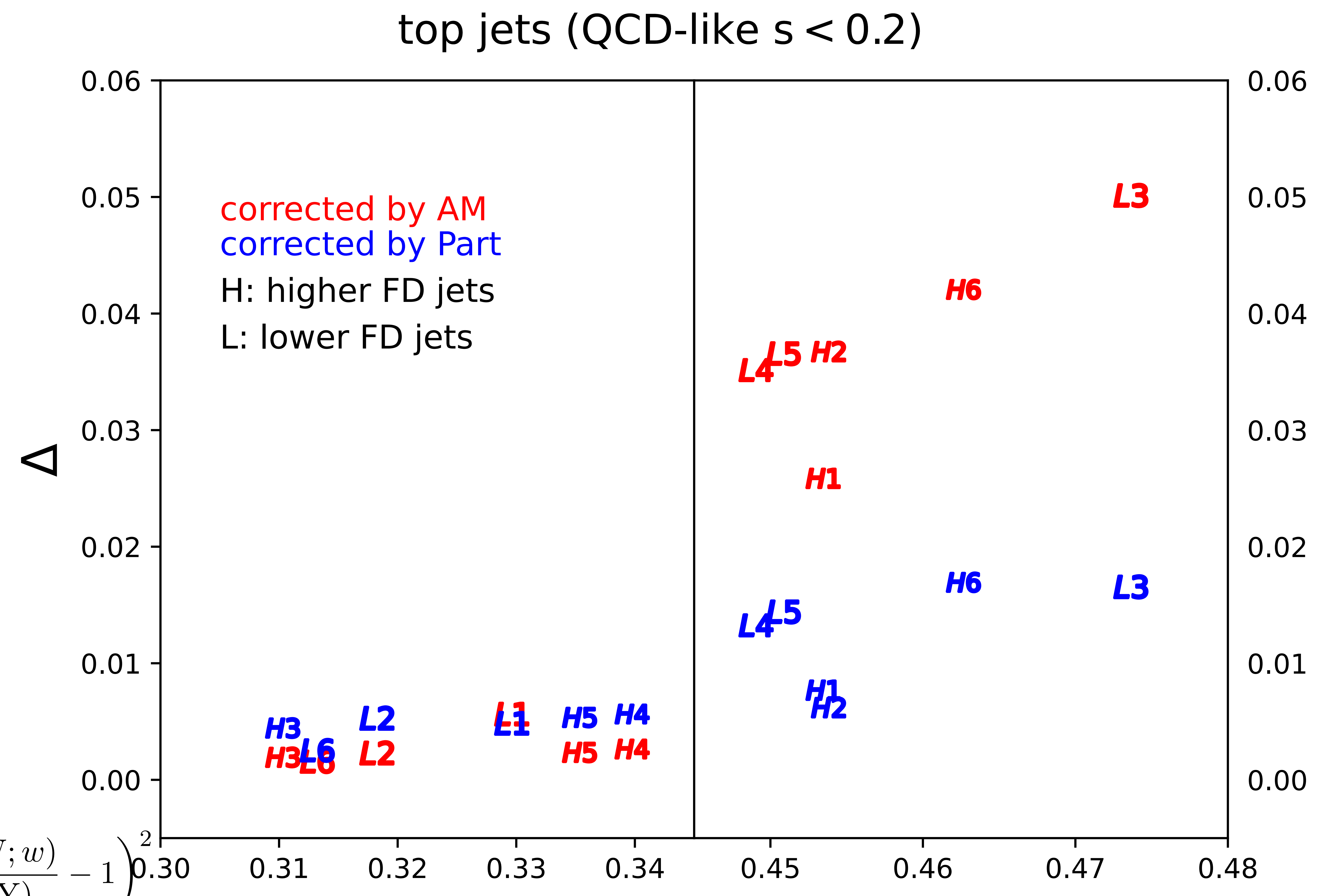
Amon Furuichi, Sung Hak Lim, Mihoko M. Nojiri JHEP 07 (2024) 146 JHEP 07(2025) 111



$s_{gen}(x)$: generator classifier output → estimated probability ratio $w = \dfrac{s_{gen}(x)}{1 - s_{gen}(x)}$

Transformer: no human bias, but poor training stablity-MLP using highlevel input : stable prediction, good for subtle generator comparison

goodness of the reweighiting of selected >3 point EFP distributions



deviation of event weights

$$\Delta_{\mathrm{L}}(\mathrm{FD}_n) := \left(\frac{N_{\mathrm{L}}(\mathrm{FD}_n|\mathrm{HW};w)}{N_{\mathrm{L}}(\mathrm{FD}_n|\mathrm{PY})} - 1\right)^2$$

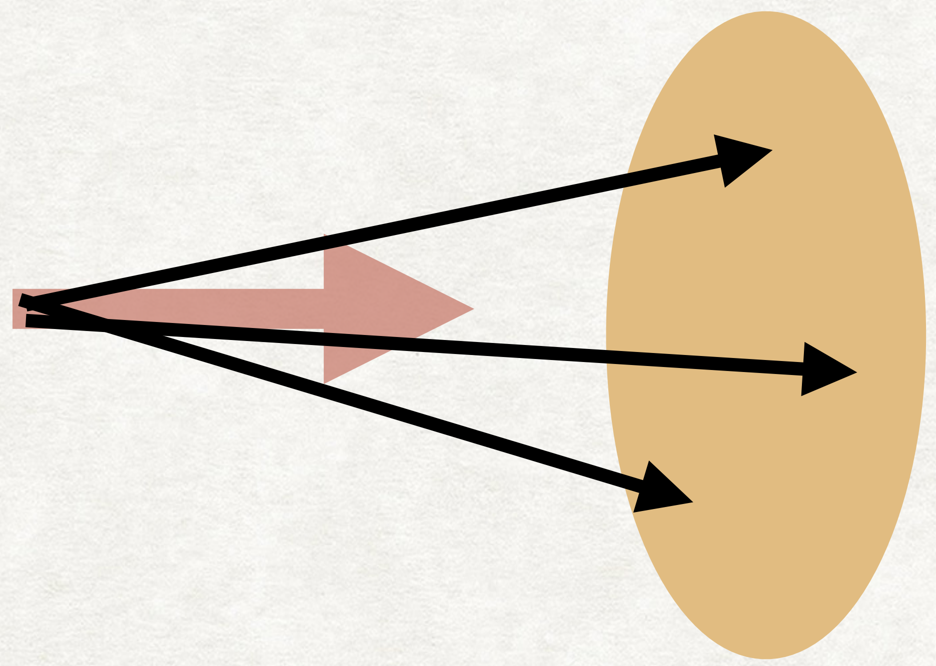
2. RESTRICTIVE TRANSFORMERS

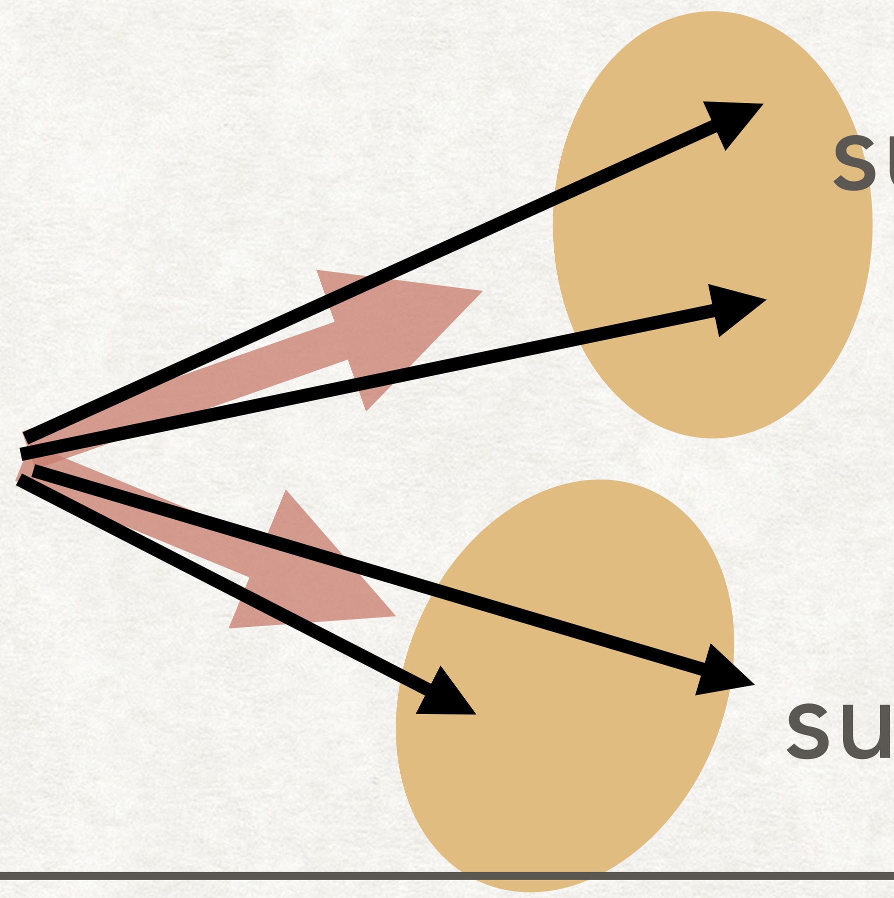
TRANSFORMER USING PHYSICS STRUCTURE

A. TRANSFORMER FOR PARTON SHOWER X HADRONIZATION

"Ahmed Hammad, & MN arXiv 2404 14677 JHEP 06 (2024) 176

- **Hard Process** = Partons(quarks and gluons) {y}
- **a jet:** P(hadrons in jets | parton ~ jet) = $P(\{x_i\} \,|\, \{y\})$
- **jet with substructure** $P(\{x_i\} \,|\, \{y_\alpha\})$

