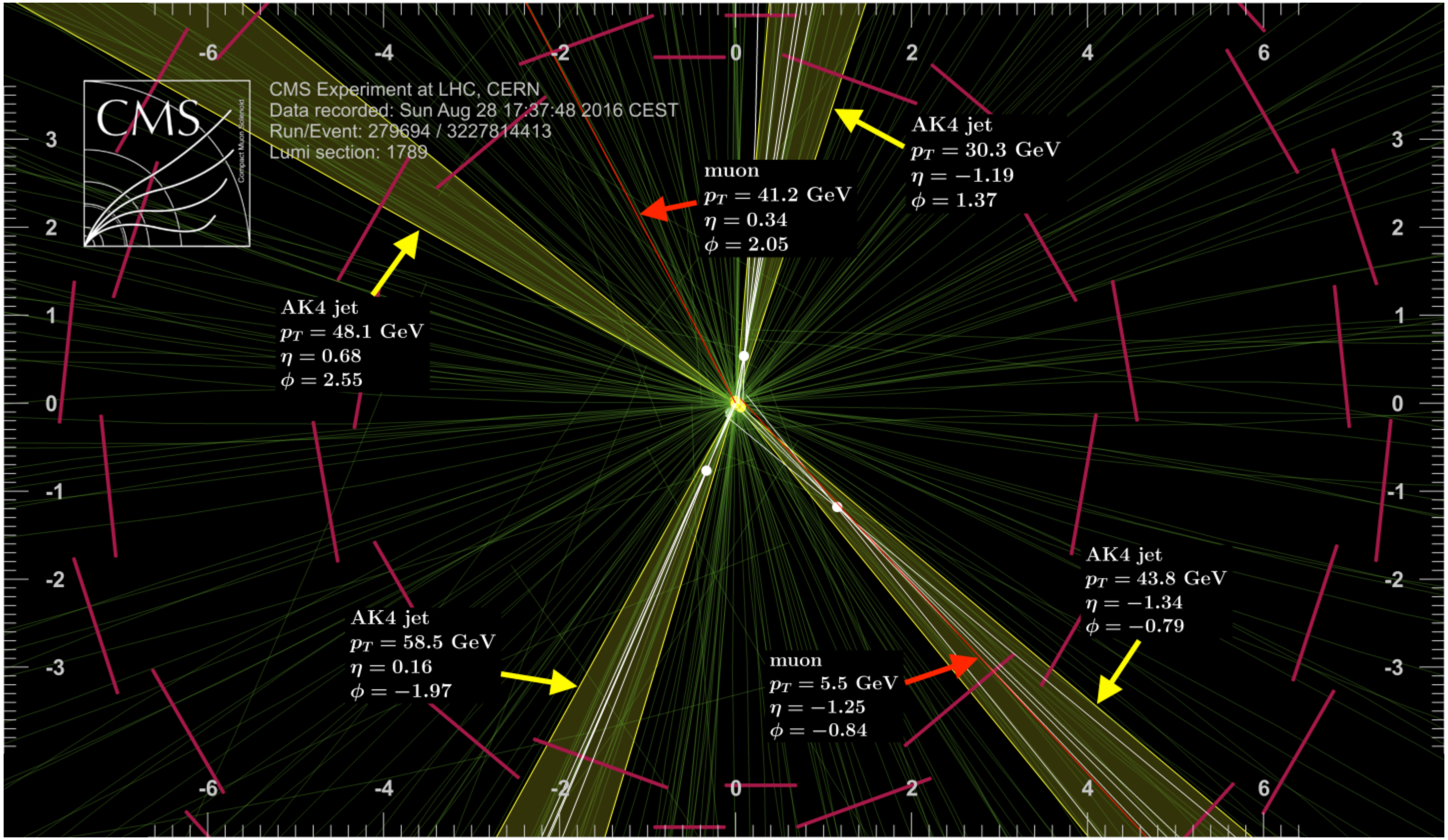
Transformer models for Heavy Flavor Jet Identification in CMS

