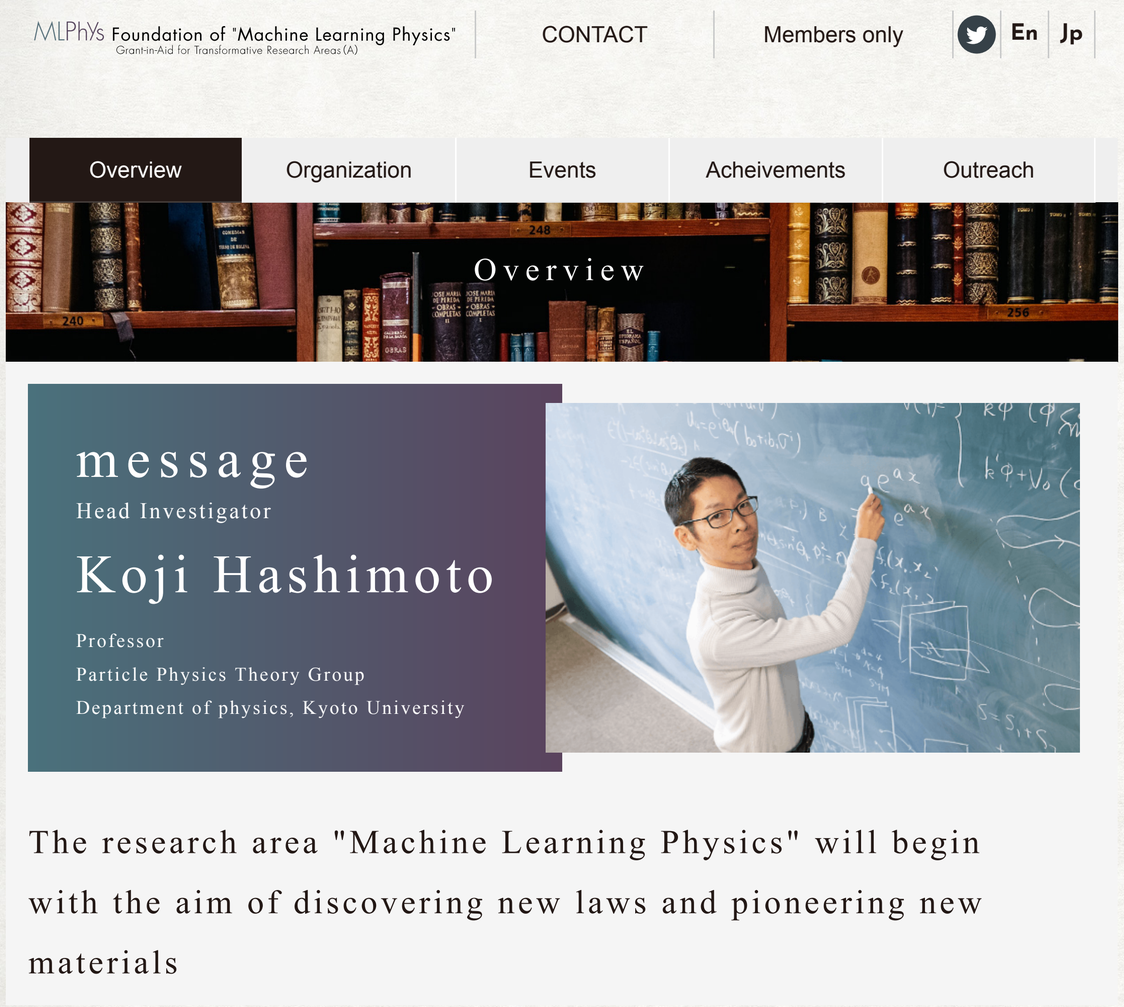
DESIGNING TRANSFORMER FOR PARTICLE PHYSICS

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

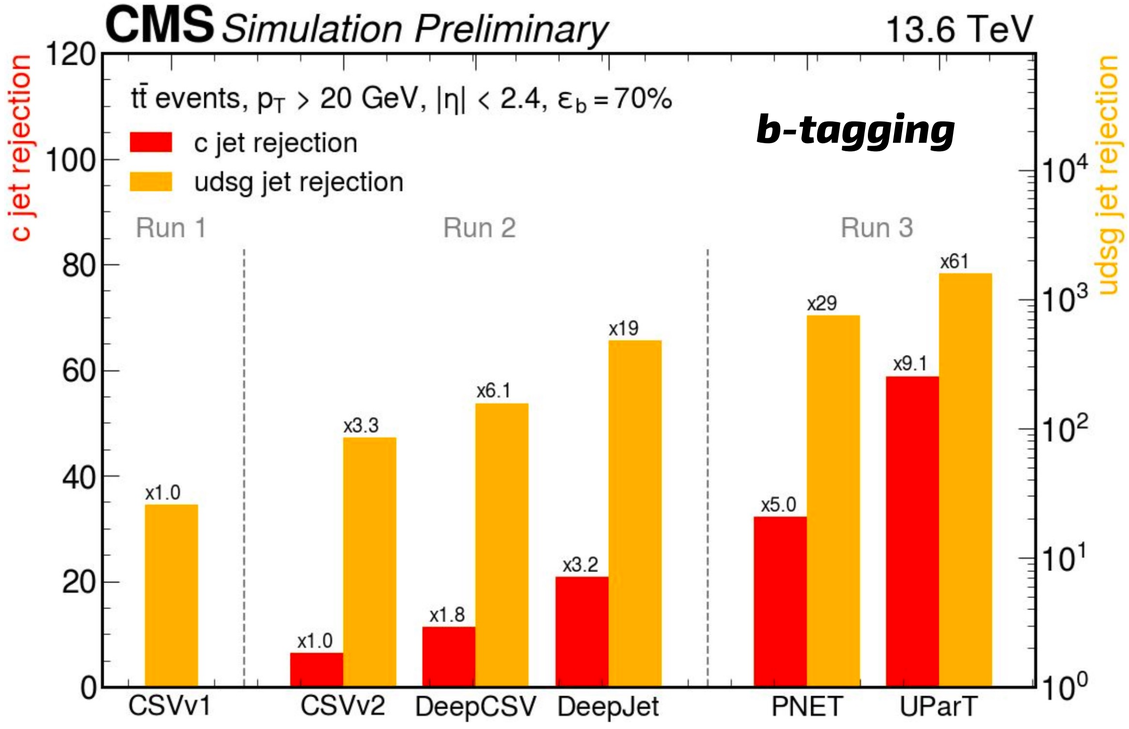
PD. Ahmed Hammad
2017-2020: Ph.D Basel University, Basel Switzerland
2020-2023: SeoulTech, Korea
2023- KEK