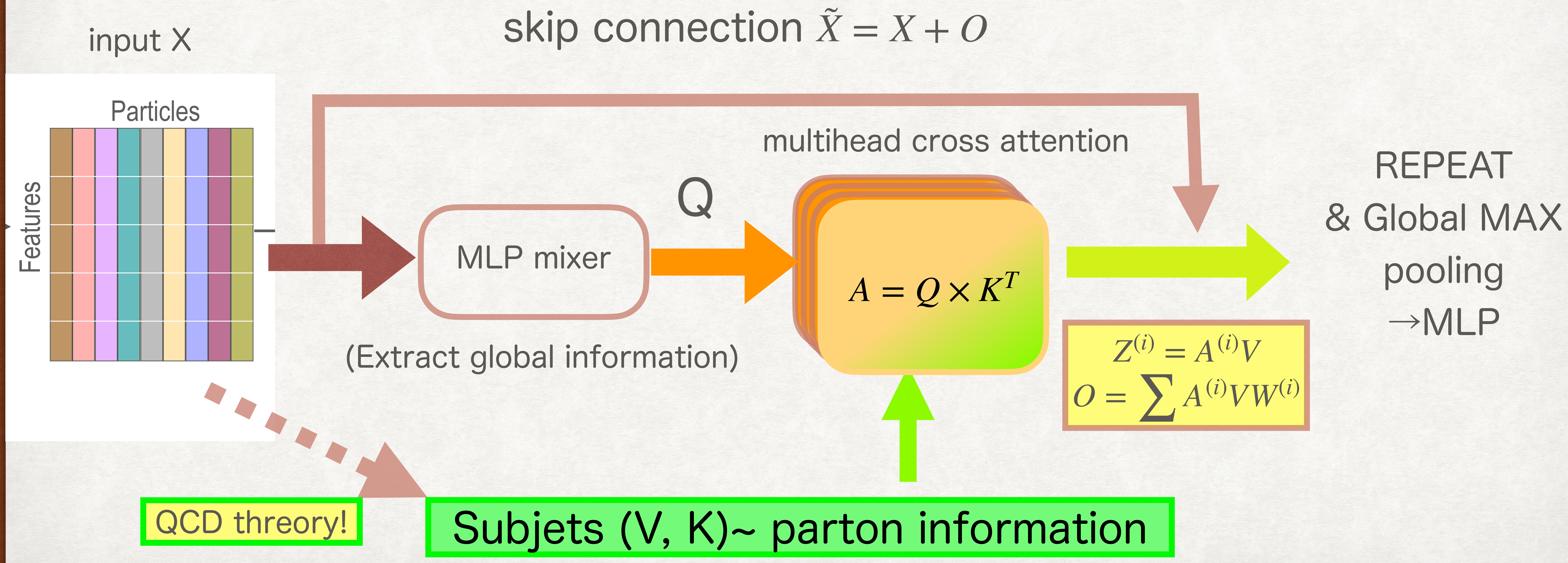




We need the network forcusing on partons(subjets/jets) vs hadrons

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

Direct transformer networks to calculate attention between subjet vs particles

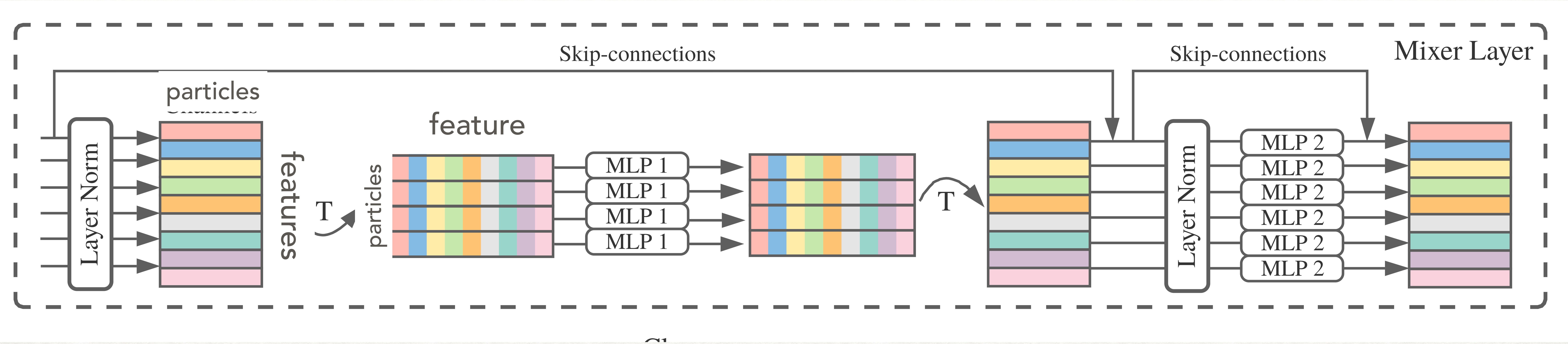


THE PERFORMANCE FOR TOP VS QCD CLASSIFICATION

	Accuracy	AUC	$1/\epsilon_B(\epsilon_s = 0.5)$	$1/\epsilon_B(\epsilon_s = 0.3)$	Parameters
		Lorentz invariance based networks			
PELICAN[35]	0.9426	0.987	–	2250 ± 75	208K
LorentzNet[70]	0.942	0.9868	498 ± 18	2195 ± 173	224K
L-GATr[71]	0.942	0.9870	540 ± 20	2240 ± 70	–
		Attention based networks			
ParT[49]	0.940	0.9858	413 ± 6	1602 ± 81	2.14M
MIParT[50]	0.942	0.9868	505 ± 8	2010 ± 97	720.9K
Mixer[21]	0.940	0.9859	416 ± 5	–	86.03K
OmniLearn[72]	0.942	0.9872	568 ± 9	2647 ± 192	1.6M
Plain Transformer*	0.927	0.979	362 ± 7	780 ± 73	1.7M
IAFormer*	**0.942**	**0.987**	$\mathbf{510 \pm 6}$	$\mathbf{2012 \pm 30}$	**211K**