Alexandre De Moor, Congqiao Li, Denise Müller, Huilin Qu, **<u>Sitian Qian</u>**

on behalf of CMS Collaboration

ML4Jets 2022 @ Rutgers University, 2022/11/01

INTRODUCTION: JET TAGGING



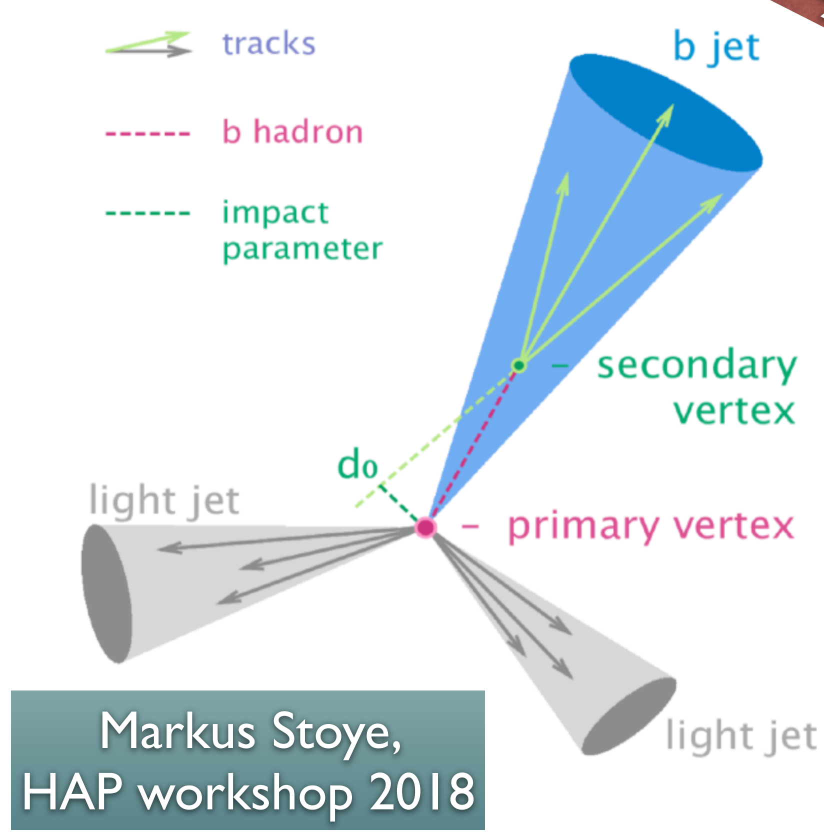
Jet tagging:

Identifying the particle that initiates the jet with experimentally observed quantities

