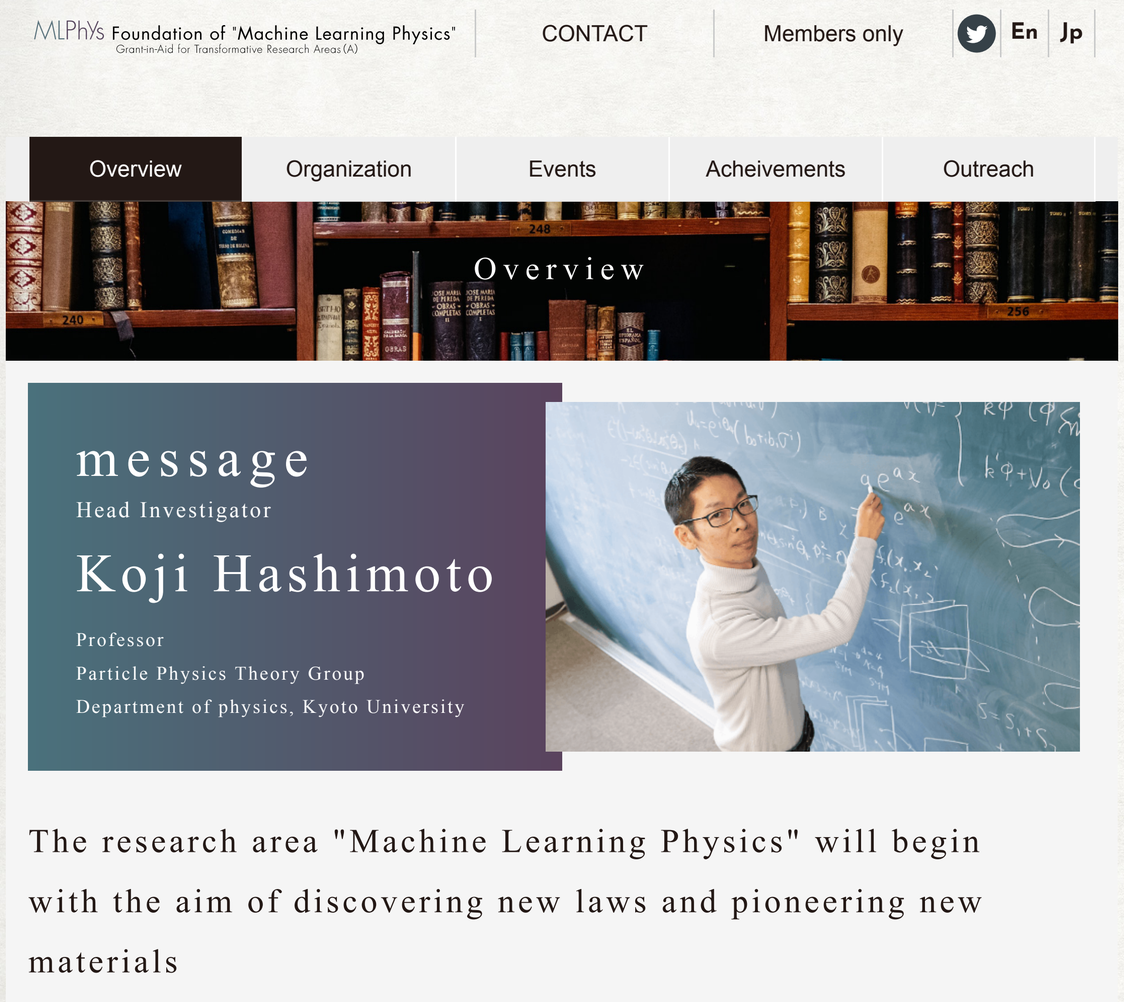
TRANSFORMER FOR PARTICLE PHYSICS

MIHOKO NOJIRI
(KEK)

with Ahmed Hammad and Sung Hak Lim

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

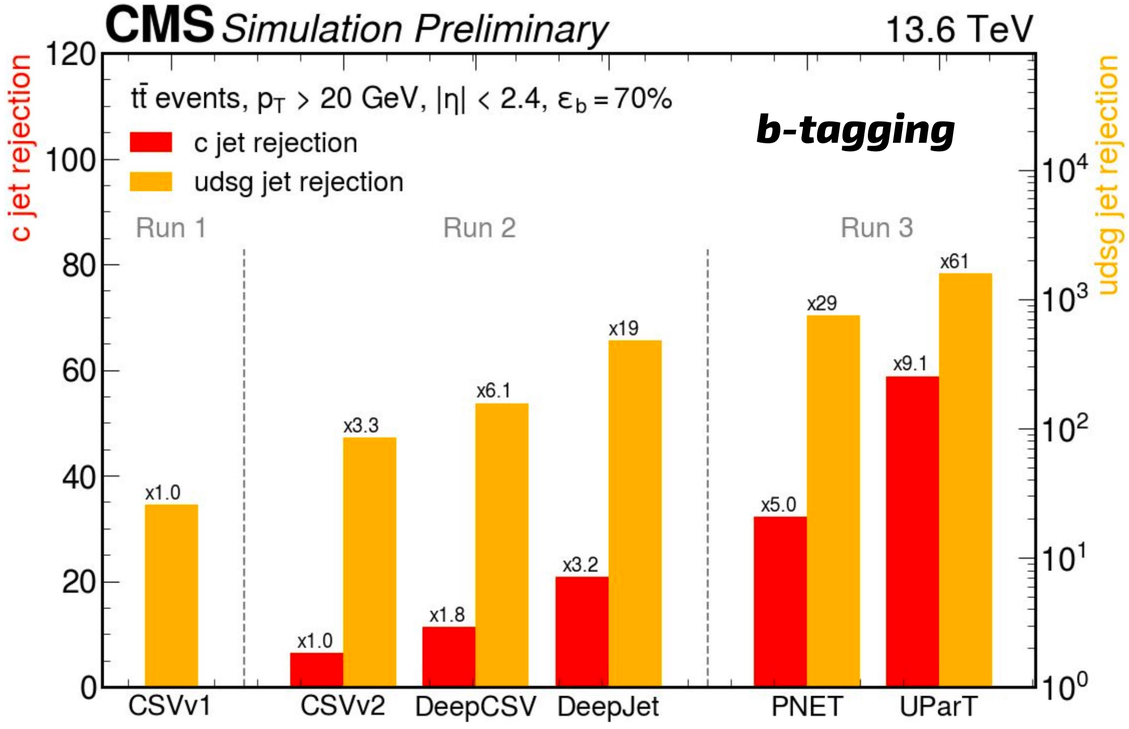
MLPhysics PD. Ahmed Hammad
2017-2020: Ph.D Basel University, Basel Switzerland
2020-2023: SeoulTech, Korea
2023- KEK