This shows “Simulated data” is build from parton->hadron picture

MLP MIXER CAN BE VERY SIMPLE



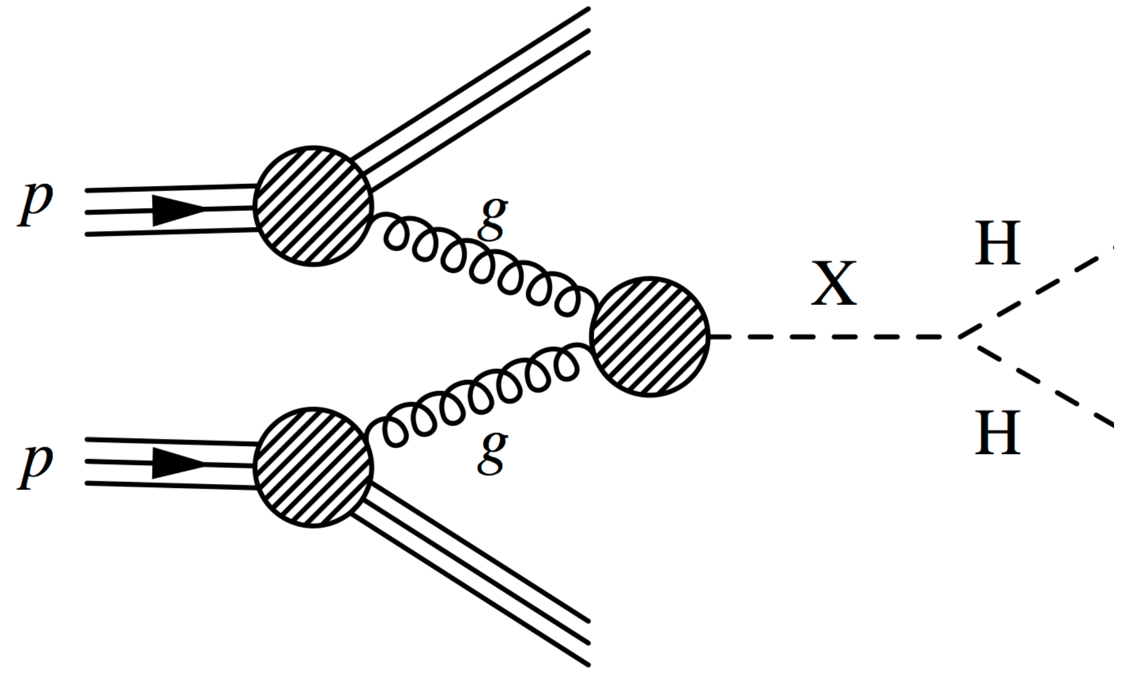
MLP 1 : operate on features

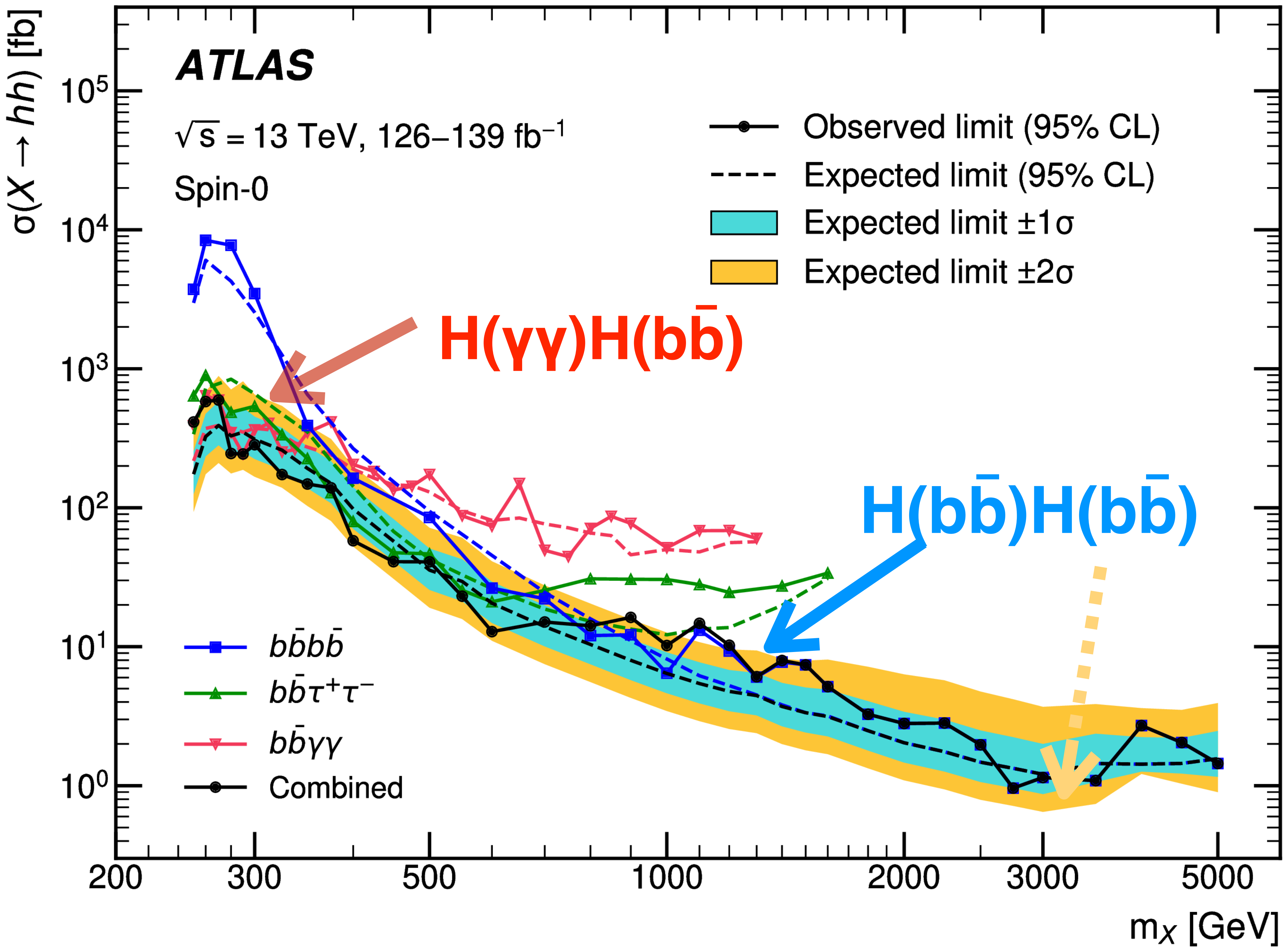
MLP 2: operate on particles

B. GLOBAL EVENT ANALYSIS AND CROSS ATTENTION

X→HH

[Phys. Rev. Lett. 132 (2024) 231801]





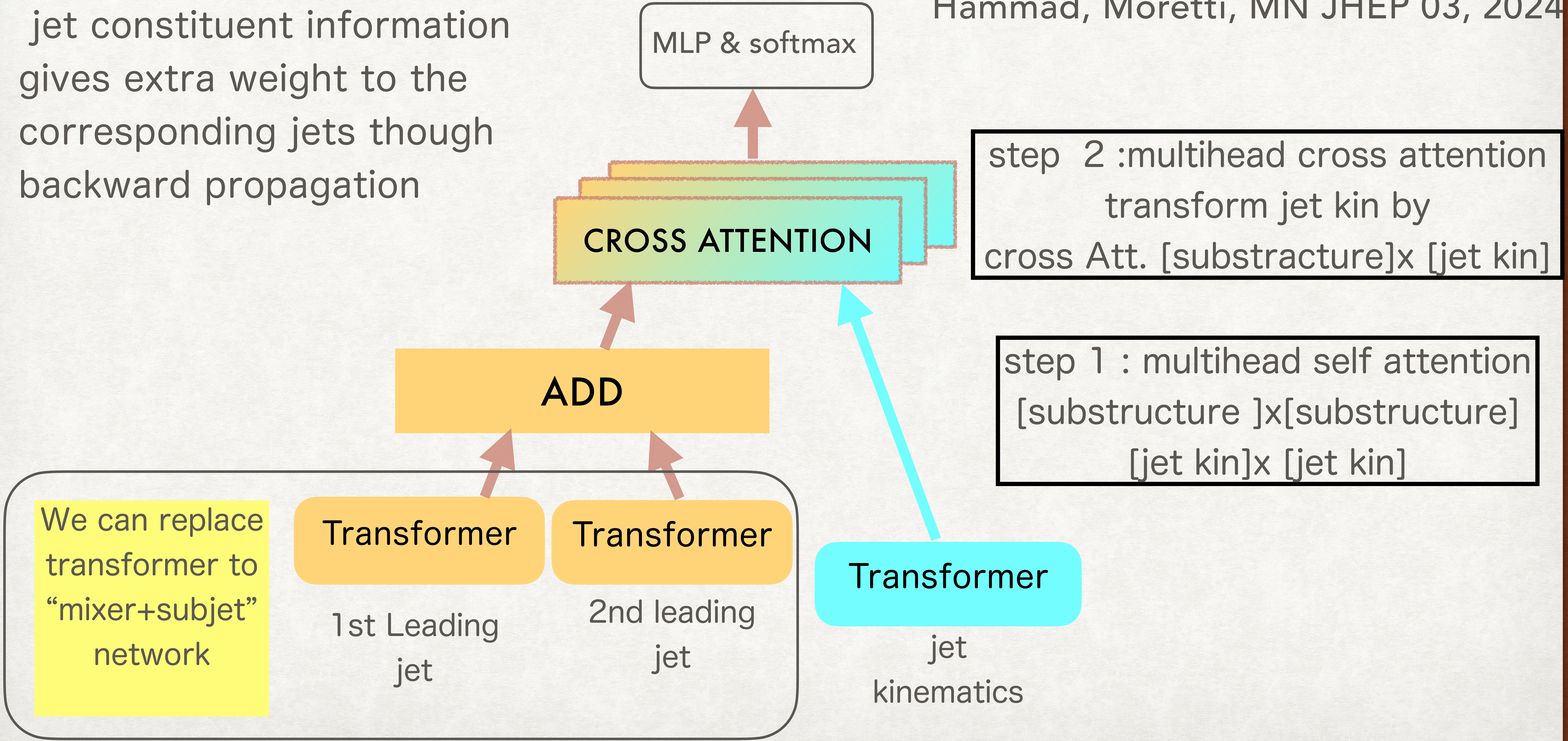
H(bb̄)H(bb̄) most sensitive channel for $m_X > 400/500$ GeV