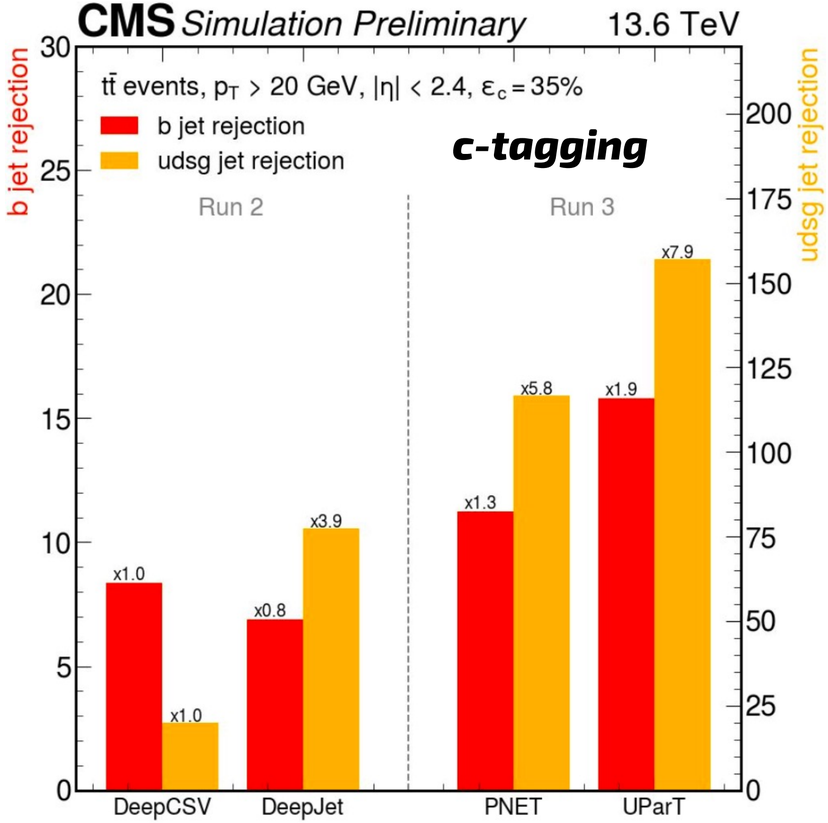
Flavor Tagging Performance

CMS-DP-2024-066

b/c-tagging performance

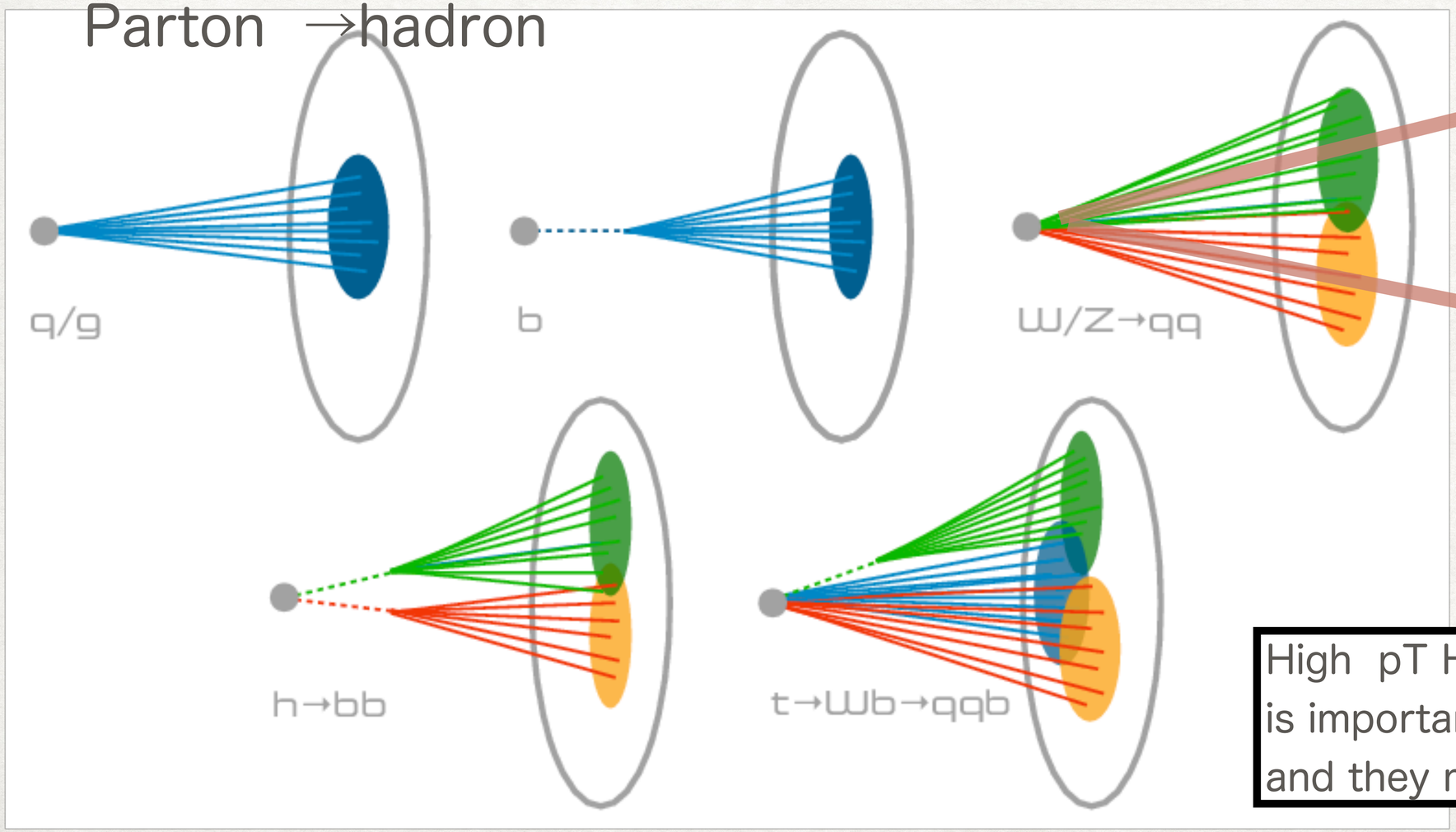
- Promising performance compared to previous taggers
  - ×3 better light jet rejection (at b-jet eff 70%) than DeepJet
  - ×2 better light rejection + ×2 better b-jet rejection (at c-jet eff 35 %) than DeepJet





JET TAGGING: WINDOW TO THE NEW PHYISICS

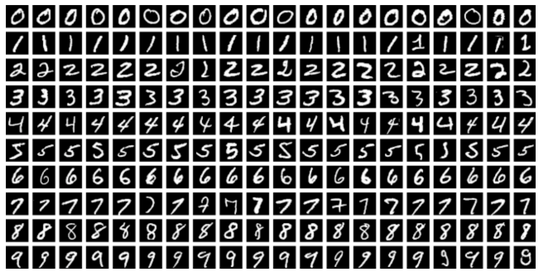
Mostly talking about top vs QCD classification in this talk



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

Training for classification

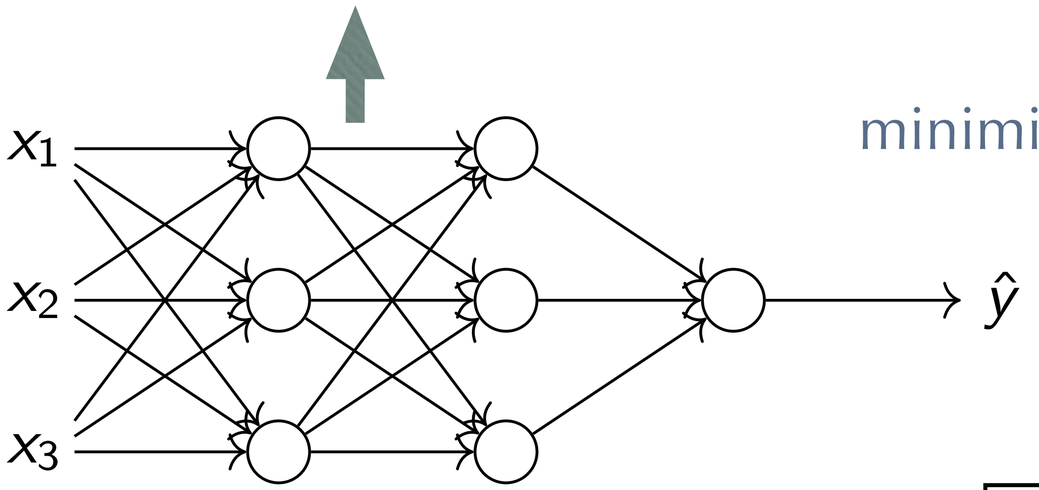
data pool of images



(28x28) の画像データをn 個

TRANSFORMER
GNN

training: $\mathbf{w_{ij}}, \mathbf{b_i}$



minimization of loss function

$$L(y, \hat{y})$$

y: truth
1→(1, 0, 0, ..)
2→(0, 1, 0,⋯)
3→(0, 0, 1, ⋯)

output

$$\hat{y} = (\hat{y}_1, \hat{y}_2 \cdot \cdot \cdot \cdot \cdot \hat{y}_{10}), \sum_i \hat{y}_i = 1$$

$$\hat{y}_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

$\hat{y}$'s represent likeliness to be y

✅表現力　expressive power

✅データを学習　learn from data

✅微分可能Simple linear algebra + activation

1. “TRANSFORMER” :SELF ATTENTION

X: n(#particle in the jet) x (feature of particle)

output size = input size

Transformation before MLP

$$X' = X + \delta X,\ \delta X = \alpha \cdot V \equiv \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \qquad Q = XW_Q, K = XW_K, V = XW_V$$

Pro

1.Attention Matrix $\alpha = QK^T$ All particle-particle attention calculated at once, while GNN takes care nearby information only

2. All feature can be included.

3.Each layer produce X^3 effect , while MLP need three steps to take into acctound the effect.