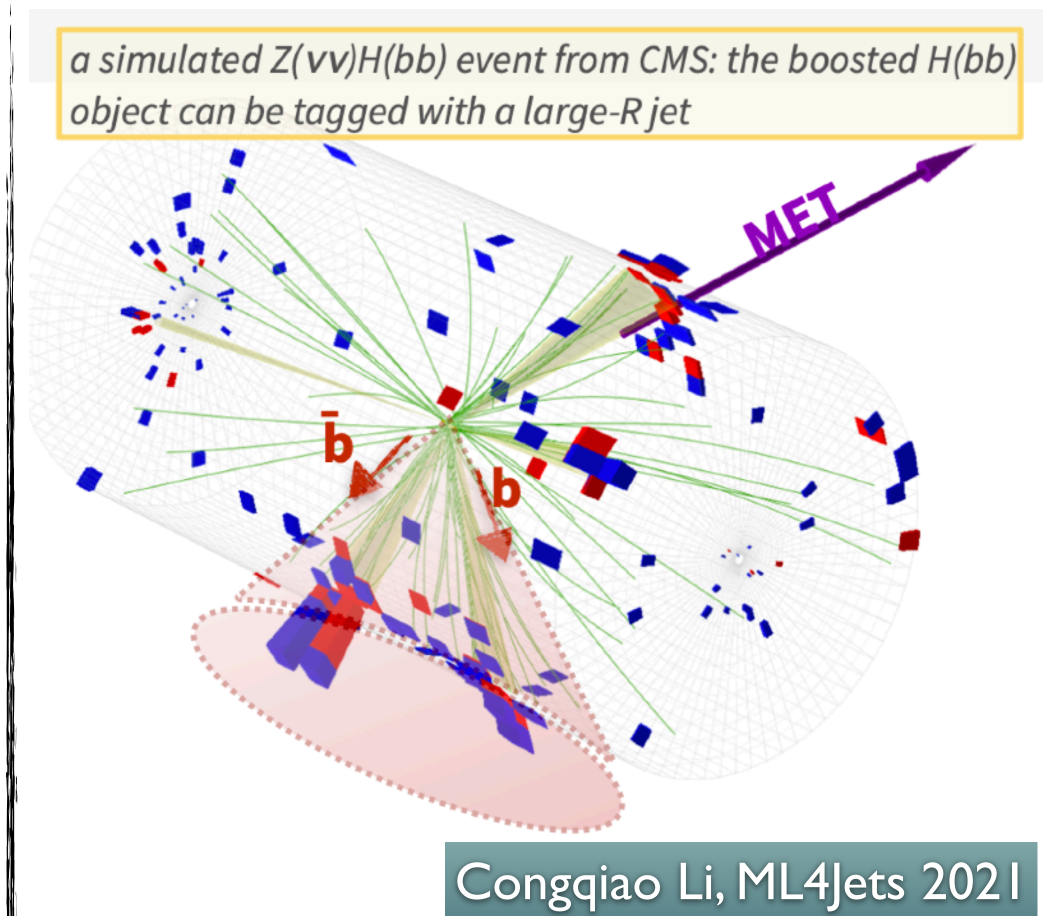
Introduction: Jet Tagging

Different types of jet tagging

Today's focus



Small radius tagging: Mainly on the flavor information (light quark/gluon/charm/beauty/tau lepton…)



Large radius tagging: Targeting on boosted heavy resonance (W/Z/H/Top…)

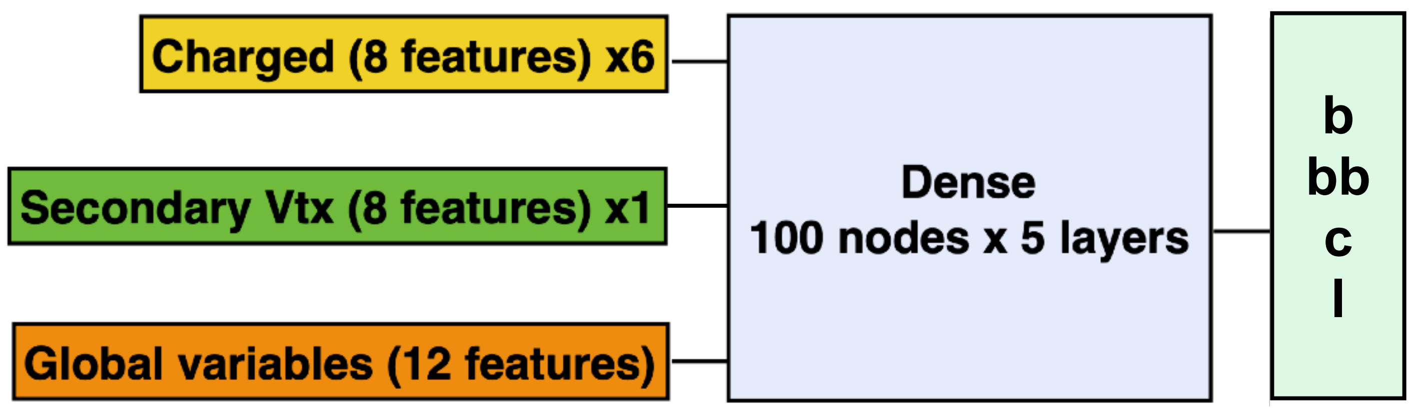
HEAVY FLAVOR TAGGING

- Problem setup of heavy flavor (HF) tagging:
  - Focus on small radius jet (in CMS: anti-kT with R=0.4)
  - Usually 3 cases:
    - Tagged as “b”
    - Tagged as “c”
    - Tagged as light flavor (“udsg”)
    - Can involve more categories including hadronic tau, “uds” vs “gluon” etc.
  - Key features of HF tagging:
    - Track displacement is of vital importance
      - Due to longer life of heavy flavor hadrons!