H(γγ)H(bb̄) complement in the low mass

Cross attention for 2 fatjet events

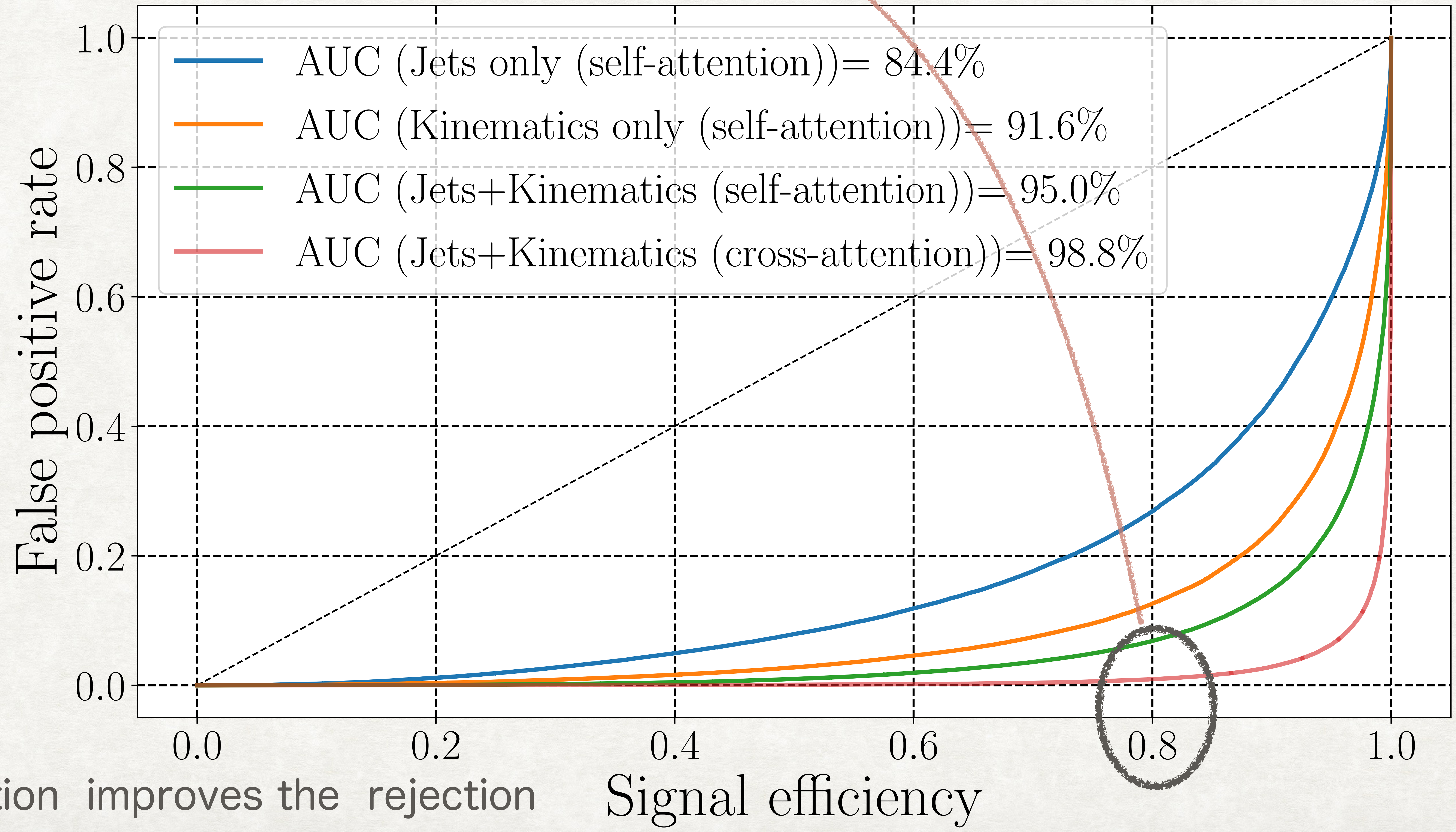
Hammad, Moretti, MN JHEP 03, 2024

- jet constituent information gives extra weight to the corresponding jets though backward propagation



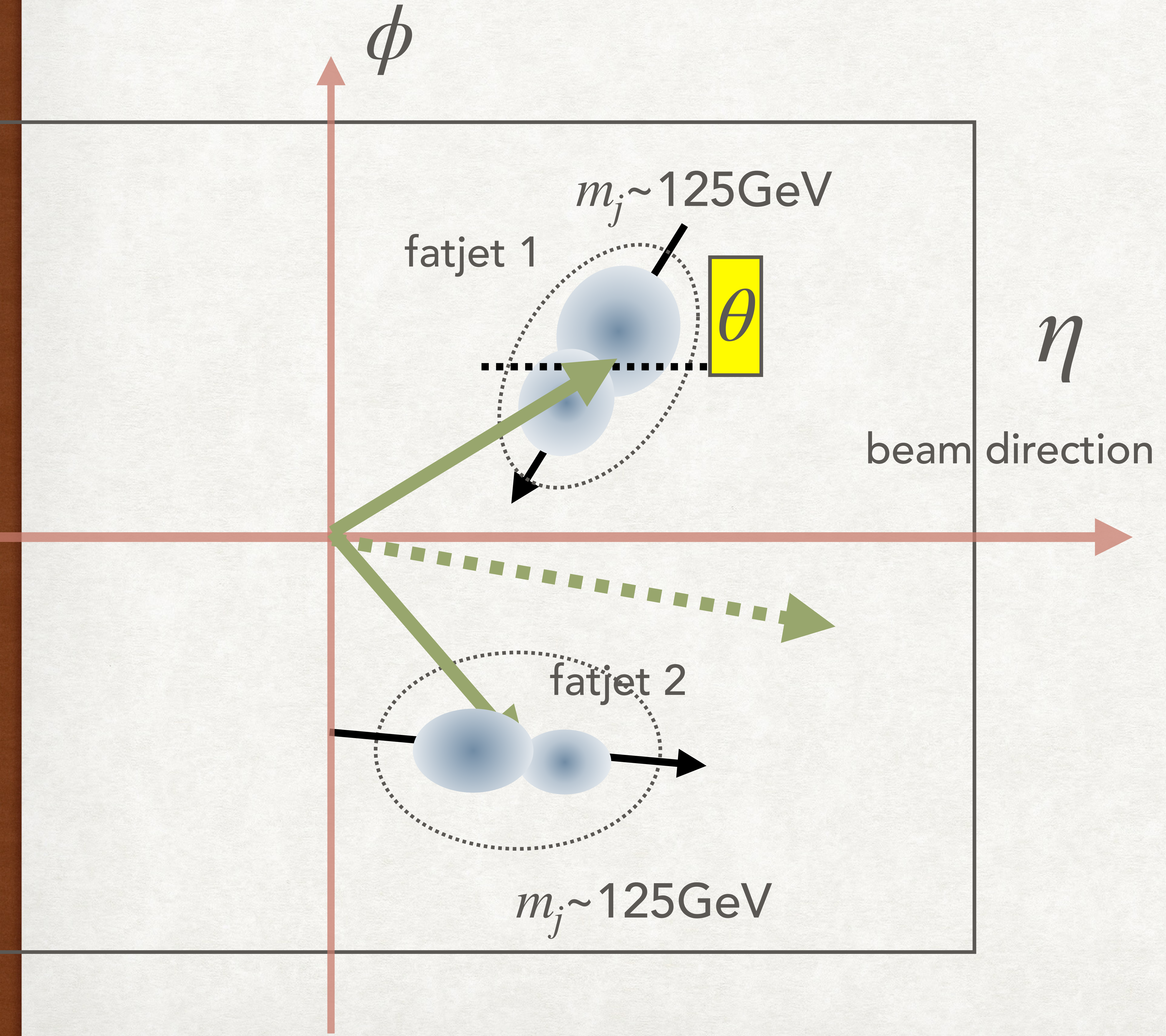
IMPROVEMENT USING CROSS ATTENTION

factor 5 improvement at the same acceptance.