4. W can be chosen so that X and δX can be added. Freedom choose attention pair.

5.skip connection-no information loss X → X’→X’’

Con

Tough for memory, calculation. Physics may allow more compact descriptions.

Features in the context of jet classification

Particle momentum

charge,particle ID

displaced vertex

Category	Variable	Definition	JET
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	
	$\log \frac{p_{\mathrm{T}}}{p_{\mathrm{T}}(\mathrm{jet})}$	logarithm of the particle's p_{T} relative to the jet p_{T}	
	$\log \frac{E}{E(\mathrm{jet})}$	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	

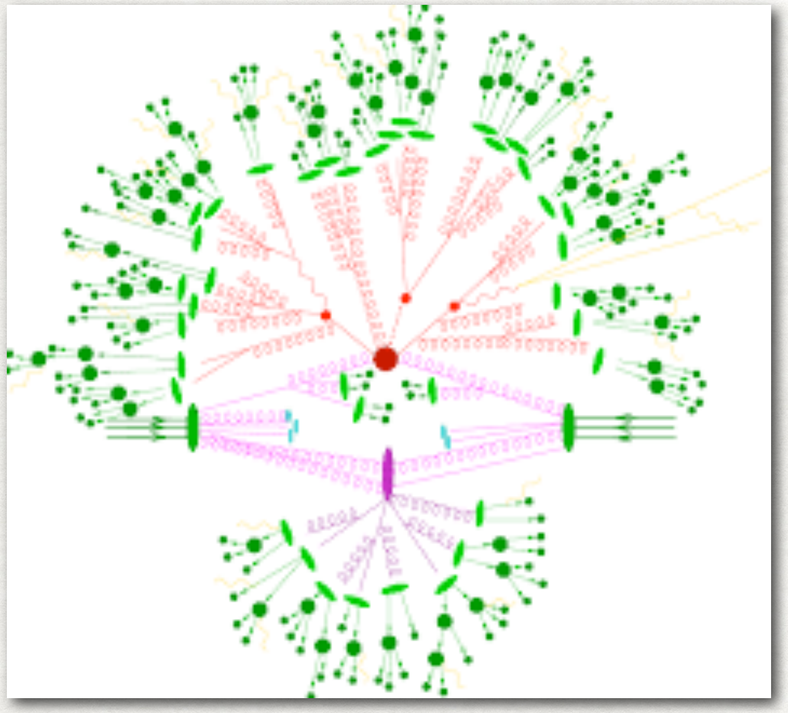
THIS TALK: IMPROVEMENT OF THE PARTICLE TRANSFORMER

	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**

Yellow bands highlight our works!

1. TRANSFORMER FOR PARTON SHOWER+HADRONIZATION

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

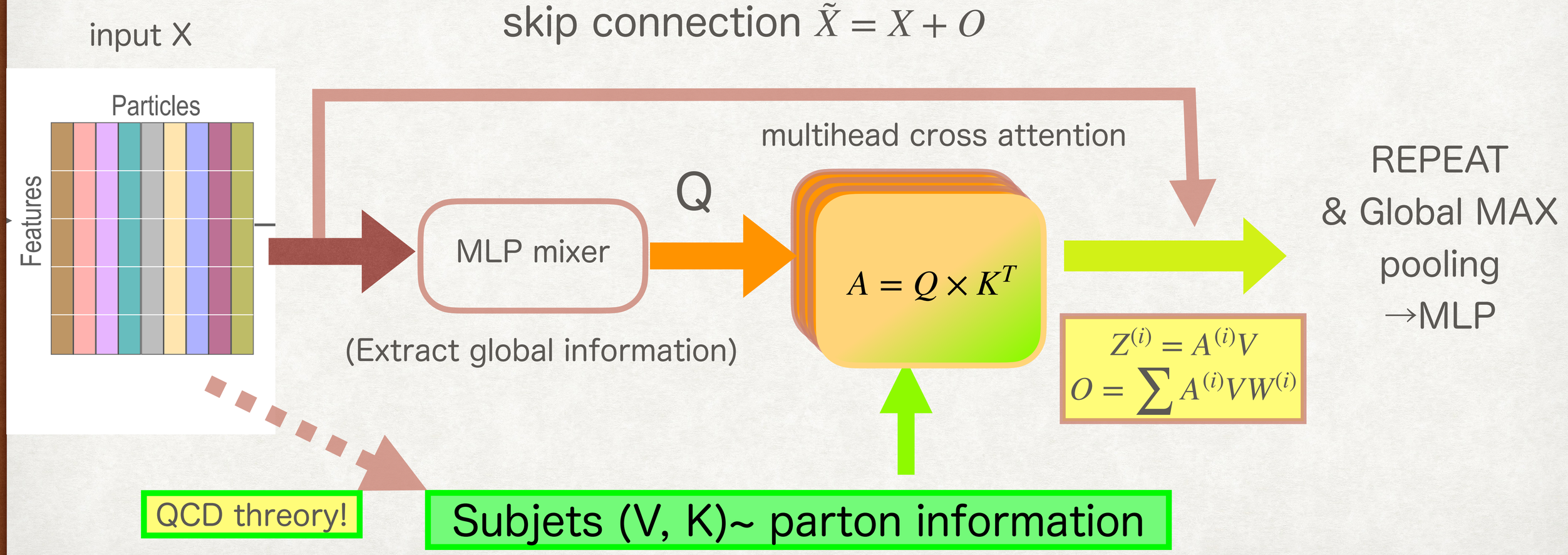


- **Hard Process** = Partons(quarks and gluons) {y}
- **a jet:** P(hadrons in jets | parton ~ jet) = $P(\{x_i\}\,|\,\{y\})$
- **jet with substructure** $P(\{x_i\}\,|\,\{y_\alpha\})$
- Maybe **several fatjets** in an event (factorization)

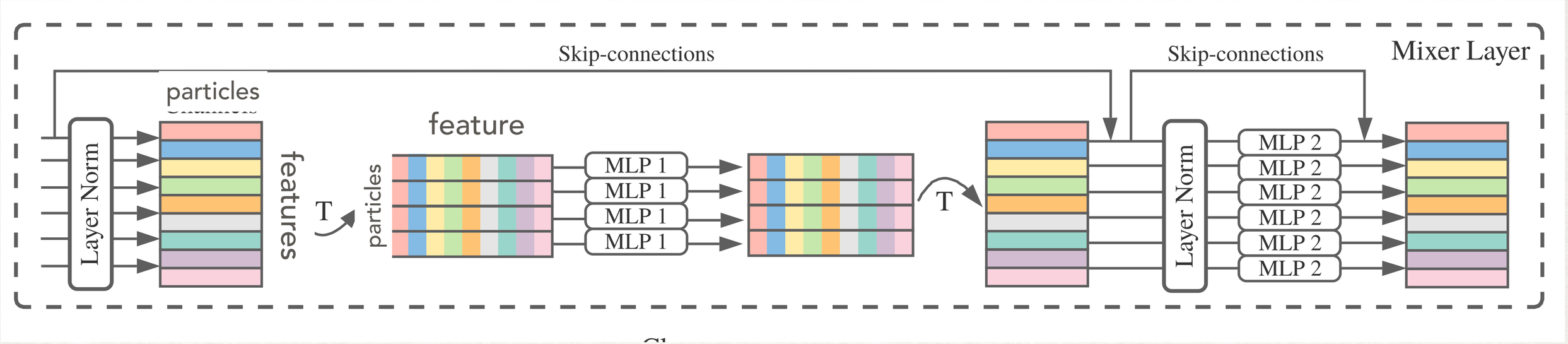
$$P(\{x_i\},\{x'_j\},\{y_\alpha\},\{y'_\beta\}) \sim P(\{x_i\}\,|\,\{y_\alpha\})P(\{x'_i\}\,|\,\{y'_\beta\})\;P(\{y_\alpha\,,y'_\beta\})$$

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

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



MLP MIXER



MLP 1 : operate on features

MLP 2: operate on particles

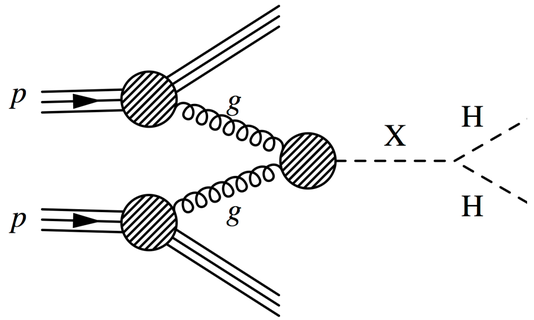
THE PERFORMANCE FOR TOP VS QCD CLASSIFICATION