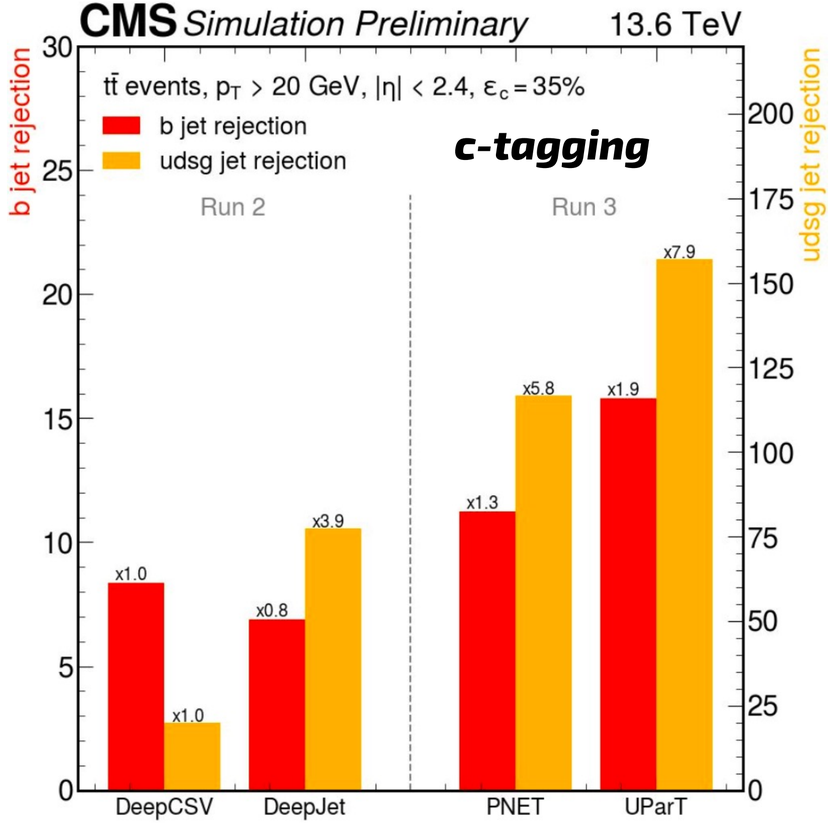
Flavor Tagging Performance

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

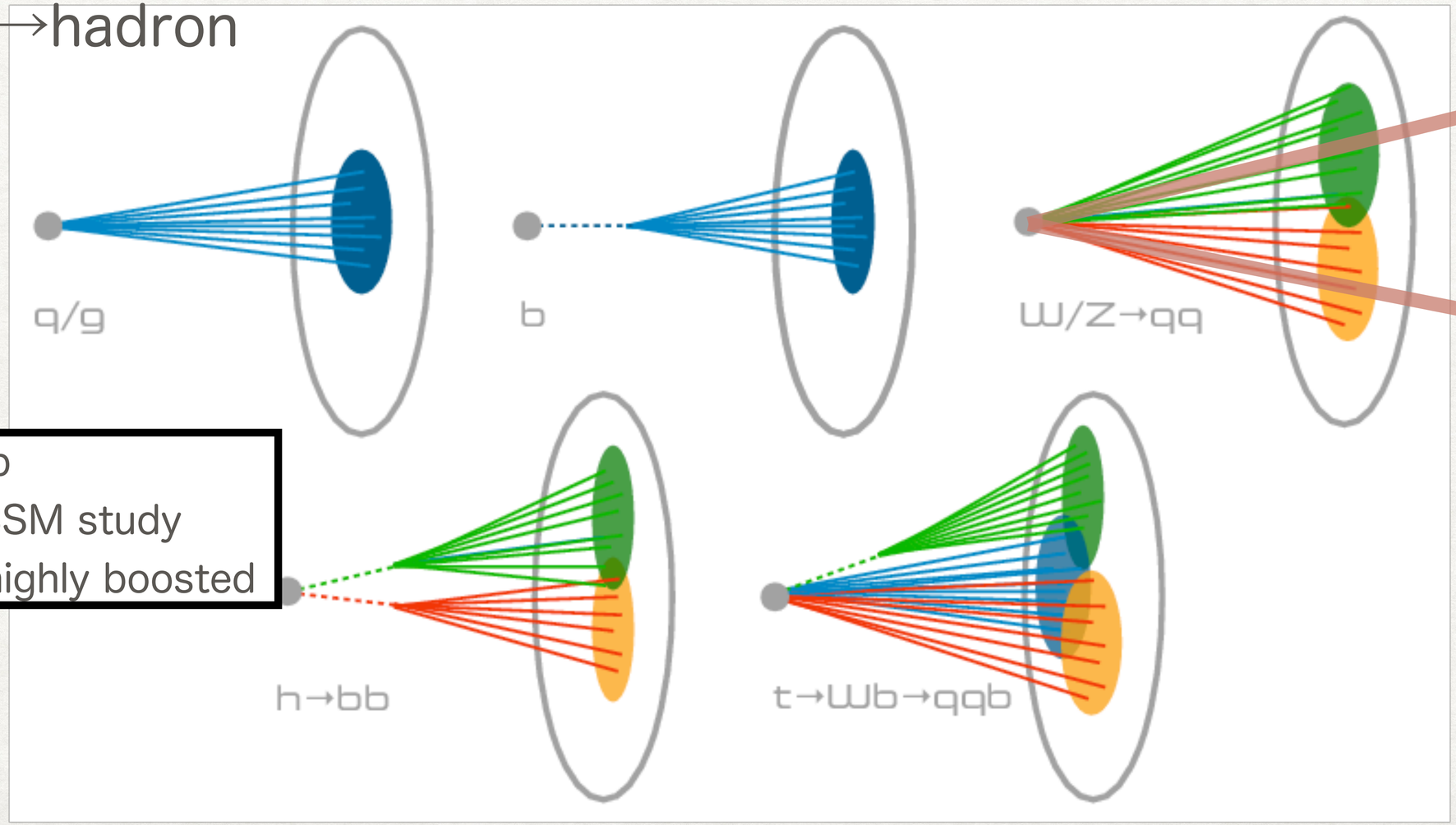




JET TAGGING: WINDOW TO THE NEW PHYISICS

Mostly talking about top QCD classification in this talk

Parton →hadron

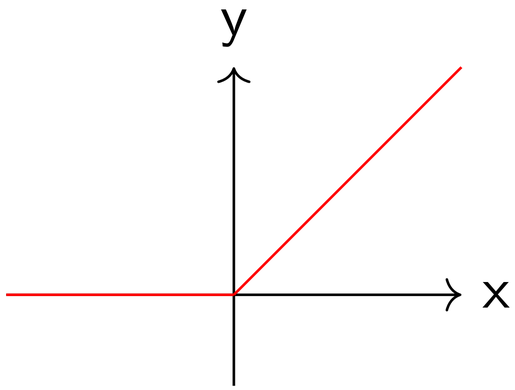


High pT H, Z, top
is important for BSM study
and they maybe highly boosted

Basic ingredient of ML

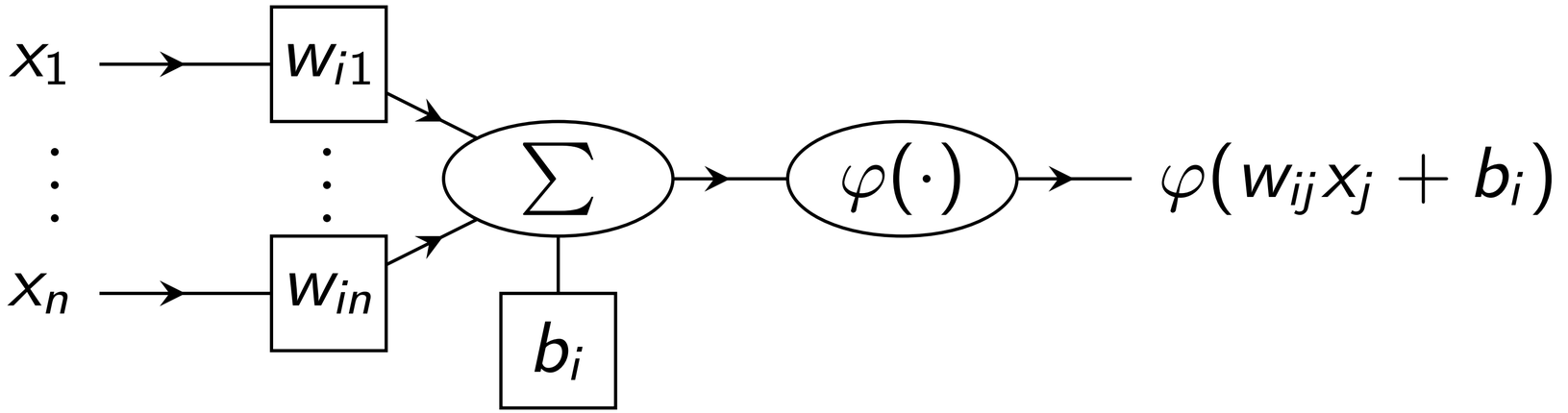
$$x_i^{(K+1)} = \varphi(w_{ij}^{(K)} x_j^{(K)} + b_i^{(K)})$$

φ: activation function —-for non lineality



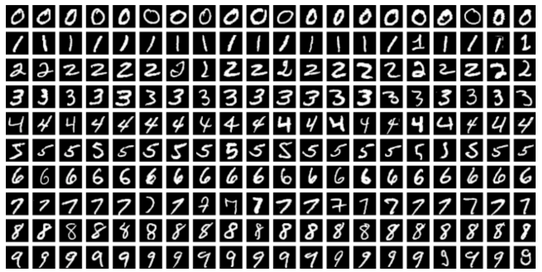
Rectified Linear Unit(ReLU)

$$\varphi_{\mathrm{ReLU}}(x) = x\theta(x)$$

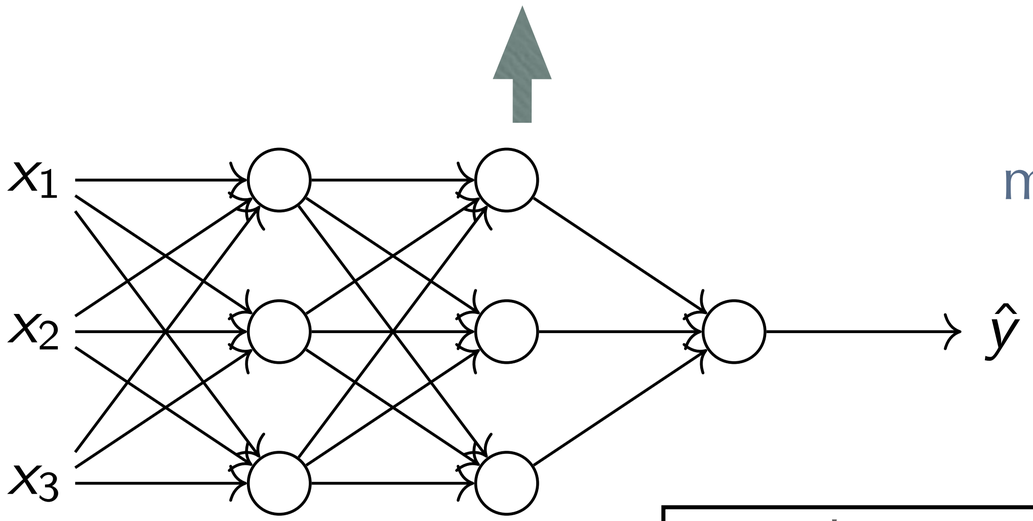


Training for classification

data pool of images



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



minimization of loss function

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

(28x28) の画像データをn 個

y: truth

1→(1, 0, 0, ..)