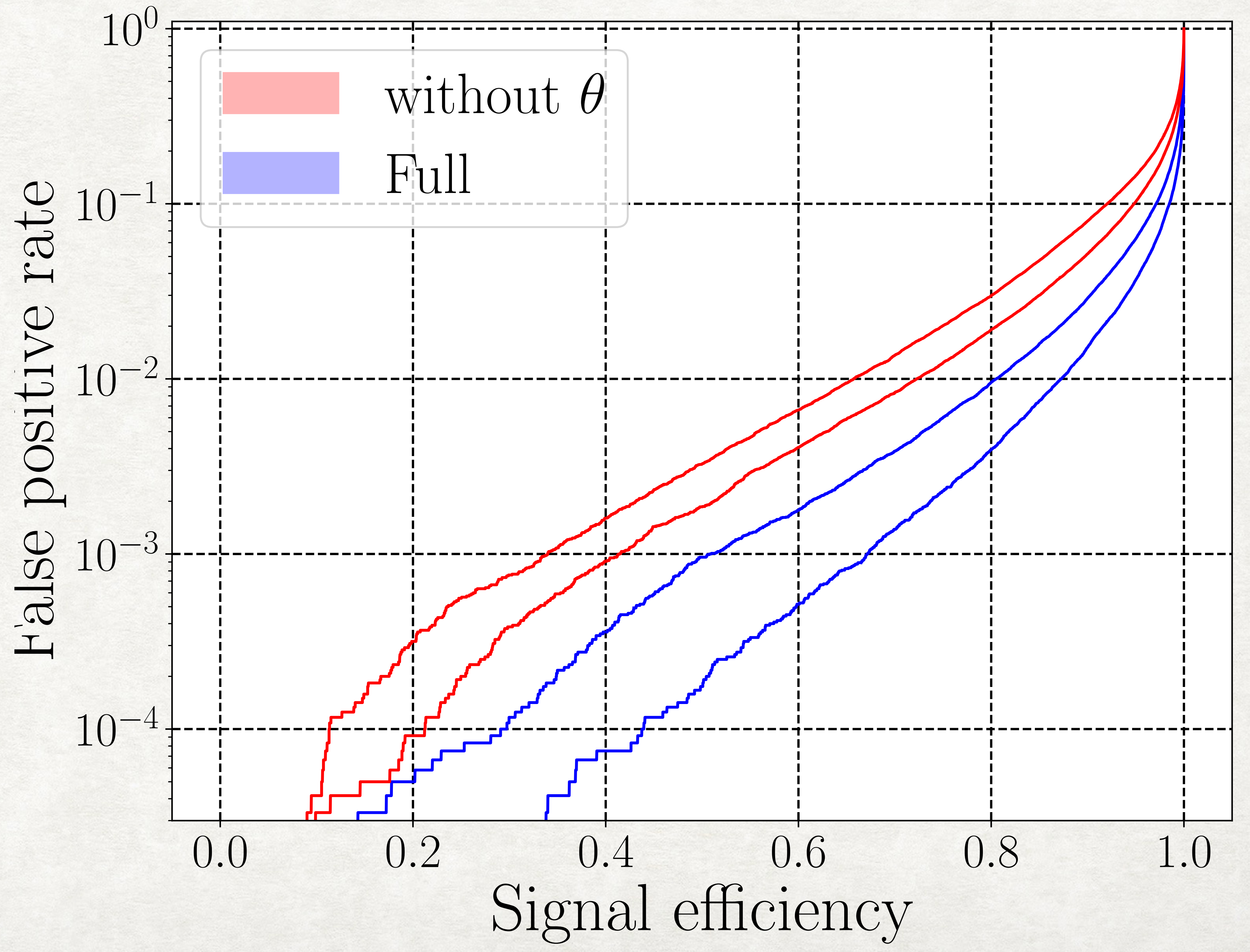


Cross attention improves the rejection efficiency significantly

JET SHAPE DEPENDENCE OF THE RESULTS

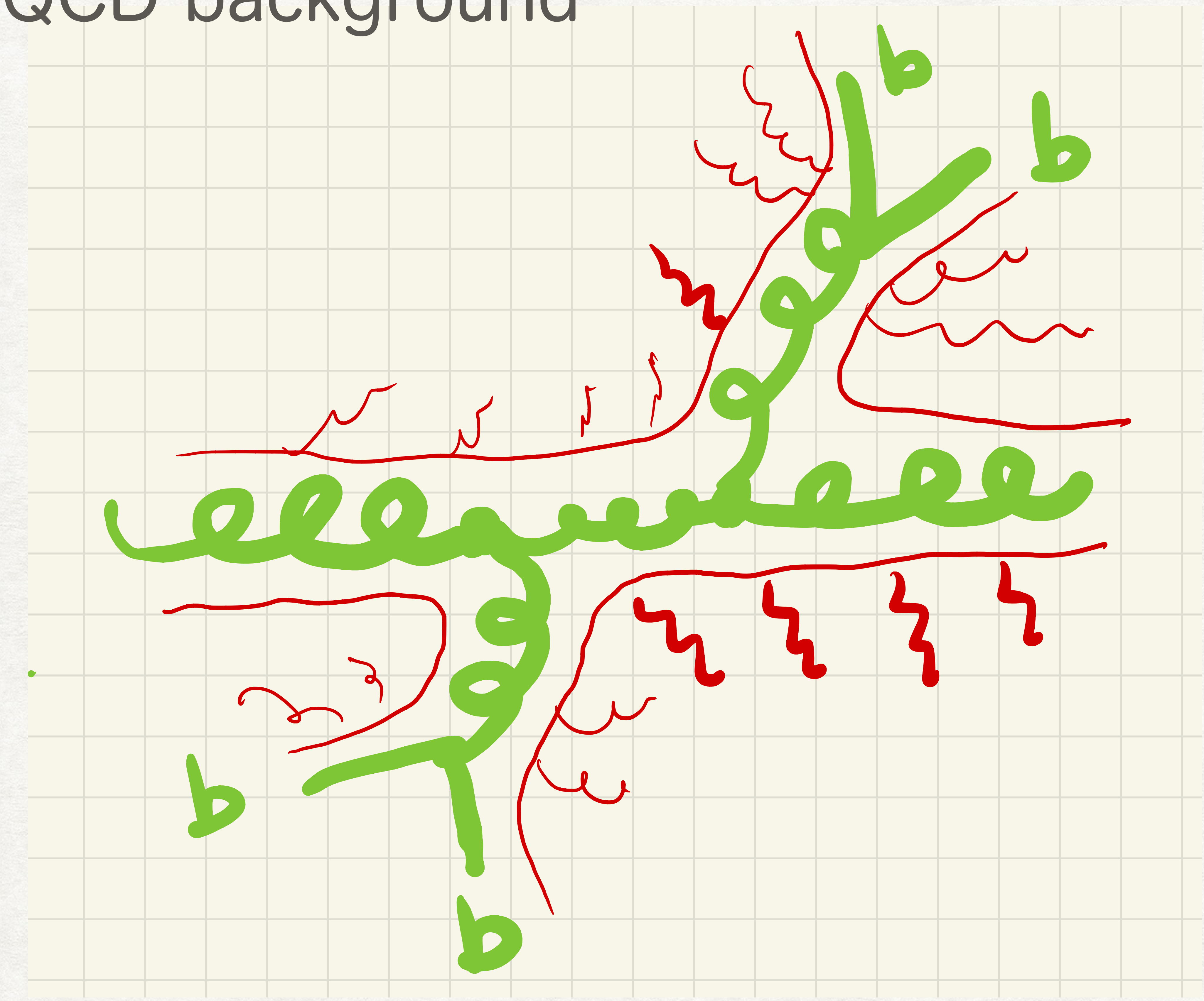


Decay correlation is important because background is correlated



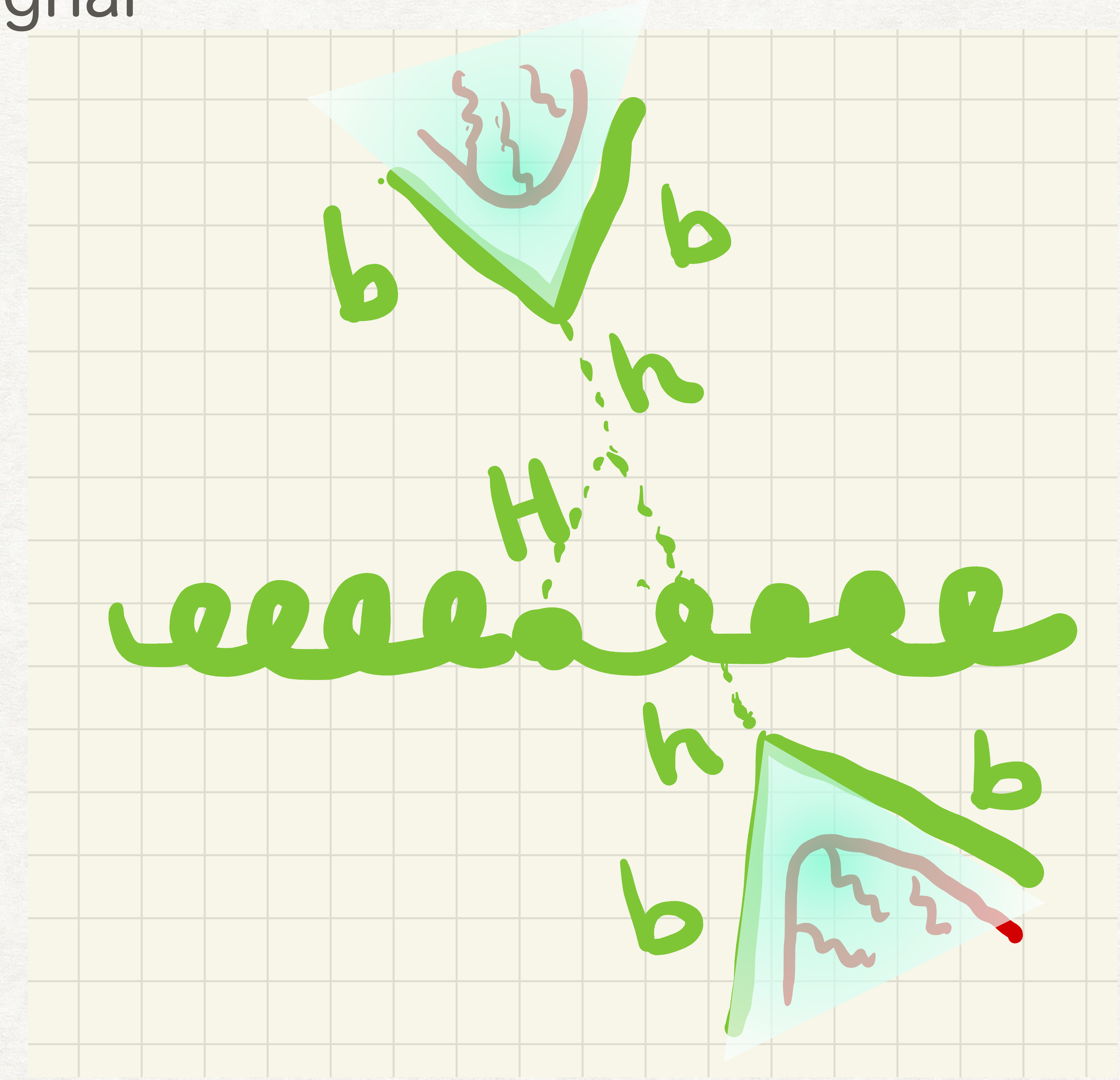
event color structure

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.

C. IAFormer (=InterAction transFormer)

3-1. Improvement of attention matrix.