Heavy Flavor Tagging @ CMS

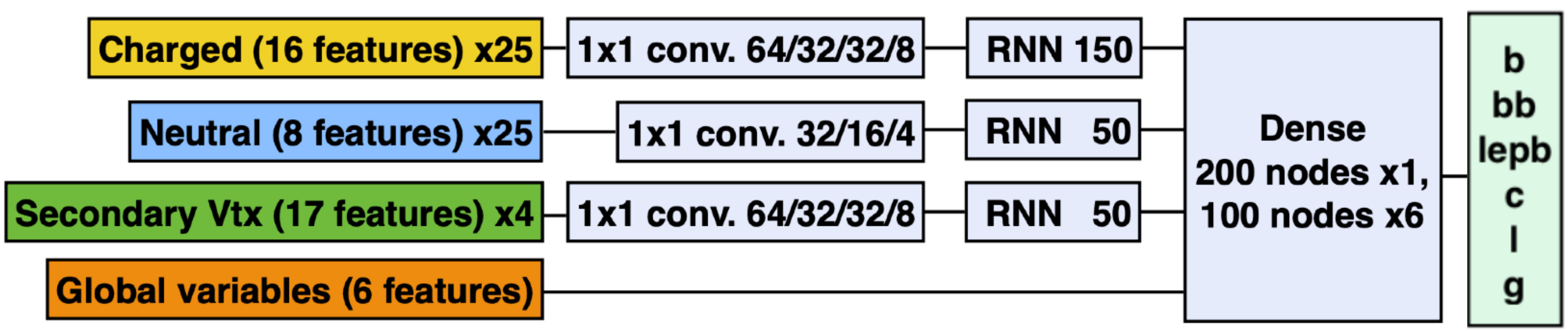
ArXiv: 1712.07158
CSVv2, cMVAv2 & deepCSV

- Brief review of CMS HF tagging evolution:
  - ~ 2016: **CSVv2(NN)** and cMVAv2(BDT)
    - CSVv2 is based on a very simple 1 layer NN
  - 2017: deepCSV



  - Mid 2017: deepJet (deepFlavour)

ArXiv: 2008.10519



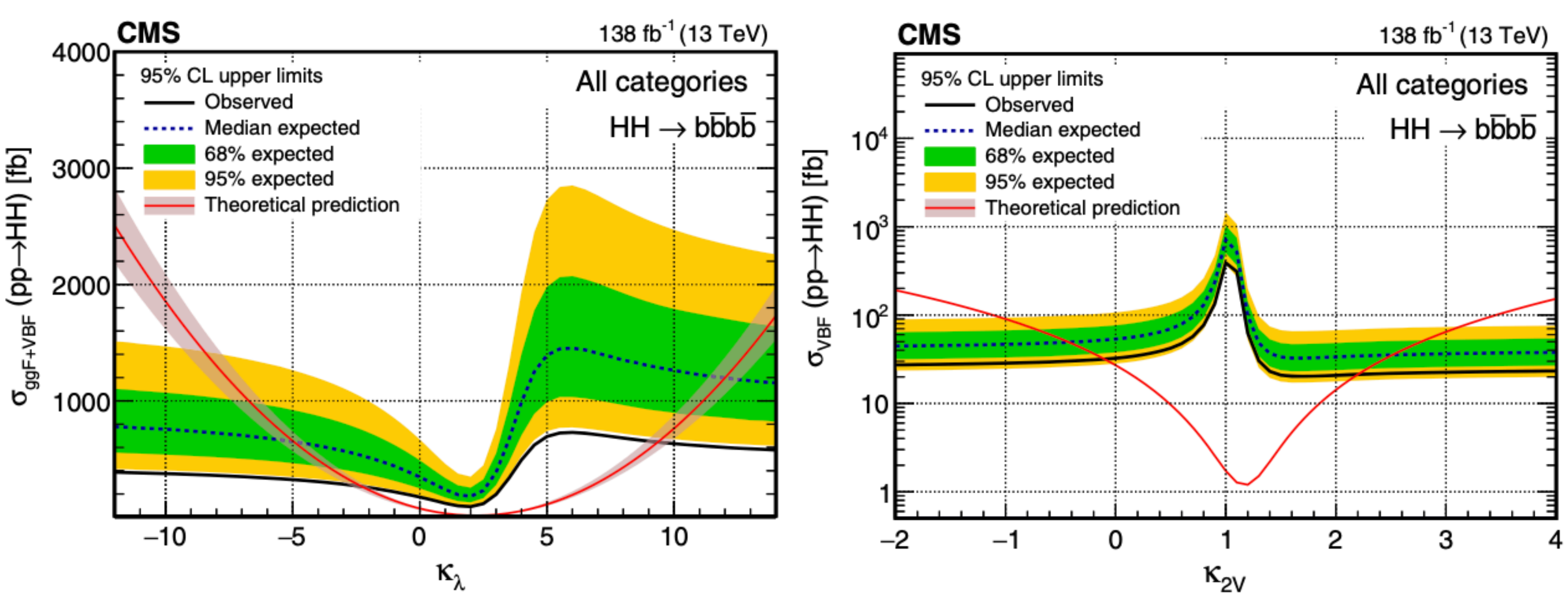
Heavy Flavor Tagging @ CMS

- Brief review of CMS HF tagging evolution:
  - ~ 2016: **CSVv2(NN)** and cMVAv2(BDT)
  - 2017: deepCSV
  - Mid 2017: deepJet (deepFlavour)