2→(0, 1, 0,⋯)

3→(0, 0, 1, ⋯)

output

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

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

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

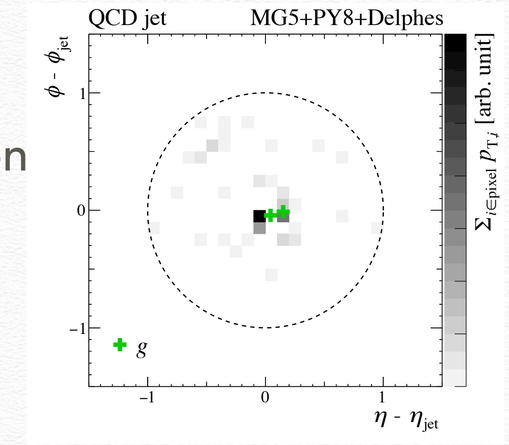
✅表現力　expressive power

✅データを学習　learn from data

✅微分可能Simple linear algebra + activation

Network design is important

QCD jet
(in W mass region



H jet

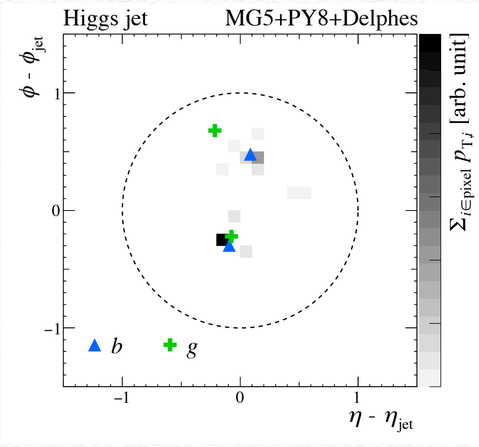
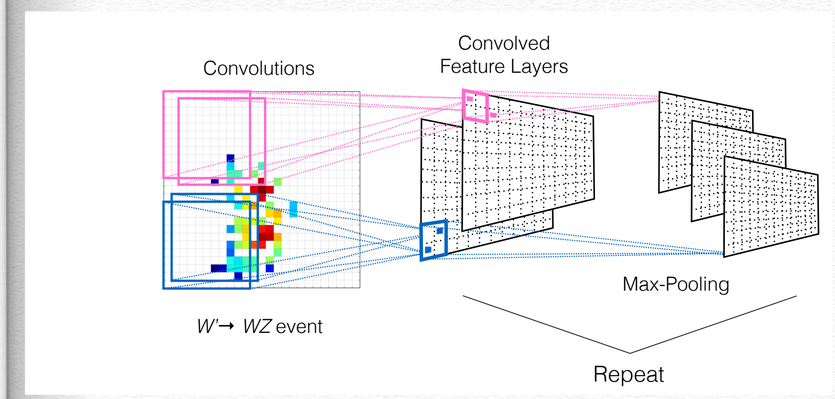


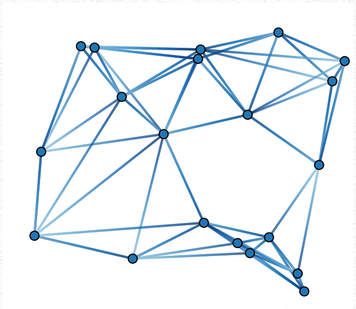
Image （Position and energy）←CNN

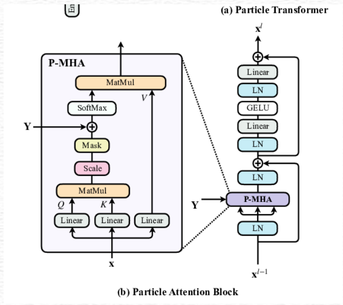
Graph Neural network
(particle →vertex)

transformer
correlation of all particles



sparse data





CNN Oliverira et al
(1511.05190)

1902.08570 Particle Net

Dreyer et al LundNet (1807.04758)

Gong et al LorentzNet(2201.08187)

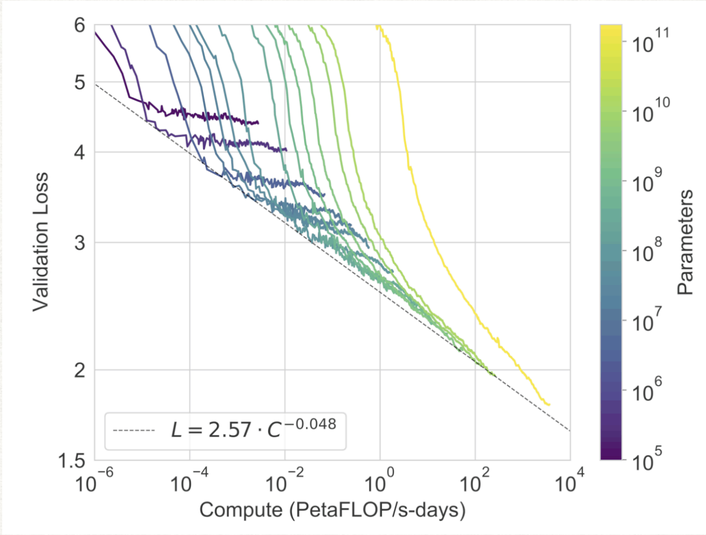
Bogatskiy et al PELICAN (2211.00454)

Particle transformer
2202.03772

TRANSFORMER

- Successful language model— good at long correlation(context)
- good at long term memory (keeping informaion)

Scaling low of LLM
(Large model and large data →improve forever



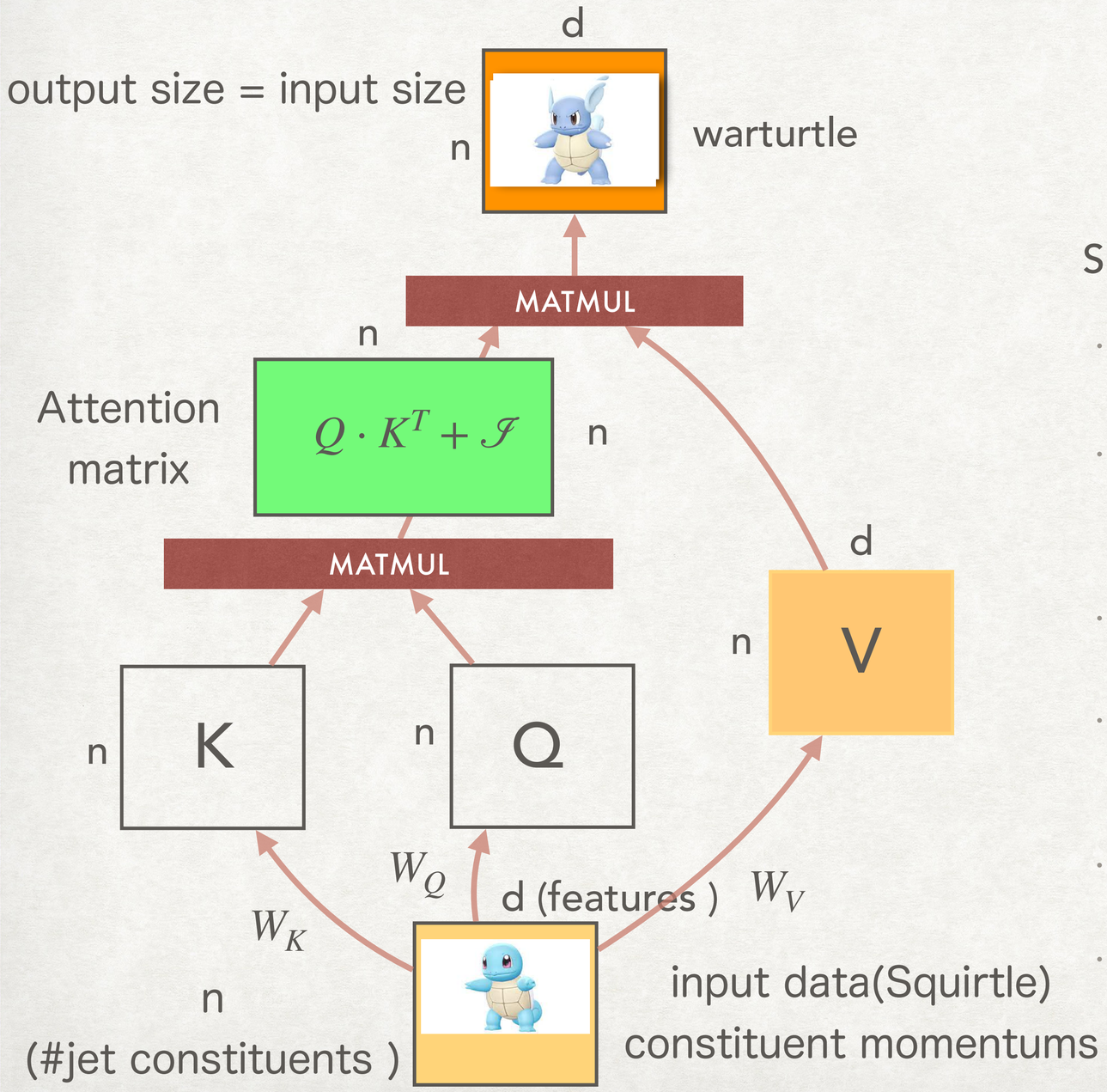
Recent Apple paper
(They just know patterns very well)

大規模言語モデルの限界について最近の Apple の論文が話題だと聞いたのですが、どの論文のことでしょうか。

Appleの研究者による注目の論文は、大規模言語モデル（LLMs）の「推論能力」における根源的な限界を示唆するものです。特に、**「Large Reasoning Models（LRMs）」が高度に複雑な問題に対して"完全な精度崩壊（complete accuracy collapse）"を起こす**ことを報告しています。具体的には以下の内容が取り上げられています。

https://chatgpt.com/share/6854a7be-2768-800c-a9a4-34dfd0431c4a

“PARTICLE TRANSFORMER” :SELF ATTENTION LAYERS



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

SETUP

- input X: n(particles in the jet) x d(features)
- n x n Attention matrix $A = QK^T$
  $K = W_K X, Q = W_Q X$ (W trainable parameters)
- Multiply $V = W_V X$ to get n x d output δX
- stack attention layers for X → X’=$X + \delta X$→X’’
  skip connection help to keep original information
- completely built from Low level feature
- using all correlation inclusing long range to short range.

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	

PHYSICS APPLICATION OF MACHINE LEARNING: CONCERNS

1. bias and valiance trade off on the training results

ML can approximate “any function” in infinite paramter limit, therefore unstable against the fluctuation of the training sample.

2. Interpretability

It gives you the results, but the reasoning is deep inside the network and hard to extract. This is problematic when you do not understand “true” distirbution. This is often the case for collider physics, where MC do not reproduce experimental data in detail.