Esmail, Hammad, Nojiri 2025.03258

original input for attention $\alpha = \mathrm{softmax}(\mathrm{QK}^{\mathrm{T}})$

particle information

- $P_4 = (p_x, p_y, p_z, E)$: particle 4-momentum
- $\Delta\eta = \eta - \eta_{\mathrm{jet}}$: pseudorapidity difference
- $\Delta\phi = \phi - \phi_{\mathrm{jet}}$: azimuthal angle difference
- $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$: angular distance from jet axis
- $\log(p_T)$: transverse momentum (GeV)
- $\log(E)$: energy (GeV)
- $\log\left(\frac{p_T}{p_{T_{\mathrm{jet}}}}\right)$: normalized p_T (GeV)
- $\log\left(\frac{E}{E_{\mathrm{jet}}}\right)$: normalized energy (GeV)

IAFormer attention $\alpha = \mathrm{softmax}(\mathscr{I}_{\mathrm{ij}})$

$\mathscr{I}_{ij} = W \cdot I_{ij}$

I_{ij} pairwize and boost invariant quantity

- $(p_{T_a} + p_{T_b})/p_{T_j}$
- $(E_a + E_b)/E_j$
- $\Delta = \sqrt{(\eta_a - \eta_b)^2 + (\phi_a - \phi_b)^2}$
- $k_T = \min(p_{T_a}, p_{T_b}) \cdot \Delta$
- $z = \min(p_{T_a}, p_{T_b})/p_{T_a} + p_{T_b}$
- $m^2 = (E_a + E_b)^2 - |\mathbf{p}_a + \mathbf{p}_b|^2$

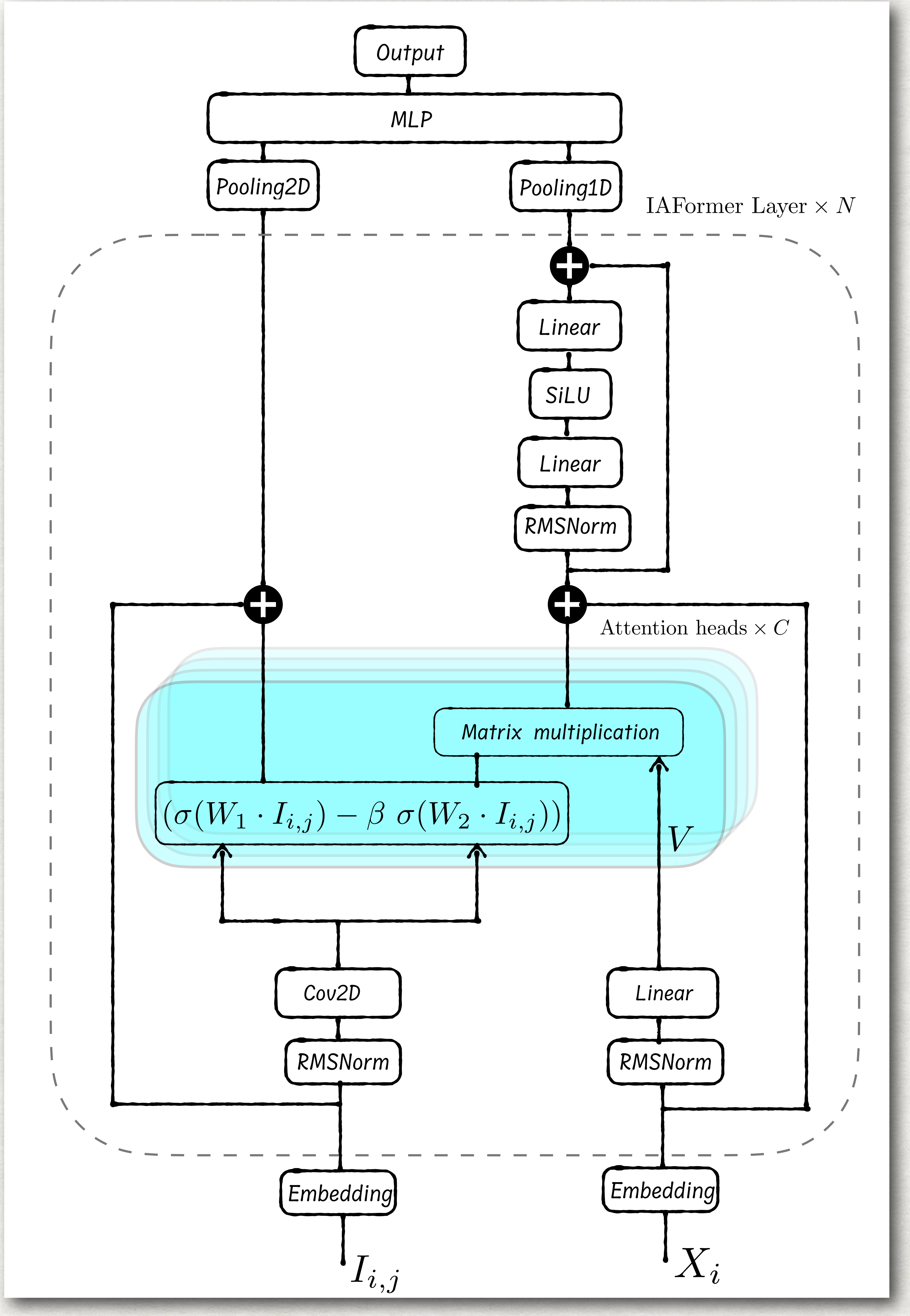
Structure of the IA-Former and results

X_i is updated by I_{ij} (cross attention)

I_{ij}, X_i are both updated by transformer.

Top vs QCD classification

	rejection efficience at 50% acceptance	
IAFormer	510 +/- 20	Tshi work
ParT	505 +/- 8	Base line
L-GATr	540 +/-20	Full Lorentz covariant



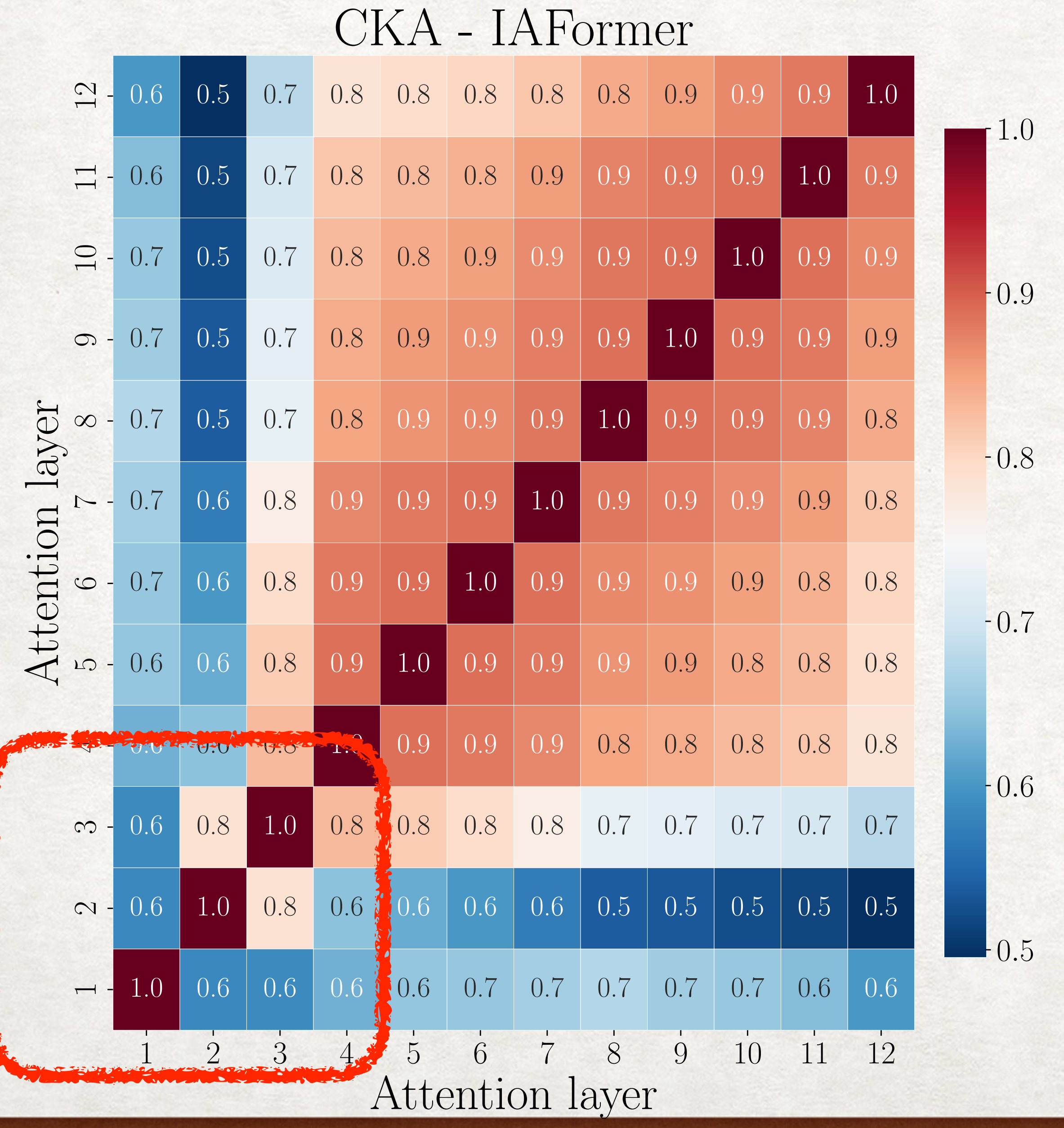
Learning pattern (CKA similarity)

d event-two layer output $X_1(d \times h_1)$ and $X_x(d \times h_1)$ → dxd matrix $M = X_1 X_1^T, N = X_2 X_2^T$. Then