ArXiv: 1712.07158
CSVv2, cMVAv2 & deepCSV

ArXiv: 2008.10519

- Rich physical results have continuously come out with the help of CMS HF tagging algorithms!



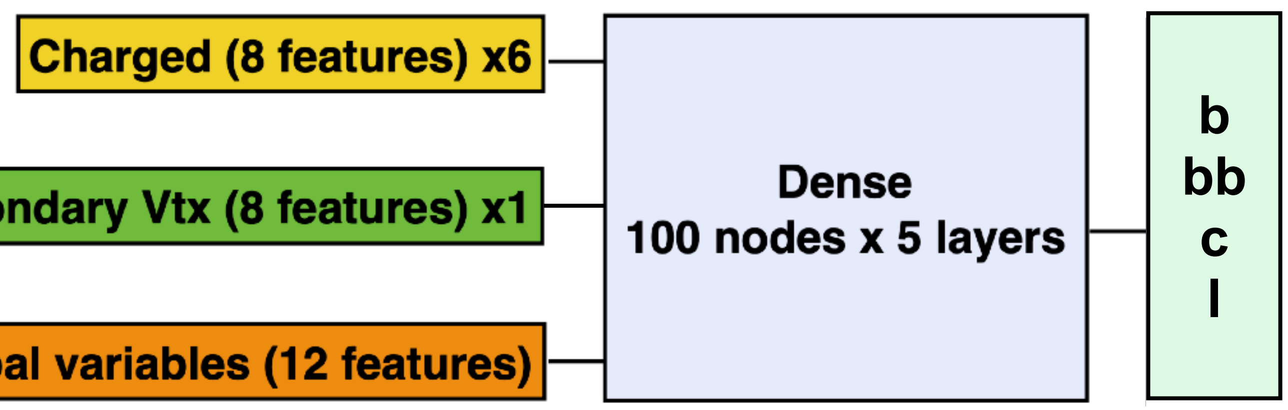
PhysRevLett.129.081802

Exciting result from CMS Full Run2 nonresonant HH4b measurement, b tagging is based on DeepJet algorithm

HEAVY FLAVOR TAGGING @ CMS

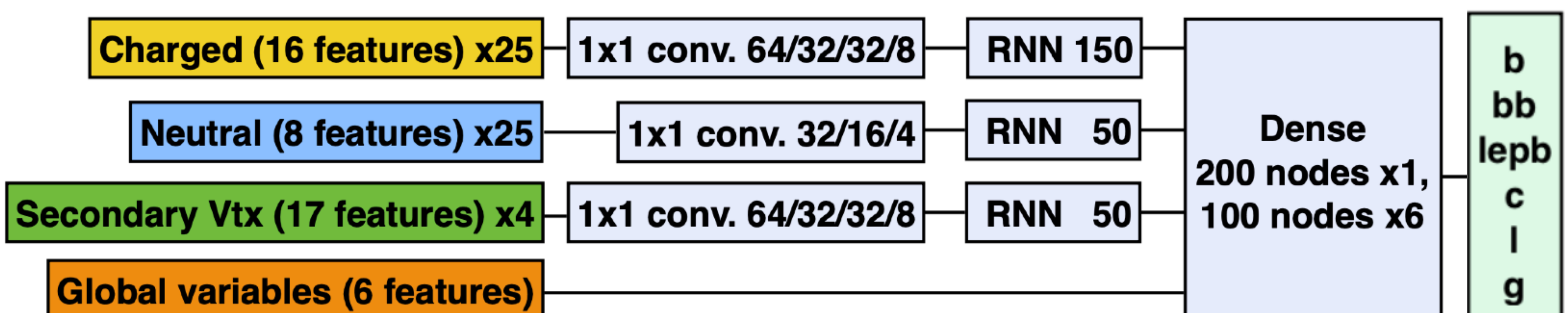
ArXiv: 1712.07158
CSVv2, cMVAv2 & deepCSV