3. scaling vs symmetry

Large data and Large network wins(it says) , but actually nature has symmetry, Permutation Invariance, boost invariance, Lorentz symmetry.. which is not easy to reconstruct from scratch.

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!

TRANSFORMER VS PHYSICS

With Ahmed Hammad

1. **fast, lightweight, while keeping performance** Reduction of network parameter **2M→100k**

   Incorporate physics picture, Cross attention between subjet vs hadron inspired by **QCD scaling behavior** **MIXER network** **arXiv 2404 14677 JHEP 06 (2024) 176**

2. **Jet analysis → event analysis.** analysis of H→hh arXiv 2401.00452 JHEP 03(2024)

3. **Respect symmetry** **Replacing "attention from generic features" →"pairwise boost invariant information " (IAFormer)**

   Reduce valiance in training → via **differential attention** **arXiv 2505.03258**

With Sung Hak Lim

4. **Identify the key parameters for classifications** → via comparison with the simple MLP model using High leavel features

5. **calibration of MC** in future. arXiv 2312.11760 JHEP 07 (2024) 146 arXiv 2503.01452 JHEP(accepted)

1. PARTON SHOWER AND NETWORK STRUCTURE

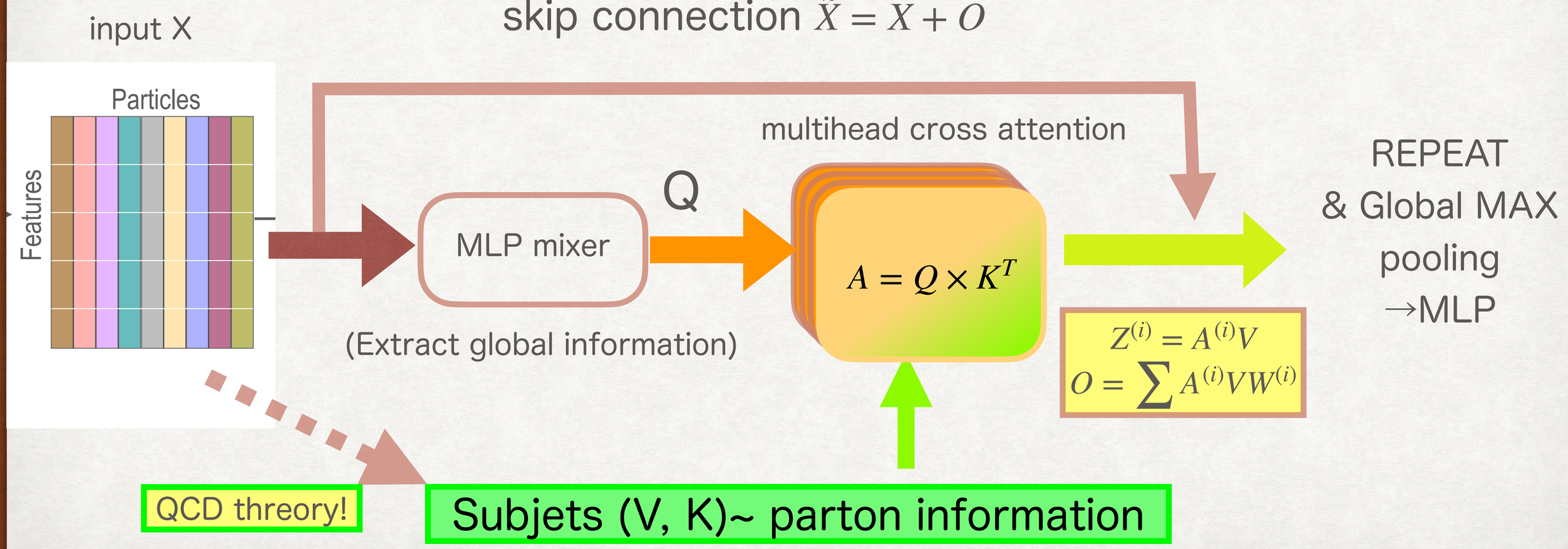
"Ahmed Hammad , MN "Streamlined jet tagging network assisted by jet prong structure" arXiv 2404 14677 JHEP 06 (2024) 176

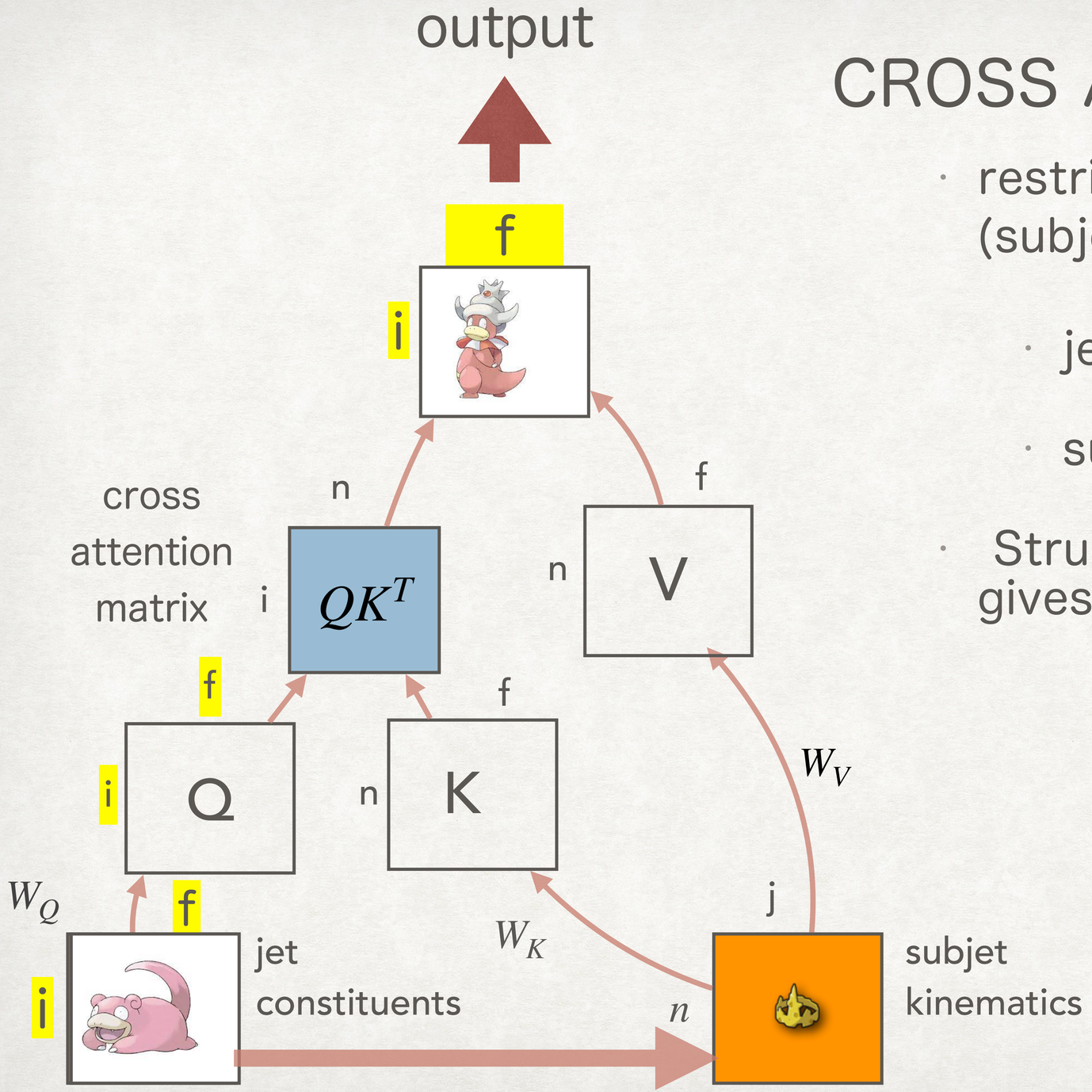
- Hard Process = Partons(quarks and gluons) {y}
- a jet: P(hadrons in jets | parton ~ jet) = $P(\{x_i\}\,|\,\{y\})$
- jet with substructure $P(\{x_i\}\,|\,\{y_\alpha\})$
- Maybe several fatjets in an event (factorization)

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

Why don't we construct the network forcusing on parton(subjet) vs hadrons

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





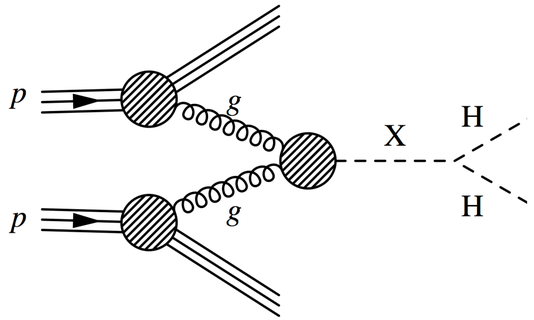
CROSS ATTENTION LAYERS

- restrict network to cross attention (subjet) x (jet constituent) $A = QK^T$
  - jet constituent Q
  - subjet ~ parton, shower K V
- Structure: High scale feature (subjet) gives extra weight to jets constituents