	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**

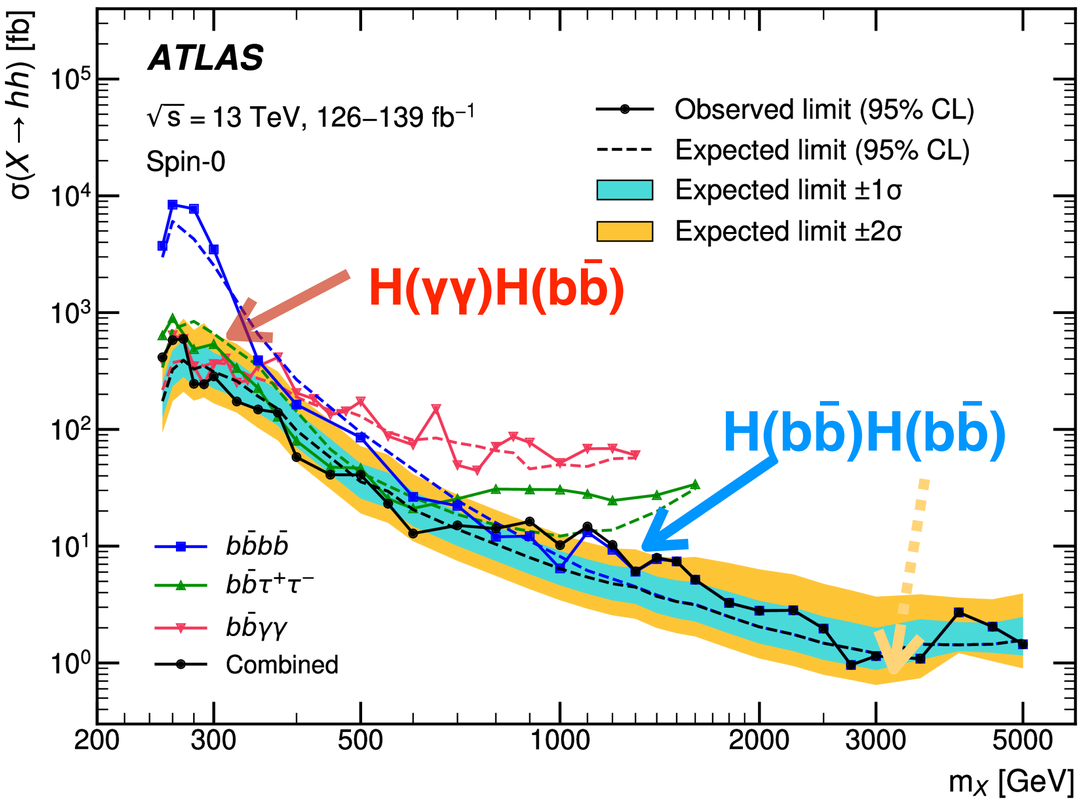
Yellow bands highlight our works!

2. GLOBAL EVENT ANALYSIS

X→HH



Phys. Rev. Lett. 132 (2024) 231801



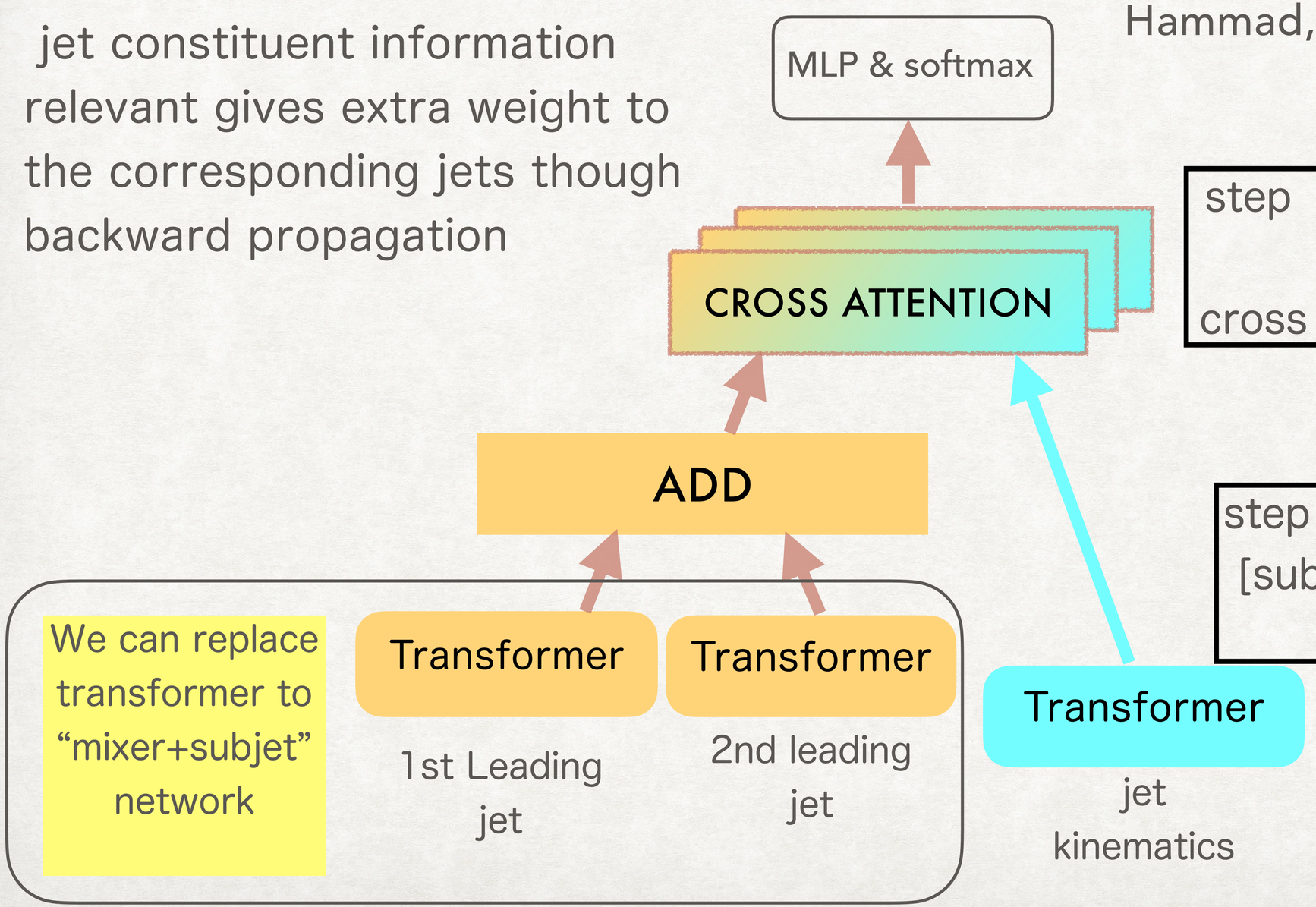
H(bb̄)H(bb̄) most sensitive channel for $m_X > 400/500$ GeV

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

cross attention for 2 fatjet events

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

Hammad, Moretti, MN JHEP 03, 2024



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

step 1 : multihead self attention
[substructure]x[substructure]
[jet kin]x [jet kin]

3. IAFormer (=InterAction transFormer)

3-1. Improvement of attention matrix.