- Brief review of CMS HF tagging evolution:
  - ~ 2016: **CSVv2(NN)** and cMVAv2(BDT)
    - CSVv2 is based on a very simple 1 layer NN
  - 2017: deepCSV



  - Mid 2017: deepJet (deepFlavour)

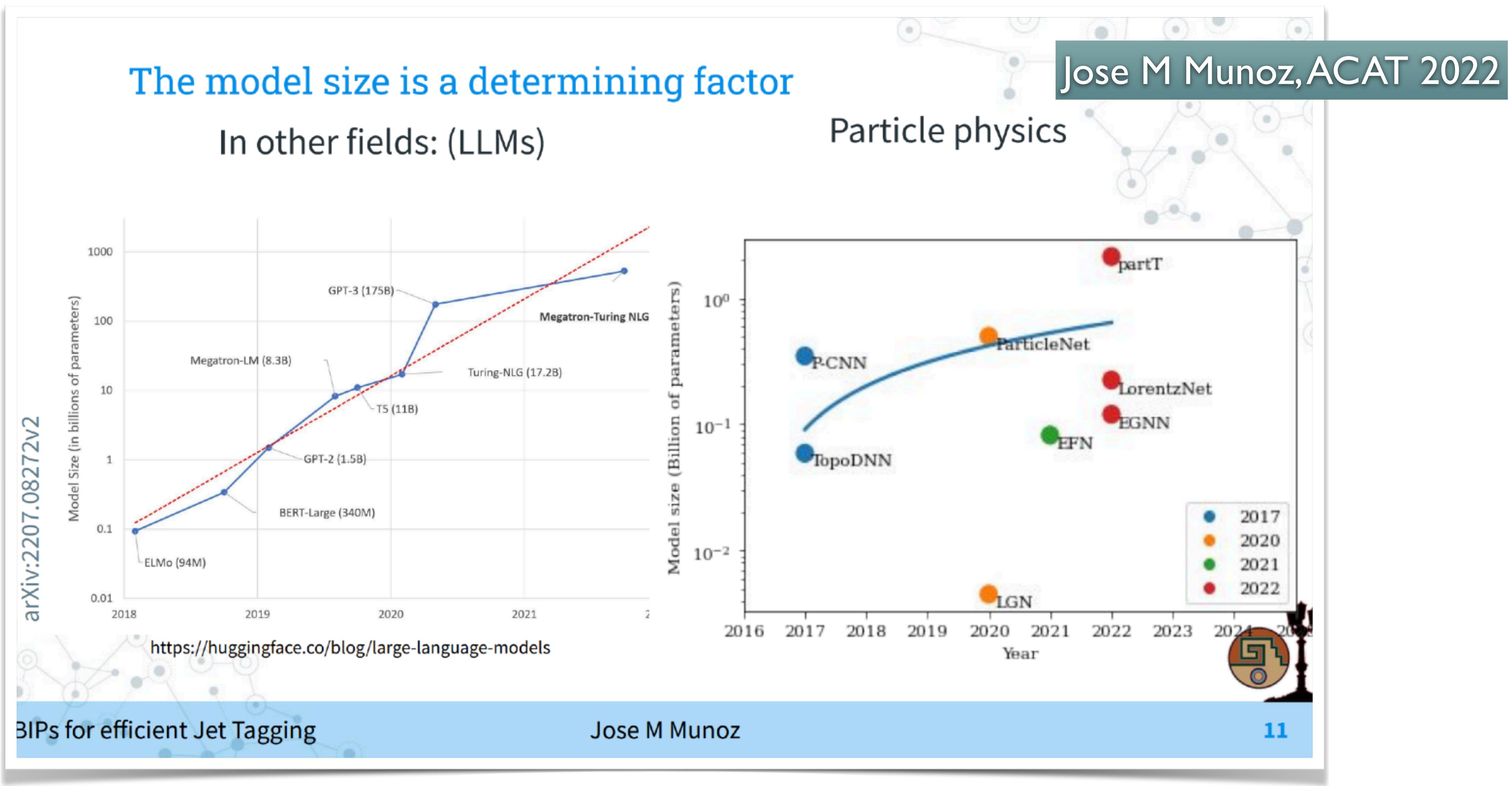
ArXiv: 2008.10519



  - **What about Run3?**

EVOLUTION OF JET TAGGING ALGO.