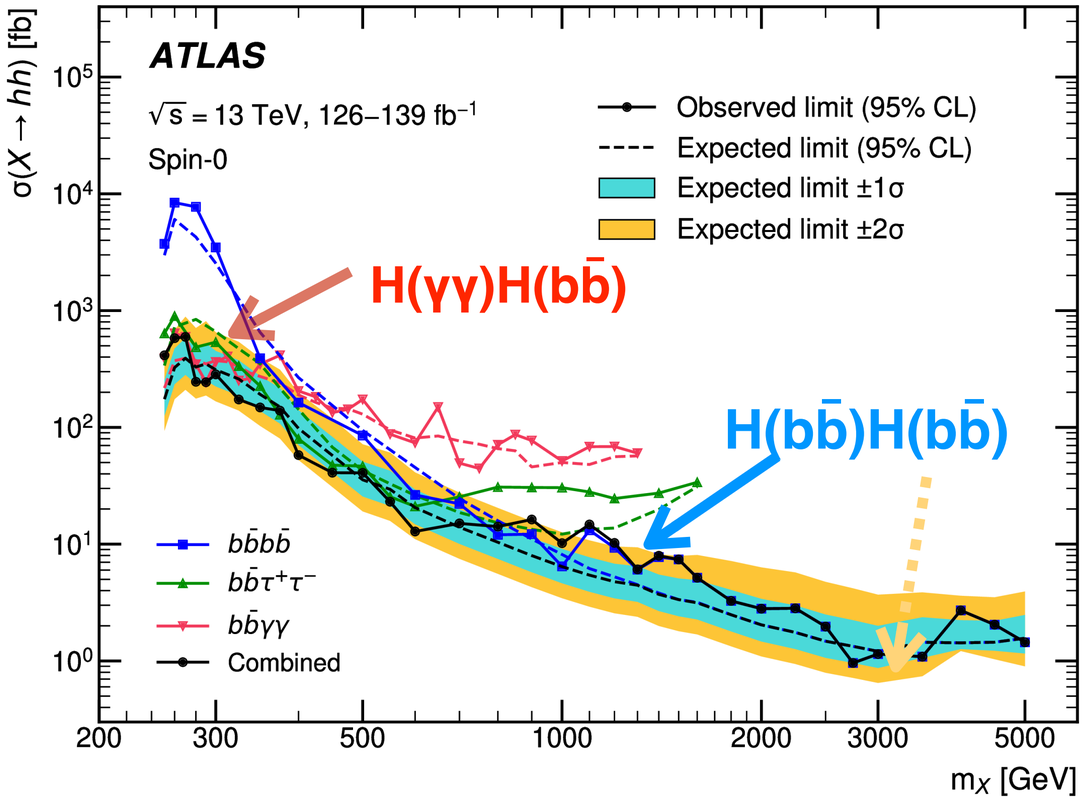
The performace is not reduced significantly from Transformer while networks size is very small becuase we respect QCD

2. CROSS ATTENTION AND GLOBAL EVENT KINEMATICS

X→HH



Phys. Rev. Lett. 132 (2024) 231801

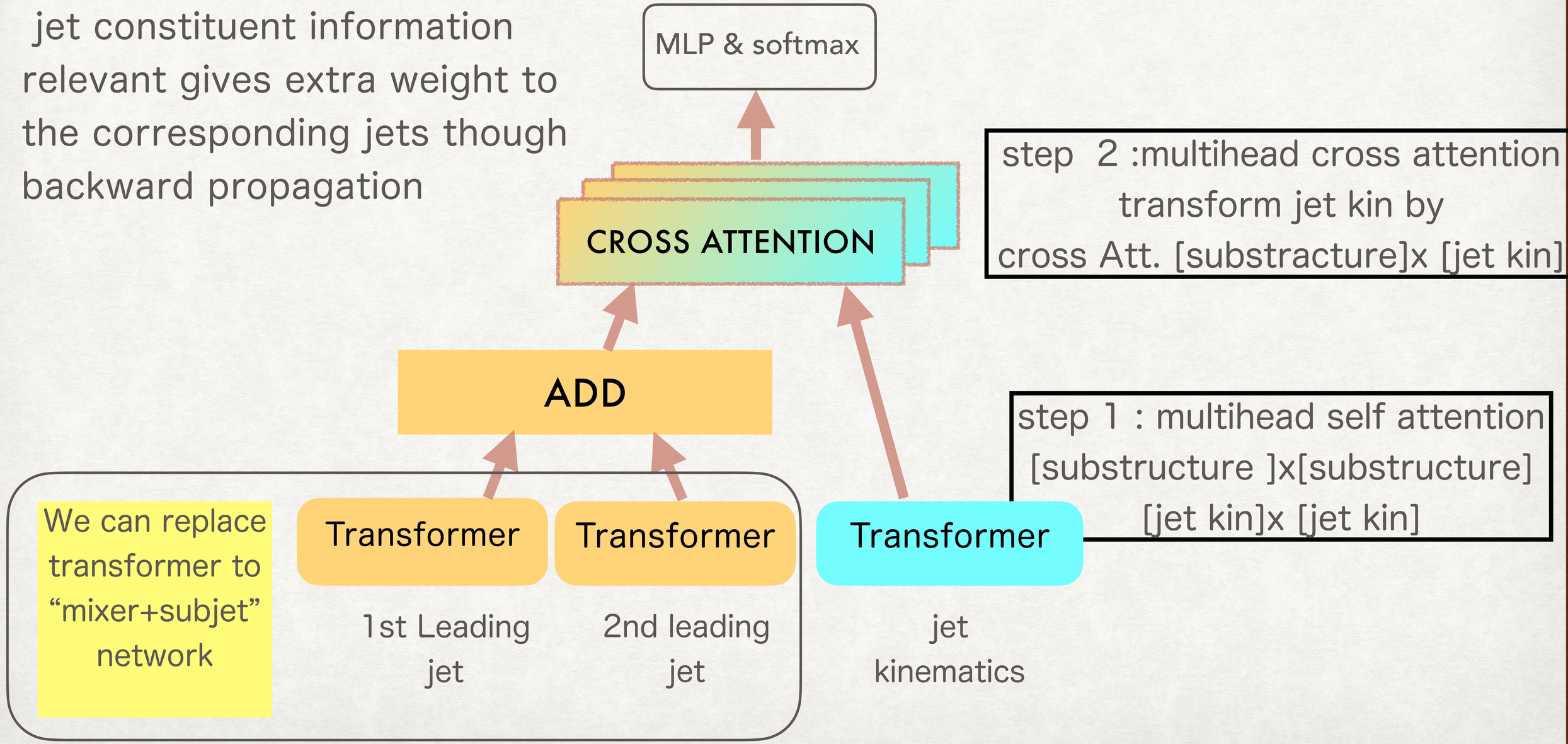


H(bb̄)H(bb̄) most sensitive channel for $m_X > 400/500$ GeV

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

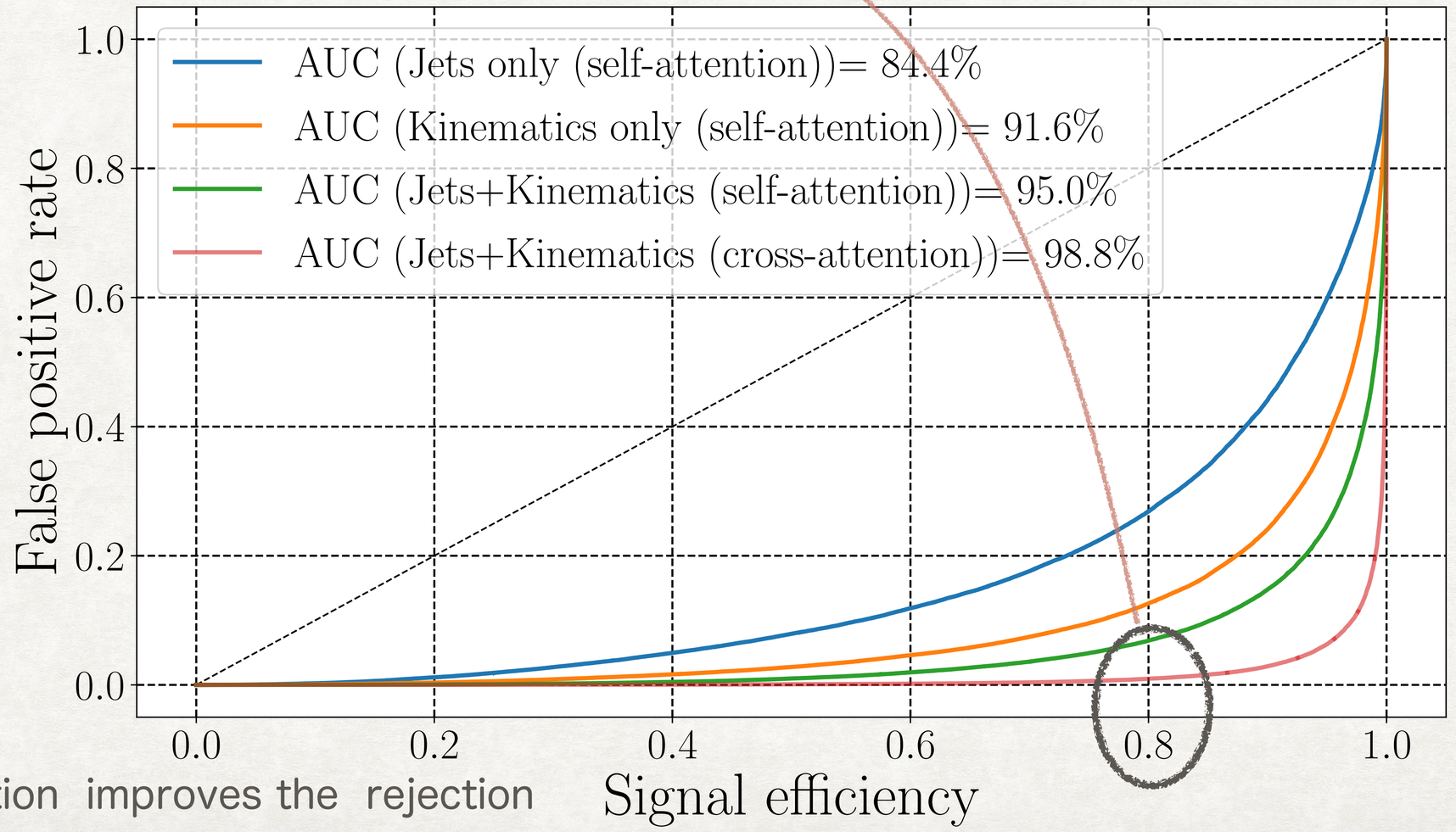
cross attention motention for 2 fatjet events

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



IMPROVEMENT USING CROSS ATTENTION COMBINIE TO NEXT

factor 5 improvement at the same acceptance.