Esmail, Hammad, Nojiri 2025.03258

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

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 $\alpha = \mathrm{softmax}(\mathcal{I}_{\mathrm{ij}})$

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

I_{ij} pairwize and boost invariant quantity

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

3-2 Introduction of Differential attention

- We have introduced a new dynamic attention called “differential attention” to the network. (see arXiv:2410.05258)
  - $\alpha = \mathrm{softmax}(\mathcal{I}) \rightarrow \alpha^{(i)} = \mathrm{softmax}(\mathcal{I}_1^{(i)}) - \beta^{(i)}\mathrm{softmax}(\mathcal{I}_2^{(i)})$

cancel the irrelevant information

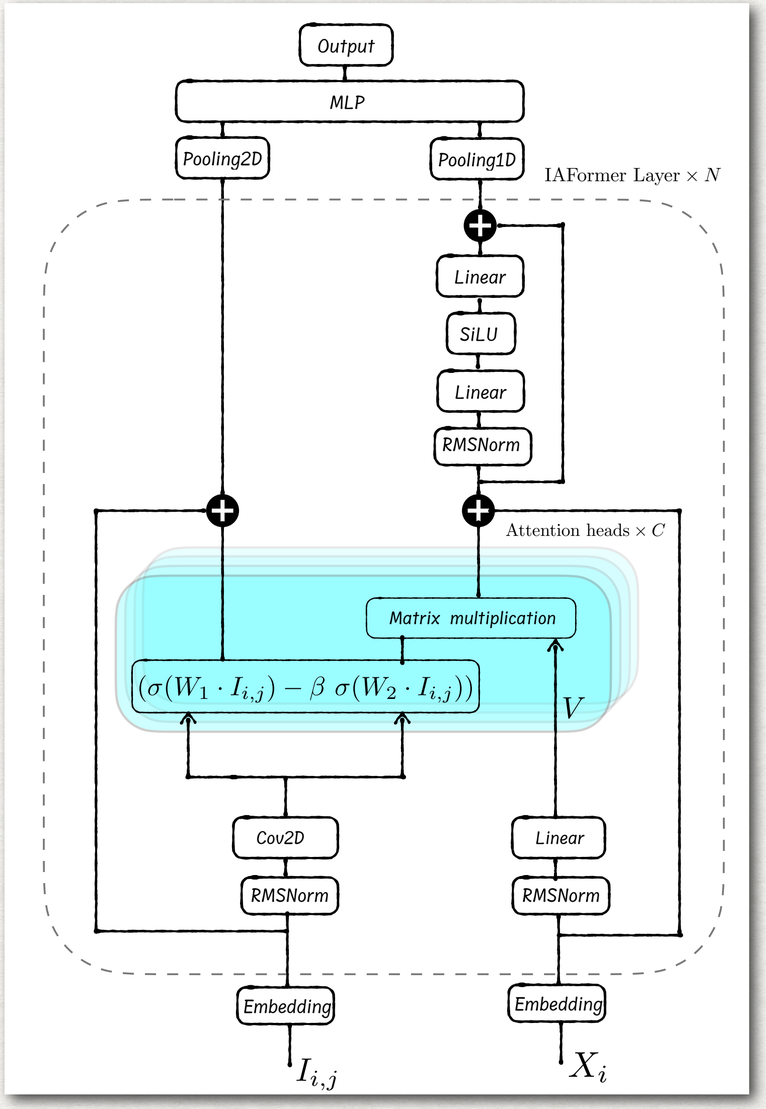
fixed sparse attention

Each layer built different filters dynamically

I_{ij}, X_i are both updated by skip connection(not like PART)

For particle transformer, Attention build from transformed X, but in our network, they develop independently.

Not sure about boost invariance is kept along with layers but leading term is OK….



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**

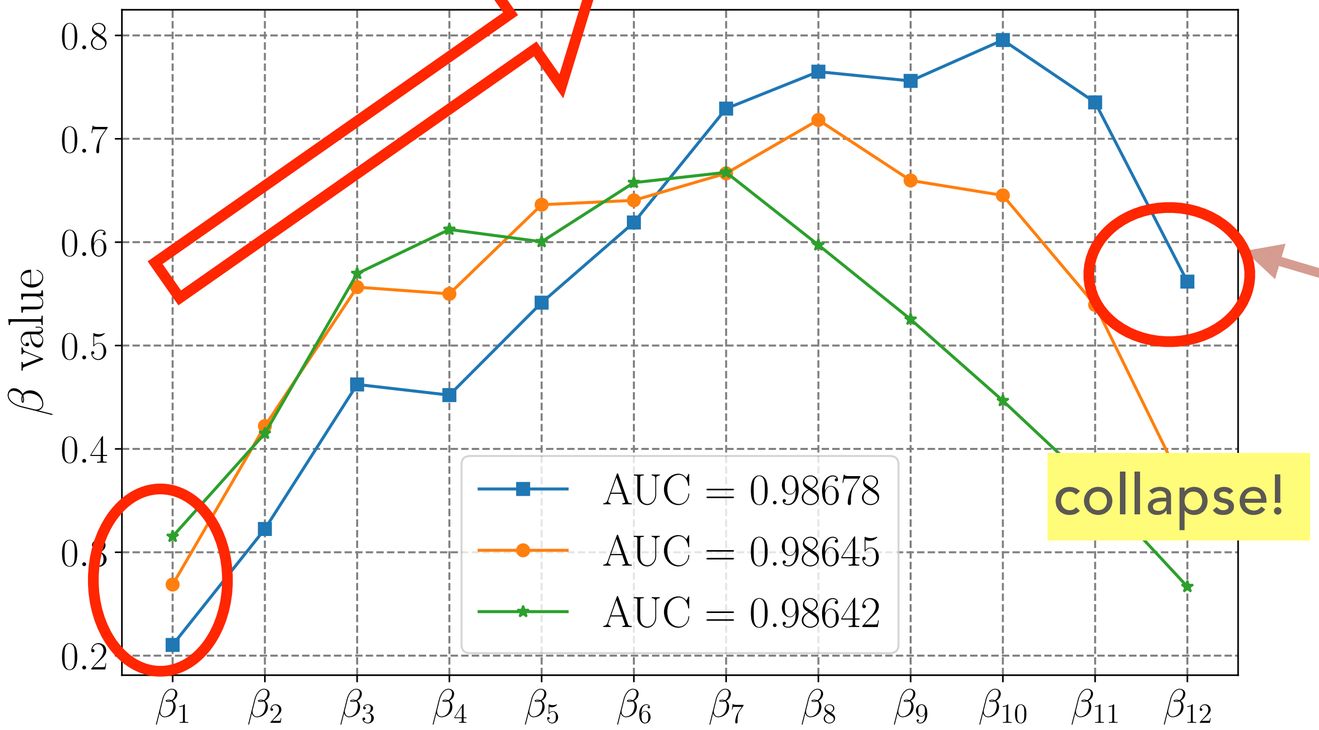
Yellow bands highlight our works!

3-4 RESULTS: BEHAVIOR OF DYNAMIC FILTERS

filtered information

Increase linearly

max



collapse!

start from small beta

Networks minimize finite positive β. We need filters

last β is large

→ higher network performance

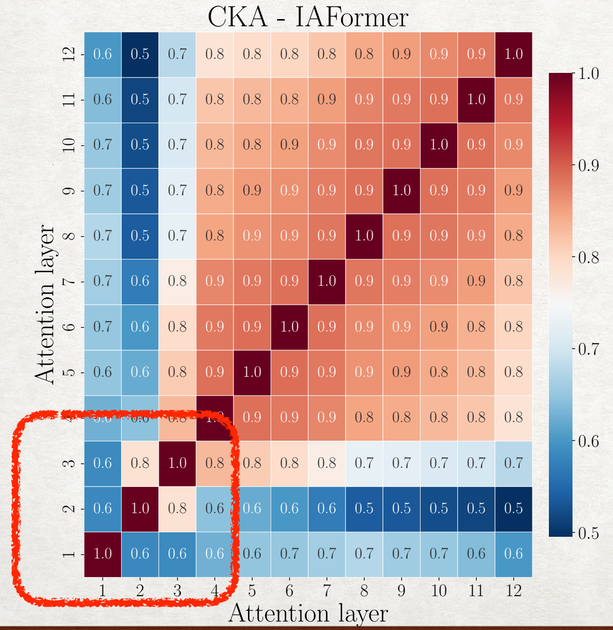
3-4 RESULT :Learning pattern analysis CKA similarity

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

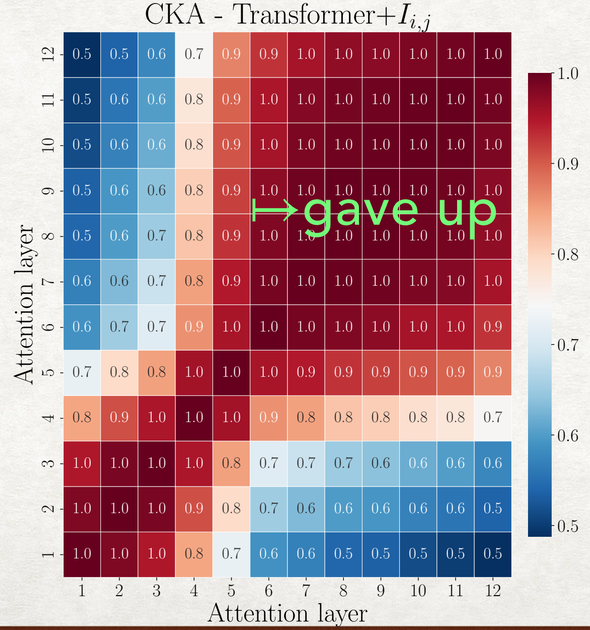
$$\mathrm{CKA(M,N)} = \frac{\mathrm{HSIC(M,N)}}{\sqrt{\mathrm{HSIC(M,M)HSIC(N,N)}}}, \qquad \mathrm{HSIC(M,N)} = \frac{1}{(d-1)^2}\mathrm{Tr}(MHNH) \qquad H = \delta_{ij} - \frac{1}{d}$$