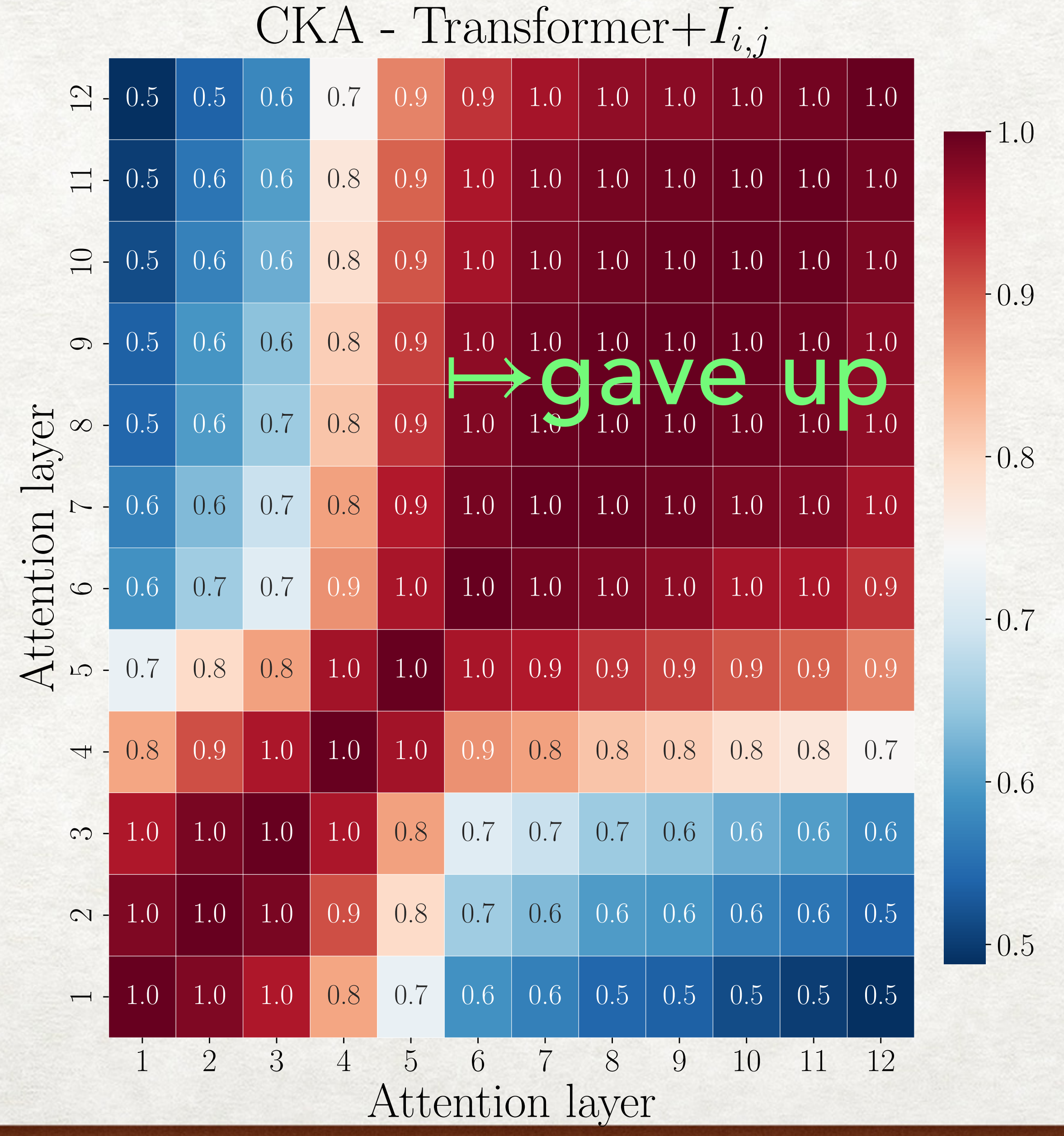
$$\mathrm{CKA(M,N)} = \frac{\mathrm{HSIC(M,N)}}{\sqrt{\mathrm{HSIC(M,M)HSIC(N,N)}}}, \qquad \mathrm{HSIC(M,N)} = \frac{1}{(d-1)^2}\mathrm{Tr}(MHNH) \qquad H = \delta_{ij} - \frac{1}{d}$$

if CKA~1, two layers are equivalent—and not needed.

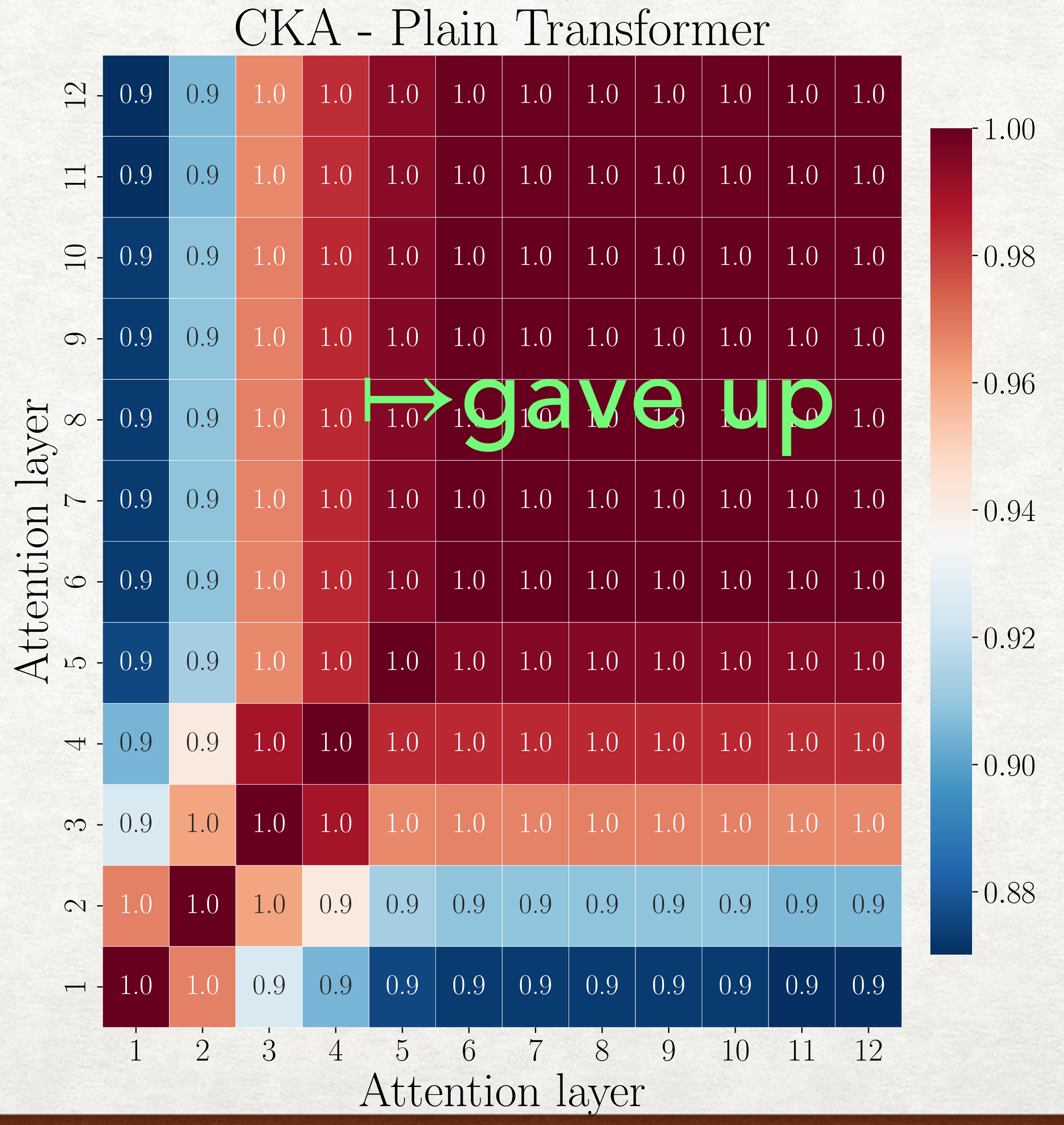
IAFormer($\epsilon_b(50\%) = 510$)



ParT $\epsilon_b(50\%) = 413$



Plain Transformer
$\epsilon_b(50\%) = 360$



IAFormer is learning efficiently

Differential attention(see arXiv:2410.05258)