Cross attention improves the rejection efficiency significantly

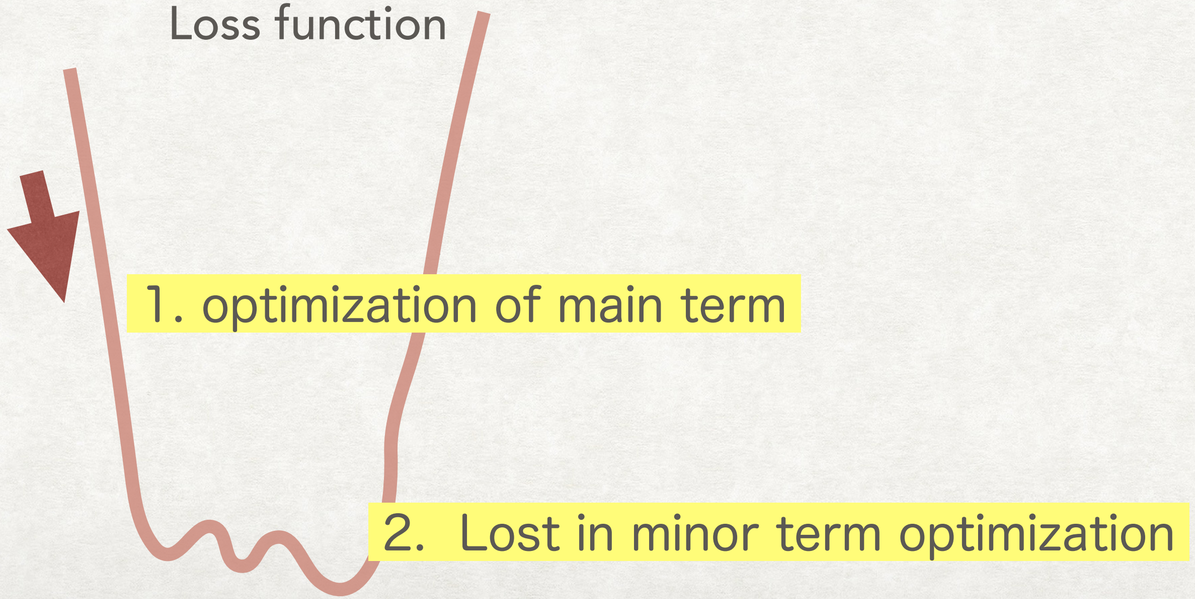
NETWORK BUDGET BEHIND CROSS ATTENTION

Isn’t that good enough to add subjet information to particle and do transformer ?

$$AV = \begin{pmatrix} Q(Sub) \times K(Sub) & Q(particle) \times K(Sub) \\ Q(Sub) \times K(particle) & Q(particle) \times K(Particle) \end{pmatrix} V = \boxed{Q(sub)\ K(sub)\ V(sub)} + \ldots$$

Dominant contribution

our network kill this term
and keep off-diagonal part only



3. (IAFormer =InterAction transFormer)

Transformer Input: $\eta, \phi, p_T \ldots$

BUT we only want to use lorentz covariant or boost invariant for LHC) information

$$\Delta R = \sqrt{(\eta - \eta')^2 + (\phi - \phi')^2} \quad z = \min(z_1, z_2)/(z_1 + z_2) \ldots$$

so using $K = W_K V, Q = W_Q V$ is too much…

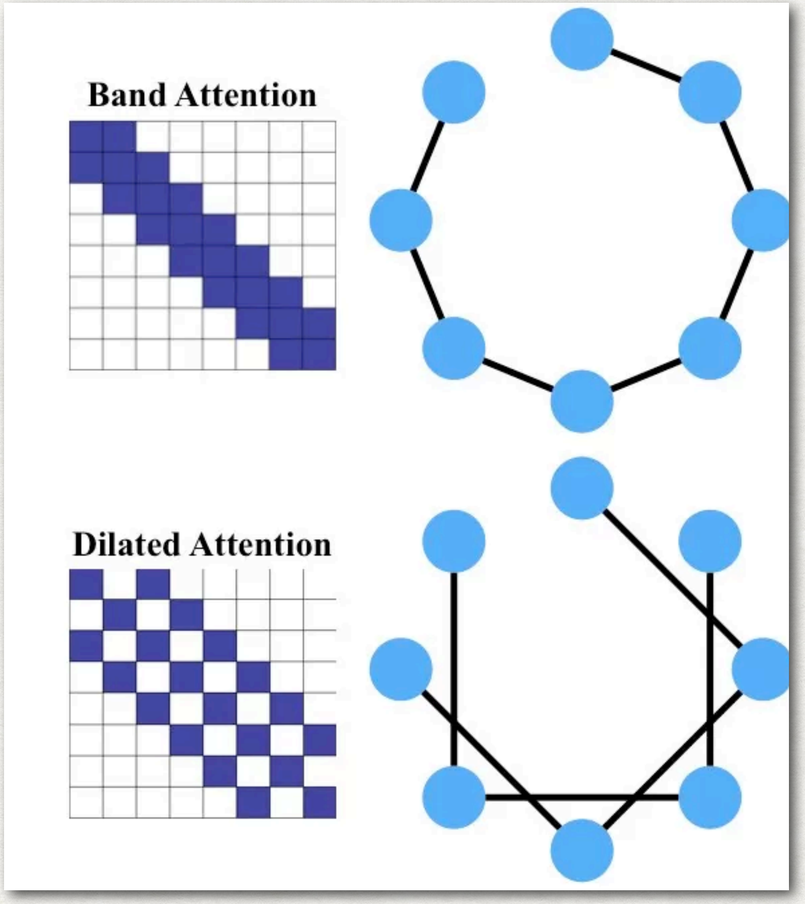
Lorentz covariant network is proposed and performance is high. but boost invariance is also known to be useful at LHC

$$\alpha = \text{softmax}\left([Q \times K^T]^{n \times n}\right) \rightarrow \alpha = \text{softmax}(\mathcal{I}^{n \times n})$$

$\mathcal{I}$: function of pairwise variable

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/

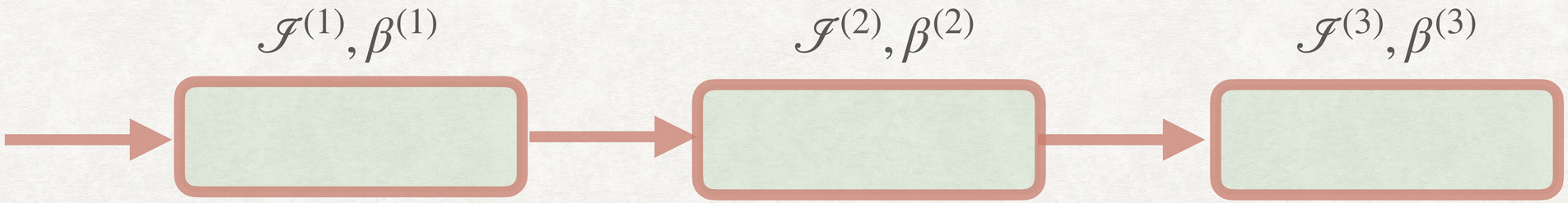
DIFFERENTIAL ATTENTION AN EXAMPLE OF DYNAMIC ATTENTION

- Dynamic attention "ask network to learn viable sparse attention pattern"
- We introduce a new dynamic attention called "differential attention" to the network. arXiv:2410.05258

strength

cancel the irrelevant information

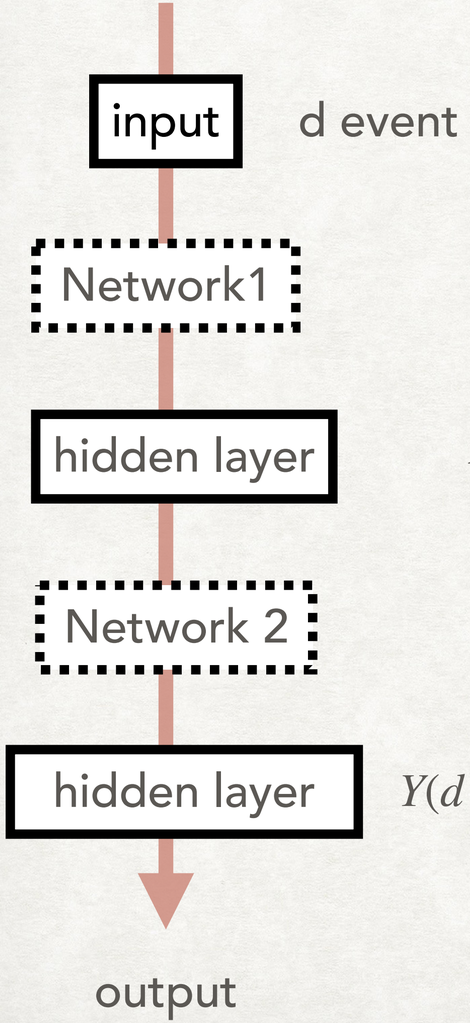
- $\alpha = \text{softmax}(\mathcal{I}) \rightarrow \alpha^{(i)} = \text{softmax}(\mathcal{I}_1^{(i)}) - \beta^{(i)}\text{softmax}(\mathcal{I}_2^{(i)})$