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

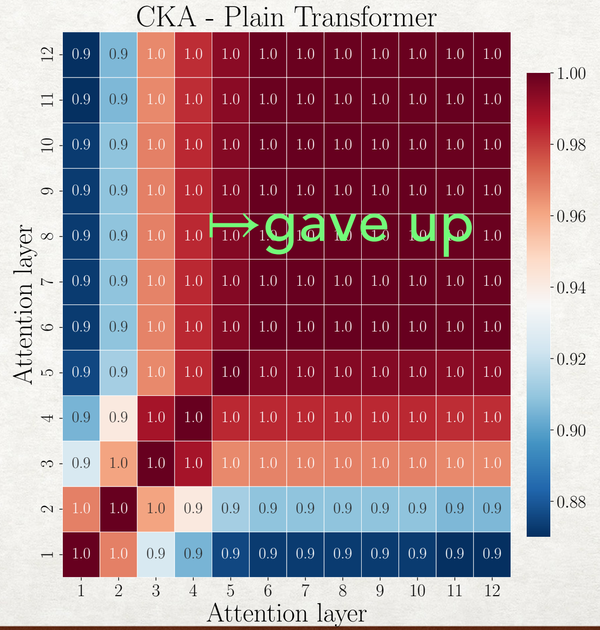
IAFormer($\epsilon_b(50\%) = 510$)



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



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

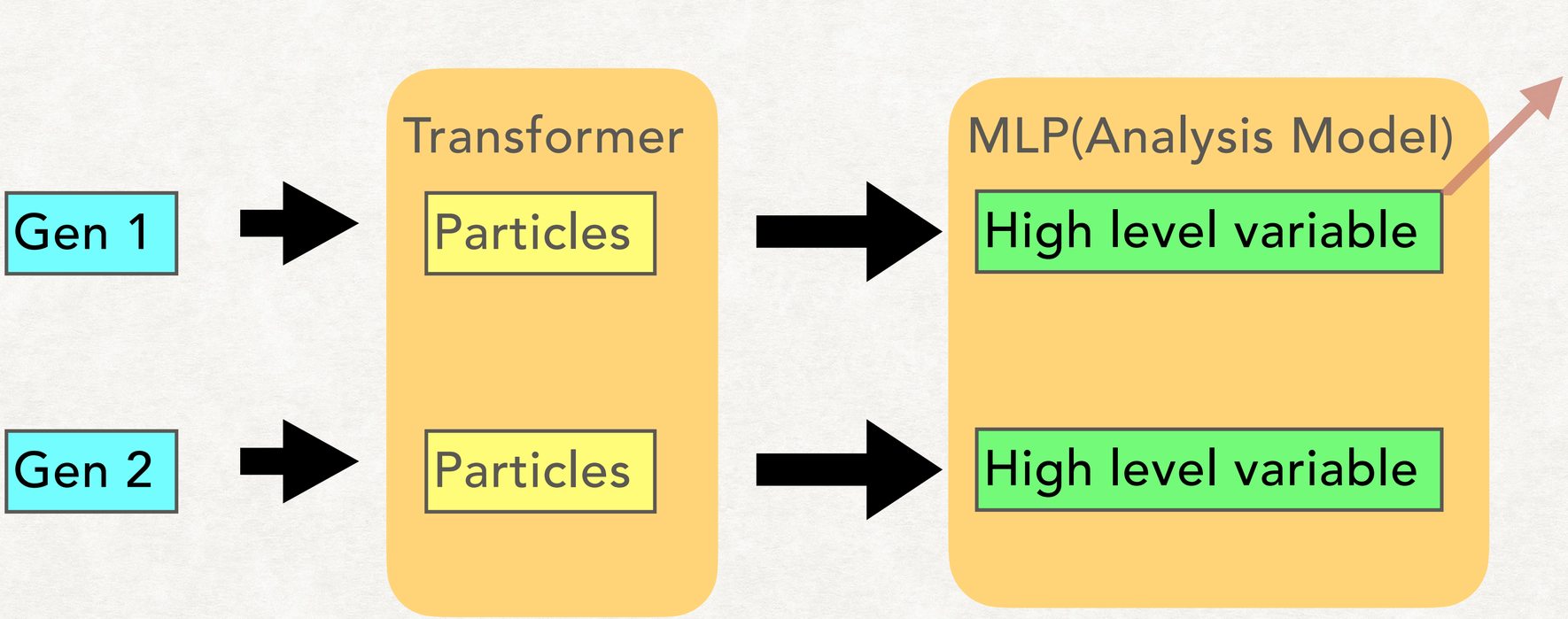


IAFormer is learning efficiently

SUMMARY: OUR MODEL: CROSS ATTENTION

	Key	Query	Value	update
Particle transformer	particle	particle	particle	particle
“Mixer”	subjet	particle	subjet	particle
“Global analysis”	particle	jet	jet	jet
“IAFormer”	boost invariant pairwise quantity		particle	particle

4. ML FOR GENERATOR COMPARISON



$s_{gen}(x)$: generator classifier output

$$w = \frac{s_{gen}(x)}{1 - s_{gen}(x)}$$ estimated probability ratio

Transformer: no human bias, poor training stablity

Analysis Model (AP):MLP using highlevel input : stable prediction, ideal for generator comparison

HL feature for generator classification

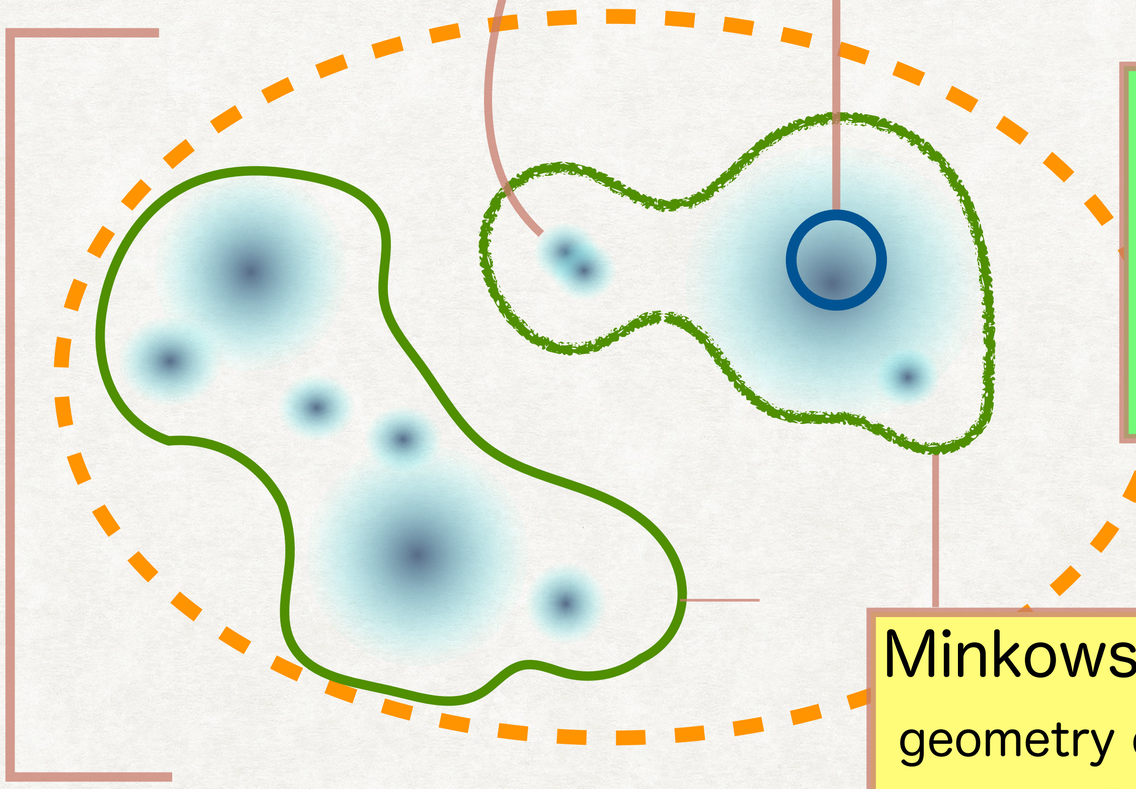
Amon Furuichi, Sung Hak Lim, Mihoko M. Nojiri JHEP 07 (2024) 146 JHEP 07(2025) 111