- $\alpha = \mathrm{softmax}(\mathscr{I}) \to \alpha^{(i)} = \mathrm{softmax}(\mathscr{I}_1^{(i)}) - \beta^{(i)}\mathrm{softmax}(\mathscr{I}_2^{(i)})$

cancel the irrelevant information

$$\to \boxed{\mathscr{I}^{(1)}, \beta^{(1)}} \to \boxed{\mathscr{I}^{(2)}, \beta^{(2)}} \to \boxed{\mathscr{I}^{(3)}, \beta^{(3)}}$$

fixed sparse attention $\Rightarrow$ Each layer built different filters dynamically

STABILITY OF THE TRAINING—SPARSE ATTENTION

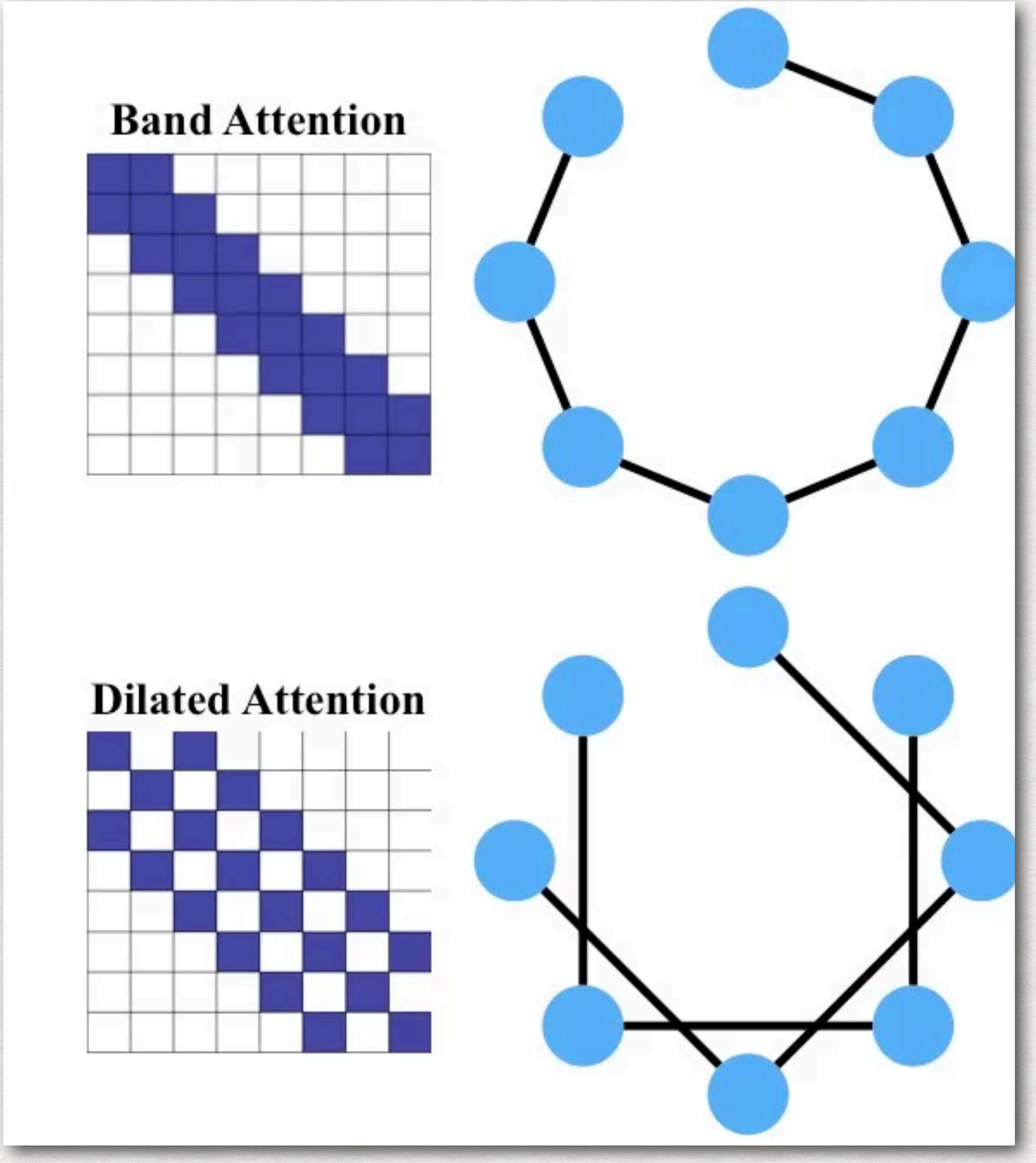
- Sparse attention: a rule to use only part of attention matrix for quick convergence and reasonings

- static attention—use “fixed patterns” to filter attention

- This is probably very important for Language model but does not looks right for particle physics.

    - **band attention** “大きなりんご big apple” “赤いりんご red apple”

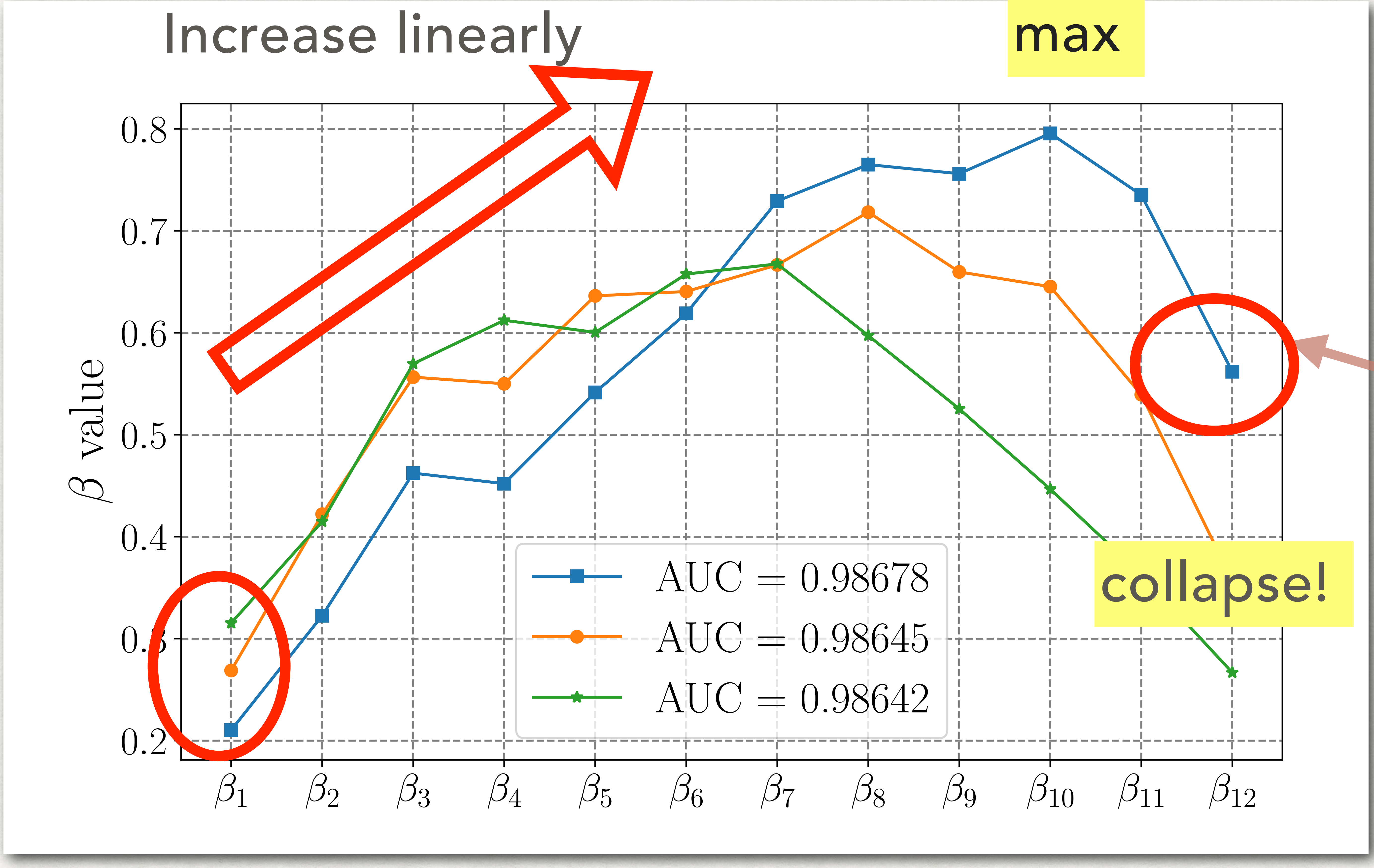
    - **Dilated attention** 赤いりんご が 落ちた のをみた (I saw a red apple falling)



https://developers.agirobots.com/jp/sparse-attention/

BEHAVIOR OF SPASE ATTENTION DYNAMIC FILTERS

filtered information



start from
small beta

Networks minimize finite positive β. We need filters

last β is large

→ higher network performance

TAKE AWAY MESSAGE

1. fast, lightweight, while keeping performance
2. Incorporate physics picture

   RESPCETING QCD
3. Jet analysis → event analysis.(H→hh)

   Cross attention is important
4. Respect symmetry Replacing “attention from generic features” →“pairwise boost invariant information “ (IAFormer)

   Symmetry
5. Reduce valiance in training

   Improved stability within DL
6. Identify the key parameters for classifications

   Identify Important variables in DL era
   Improving MC simulation