Each layer built different filters dynamically

EFFICIENCY AND INTERPRETATION

Compare hidden layers using CKA(centered kernel alignment) similaity



$$X(d \times h_1)$$

$$\rightarrow M = XX^{\dagger} \quad (d \times d)$$

$$Y(d \times h_2) \rightarrow N = YY^{\dagger} \quad (d \times d)$$

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

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

CKA <<1 if new information is captured

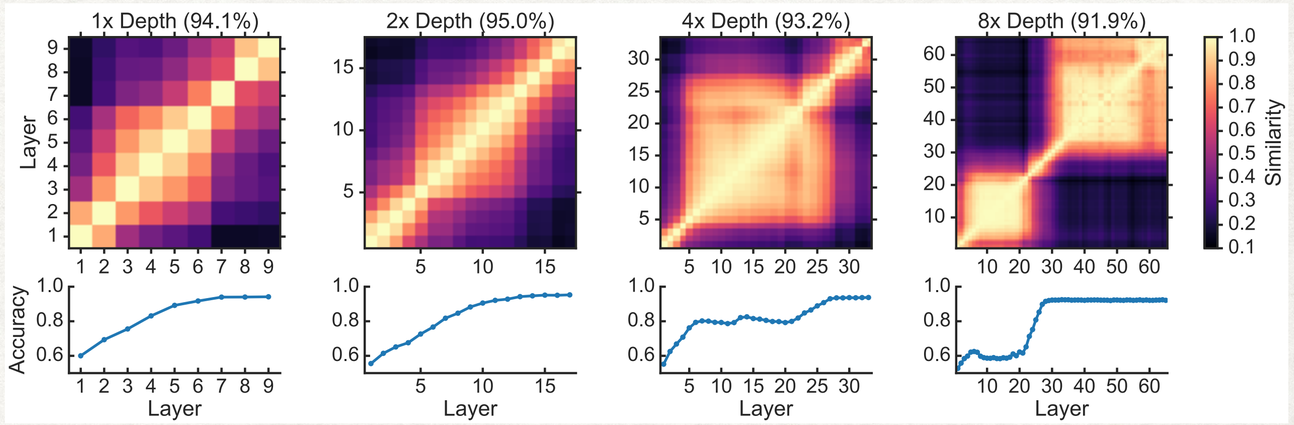


Figure 3. CKA reveals when depth becomes pathological. **Top**: Linear CKA between layers of individual networks of different depths on the CIFAR-10 test set. Titles show accuracy of each network. Later layers of the 8x depth network are similar to the last layer. **Bottom**: Accuracy of a logistic regression classifier trained on layers of the same networks is consistent with CKA.

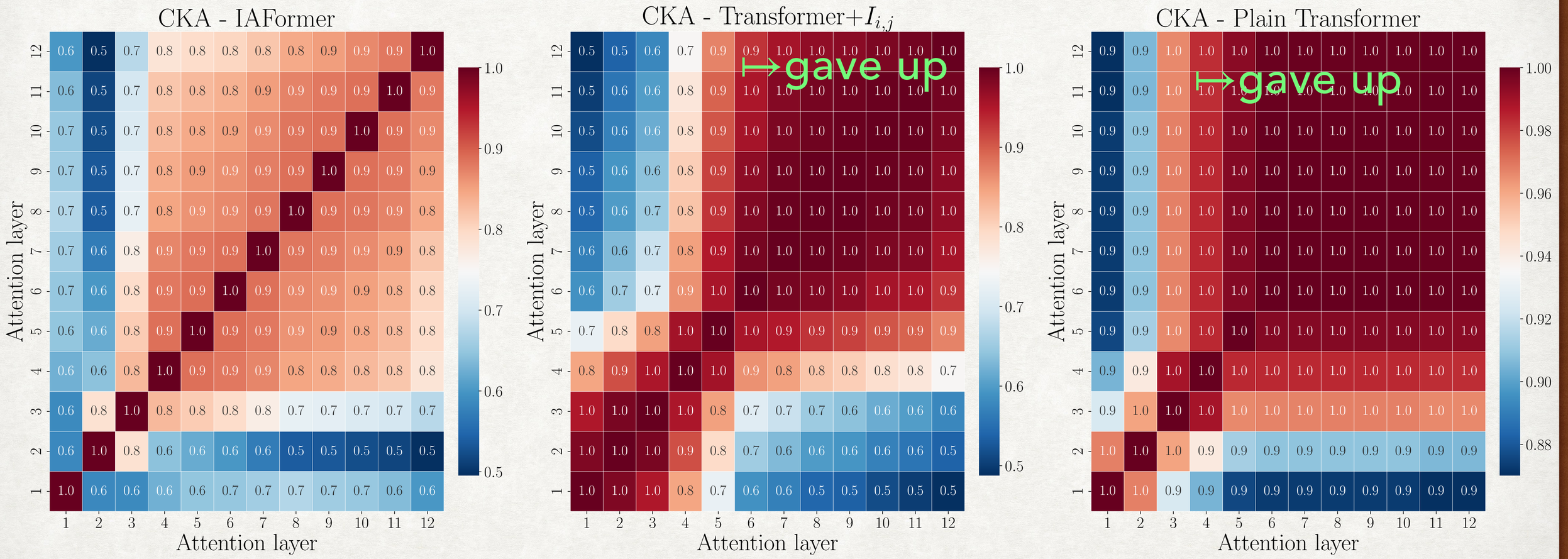
IAFormar constantly improve classification performance

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

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

Plain Transformer

$\epsilon_b(50\%) = 360$

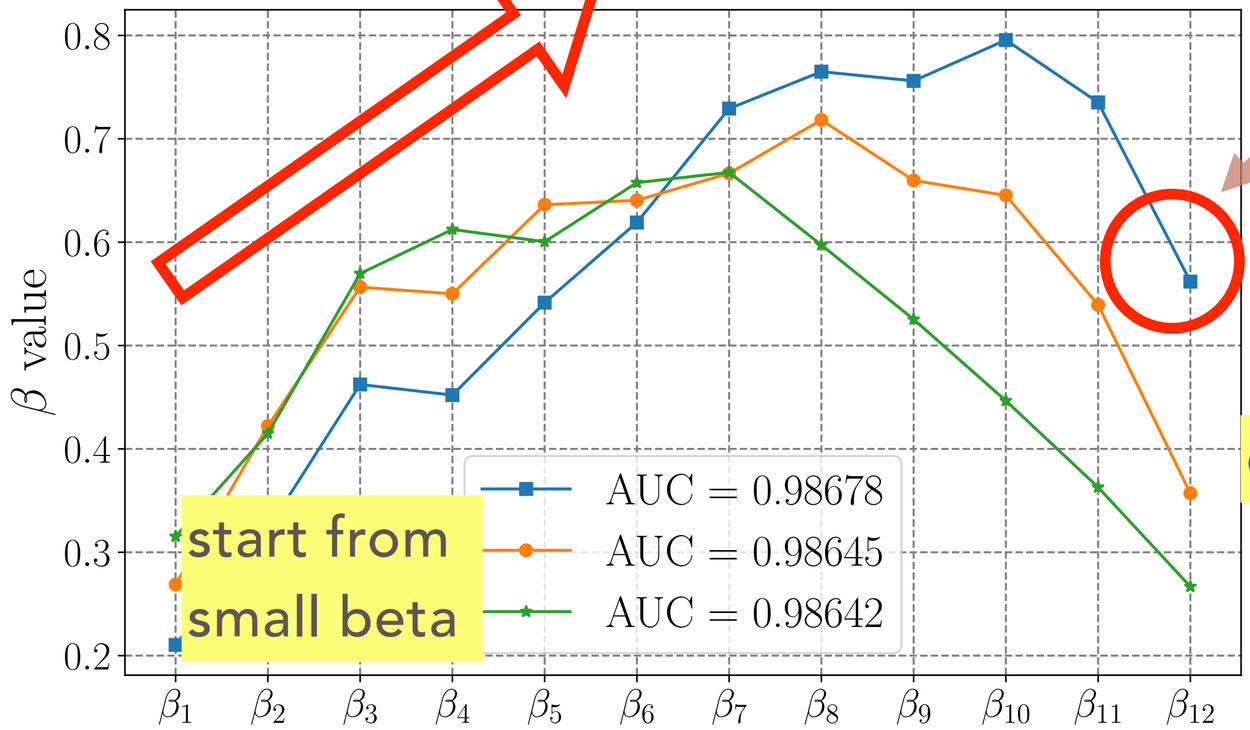


BEHAVIOR OF DYNAMIC FILTERS

filtered information

Increase linearly

max



start from small beta

collarpes!