pt distribution of constituents

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

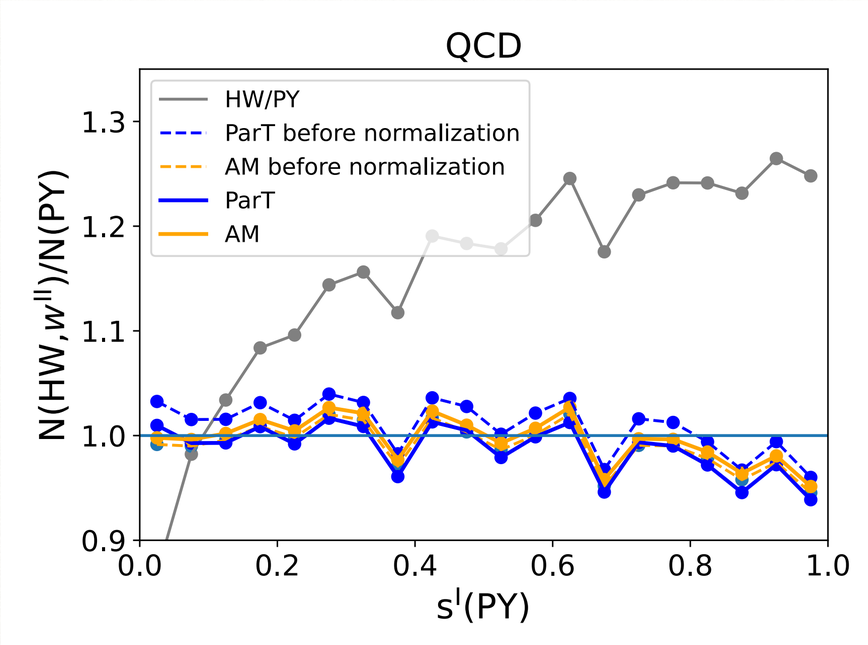
geometry of jet cosntituent distribution

PYTHIA VS HERWIG UNDER ML

$s_{gen}(x)$: generator classifier output

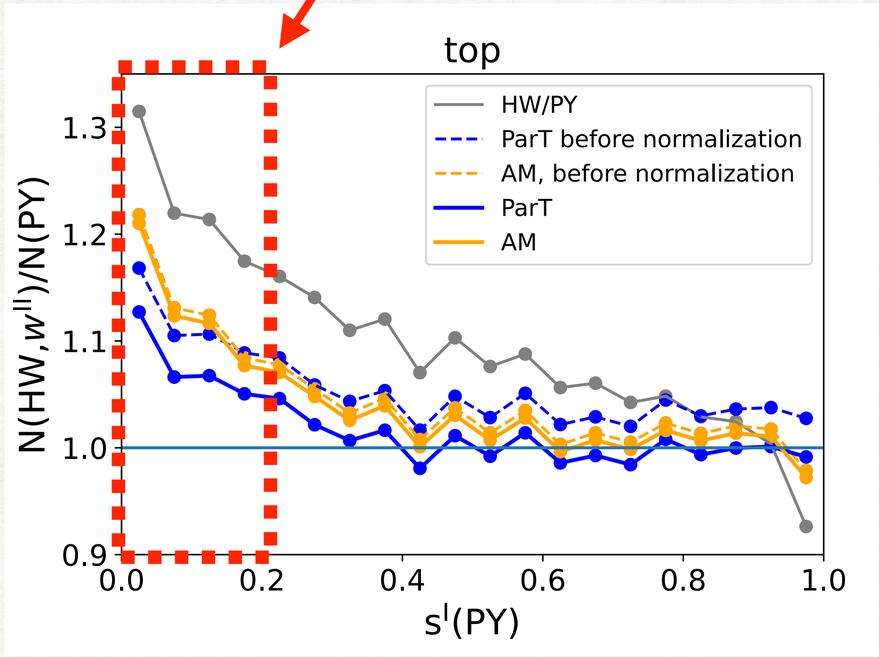
$$w = \frac{s_{gen}(x)}{1 - s_{gen}(x)}$$ estimated probability ratio

agreement of reweighted HW distribution to PY quantify DL infering



QCD-like ⟷ top-like

poor overlap between PY vs HW in QCD-like top jet



QCD-like ⟷ top-like

HL feature for generator classification

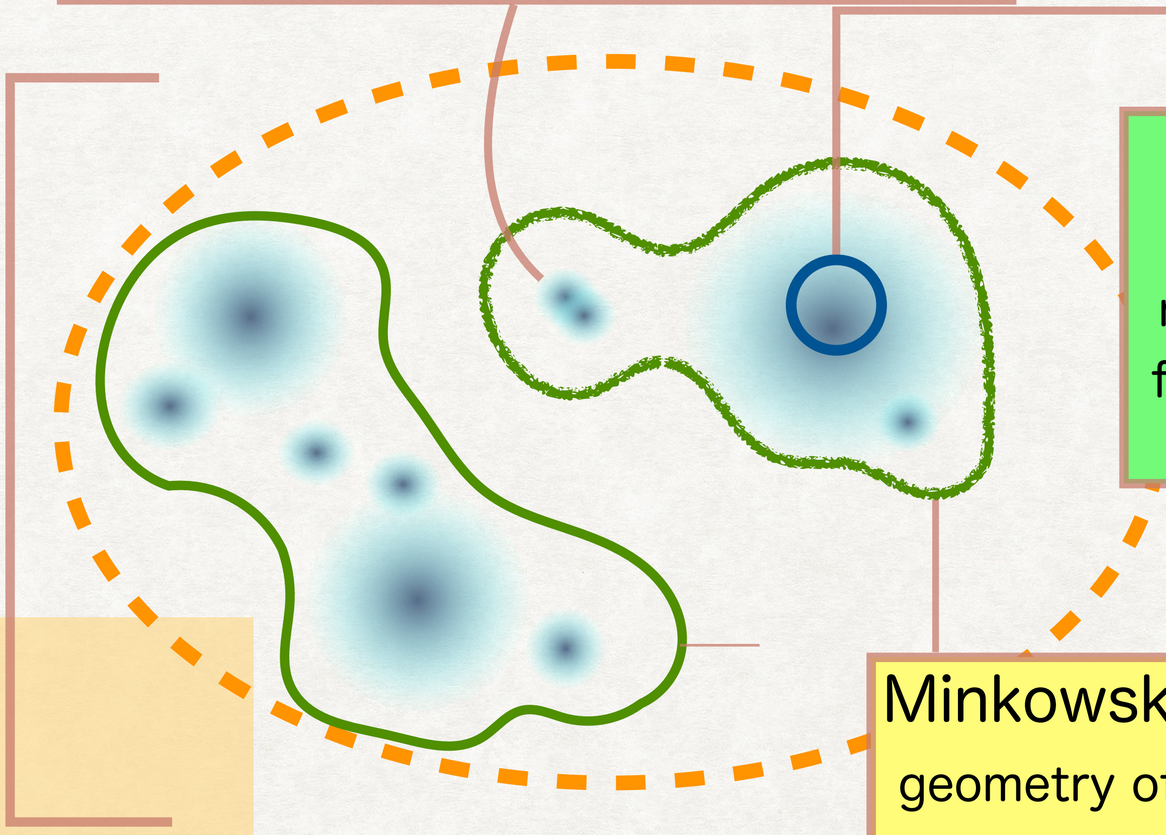
Amon Furuichi, Sung Hak Lim, Mihoko M. Nojiri JHEP 07 (2024) 146 2312.11760

pt distribution of constituents

Jet spectrum

two point Energy correlation (unlocalized sampling)