most successful network trainig

last β is large → higher network performance

4. Important feature for jet classification

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

pt distribution of constituents

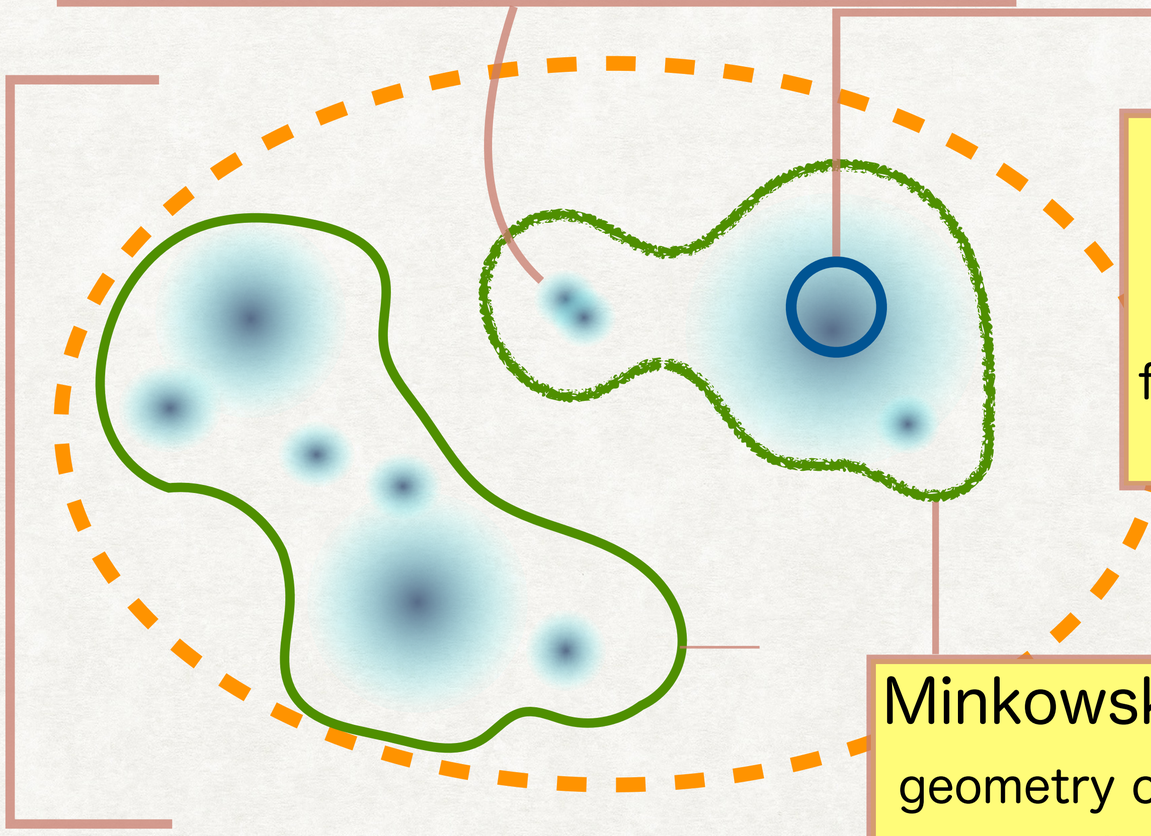
Jet spectrum

two point Energy correlation (unlocalized sampling)

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

Energy flow polinomials

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



Subjet

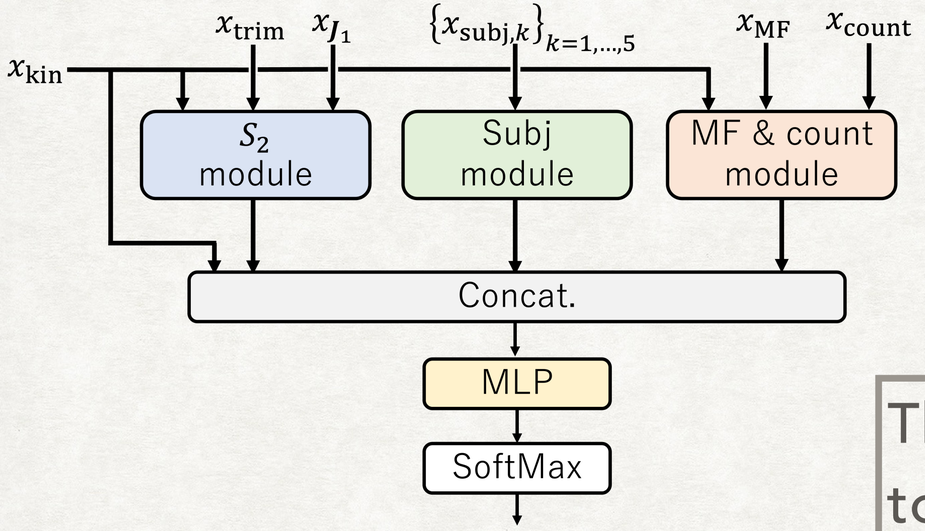
Localized sampling momentum and counting for various angular sccale R=0.1, 0.2, 0.3

Minkowski Functionals

geometry of jet cosntituent distribution

NETWORK USING HL INPUTS VS PARTICLE TRANSFORMER

Aanalysis model(AM)
modulated structure taking HLF
easy to add and remove



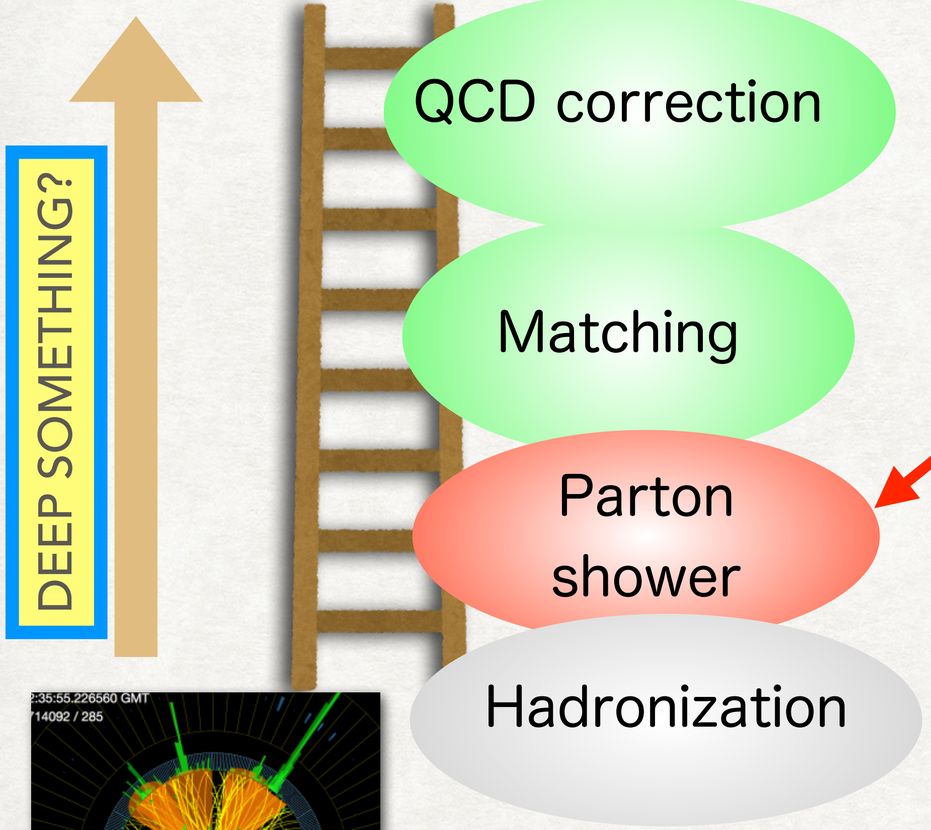
arXiv 2312.11760 JHEP 07 (2024) 146 arXiv 2503.01452 JHEP(accepted)

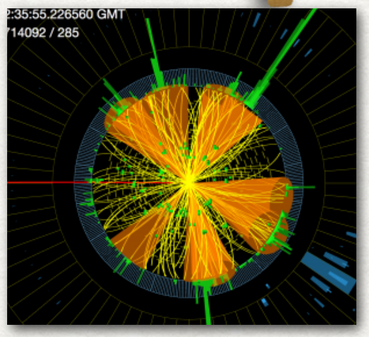
	Pythia		Herwig		PY vs HW QCD		PY vs HW TOP	
Full AM	85.7	343	61.3	**244**	**2.78**	**5.43**	2.82	5.36
ParT [33]	**90.5**	**372**	**62.6**	242	2.77	5.38	**3.07**	**5.94**

for Systematical error correction

The network using HL feature is more stable and useful to fine difference among Monte Carlo and reduce systematical errors (currenty more than 20%) because variance is low .

5. Calibration of MC data





QCD connecting BSM and events

Madgraph: Automatic Amplitude calculation in NLO level

2001 Matrix element and Parton shower matching
MLM CKKW → 2007 Madgraph

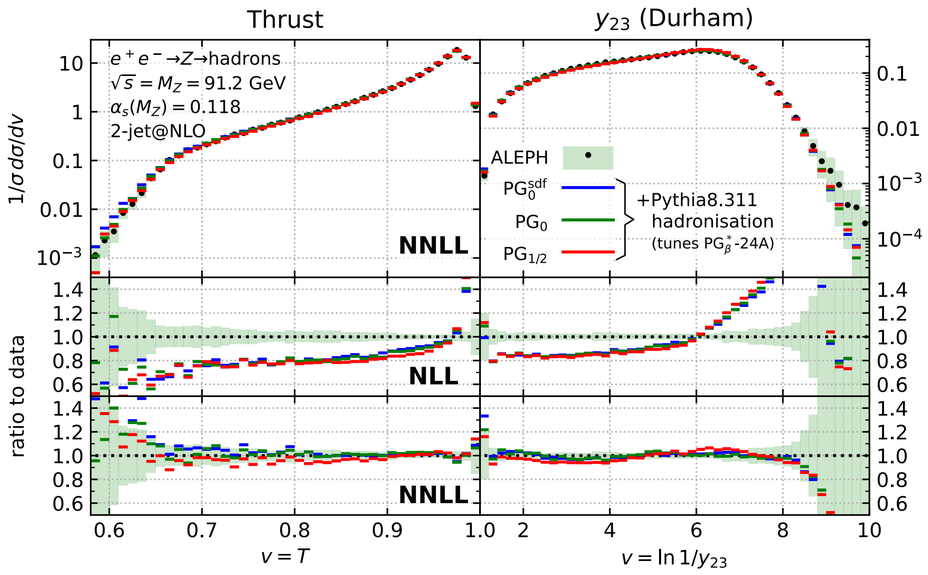
Pythia, Herwig, Sherpa evolving toward Dipole shower,
Higher order collection of PS (Panscale)

2006 QCD aware definition of jets(fastjet)

Should we give up Theory? No.

It is time to invest the parton shower and hadronization algorithm in the level relevant to DeepLearning Era

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 relevant away from the 2-jet region.
- 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 $\alpha_s^{\rm eff}$ not an $\alpha_s^{\overline{MS}}$

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