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



Subjet

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

Energy flow polynomials

$$\mathrm{EFP}_G^{(\kappa,\beta)} = \sum_{i_1\in J}\cdots\sum_{i_N\in J} z_{i_1}^{(\kappa)}\cdots z_{i_N}^{(\kappa)} \prod_{(m,\ell)\in G} \theta_{i_m i_\ell}^{(\beta)}.$$

Minkowski Functionals

geometry of jet cosntituent distribution

ENERGY FLOW POLYNOMIAL AND TOP JET CLASSIFICATION

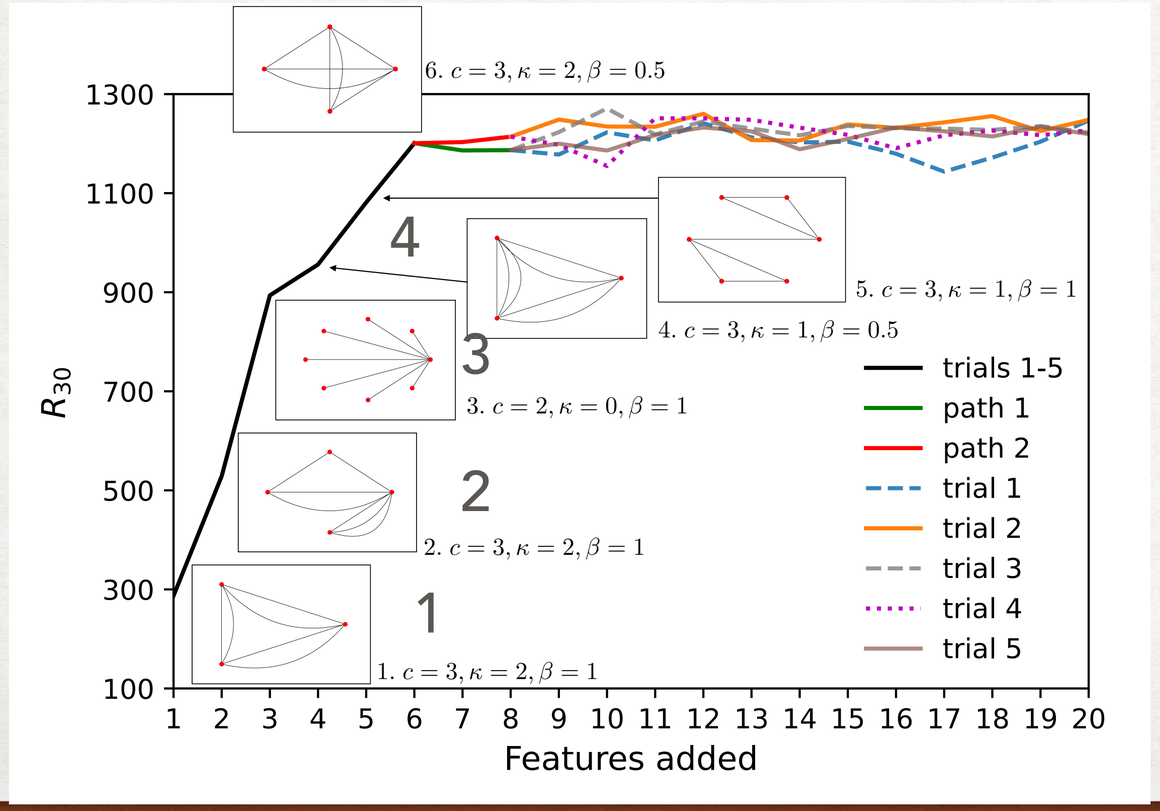
"Feature selection with distance correlation",Ranit Das, Gregor Kasieczka, David Shih

Phys.Rev.D 109 (2024) 5, 054009

Energy flow polynomial

$$\mathrm{EFP}_G^{(\kappa,\beta)} = \sum_{i_1 \in J} \cdots \sum_{i_N \in J} z_{i_1}^{(\kappa)} \cdots z_{i_N}^{(\kappa)} \prod_{(m,\ell)\in G} \theta_{i_m i_\ell}^{(\beta)}.$$

take all possible combination of N constituent and weight by common power k for energy of the particle and common power β for angle according to the graph

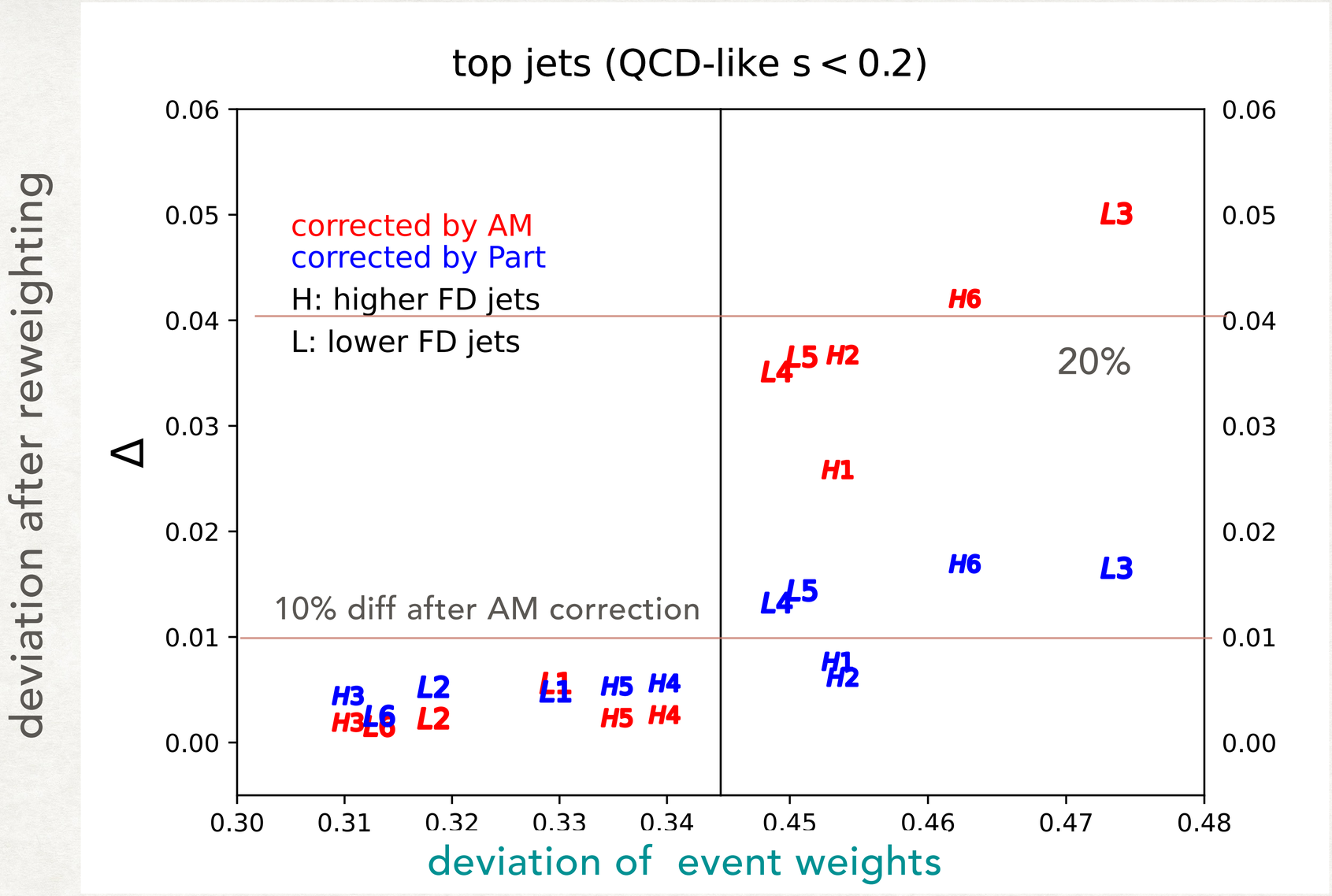


After reweigting by AM Pythia and HW difference remains for "QCD like top event" with

Higher FD1, FD2, FD6 ($\kappa = 2$, 3 or 4 points)

Low FD3, FD4, FD5 ($\kappa = 0,1$ θ power ~3.5,7, 8)

corresponding to narrow jet where a few jet constituents take most of jet energy



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

smaller overlap between PY & HW evenets

TAKE AWAY MESSAGE

1. fast, lightweight, while keeping performance

RESPCETING QCD

2. Incorporate physics picture

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