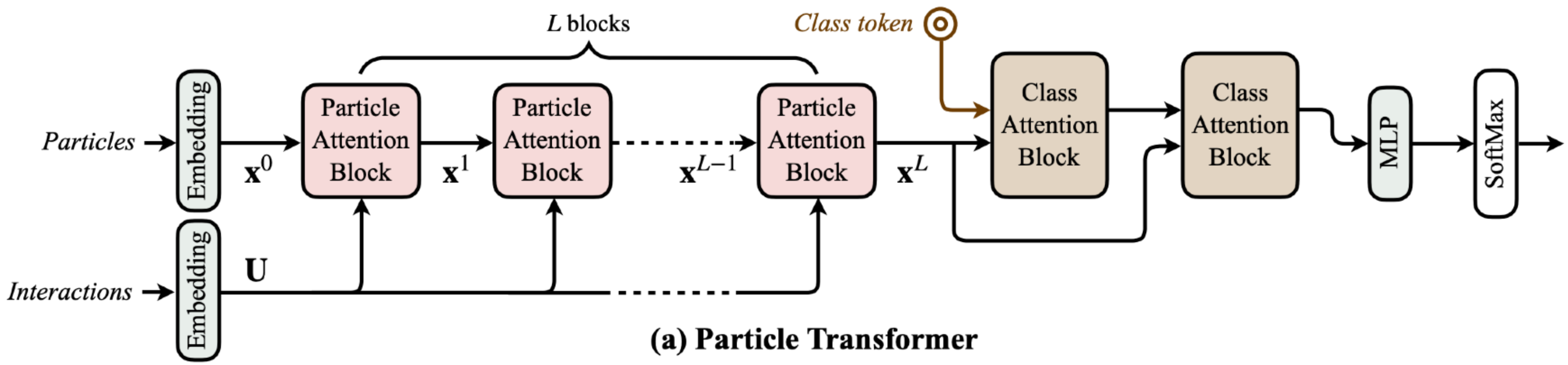
- Observation from CMS HF tagging evolution:
  - Models become larger and larger
- CMS HF tagging is not alone!



  - Larger model is now a trend in not only ML community, but also in HEP! -> Motivates the usage of largest model :) -> Particle Transformer (ParT)

PART 101

- Particle Transformer: ArXiv: 2202.03772
  - the transformer designed for particle physics

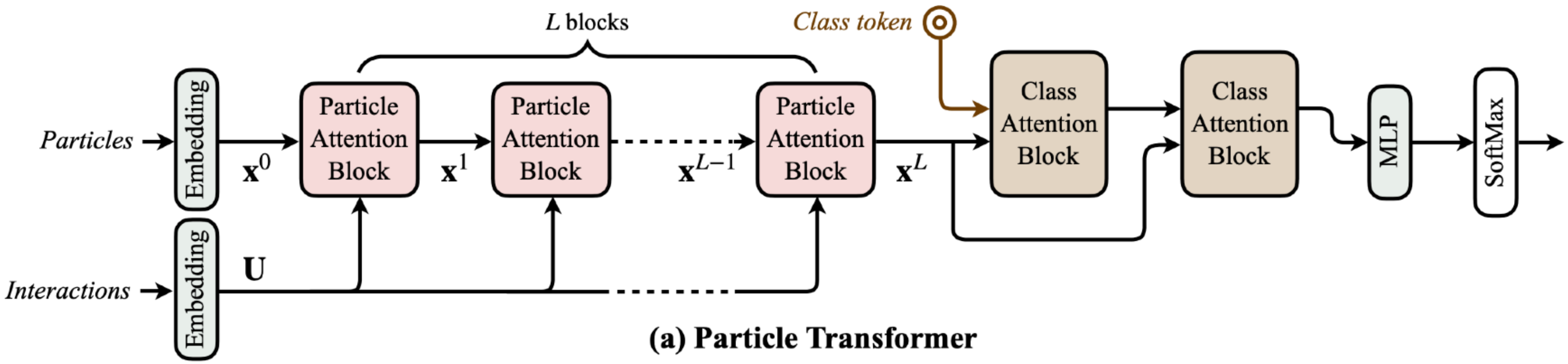


(a) Particle Transformer

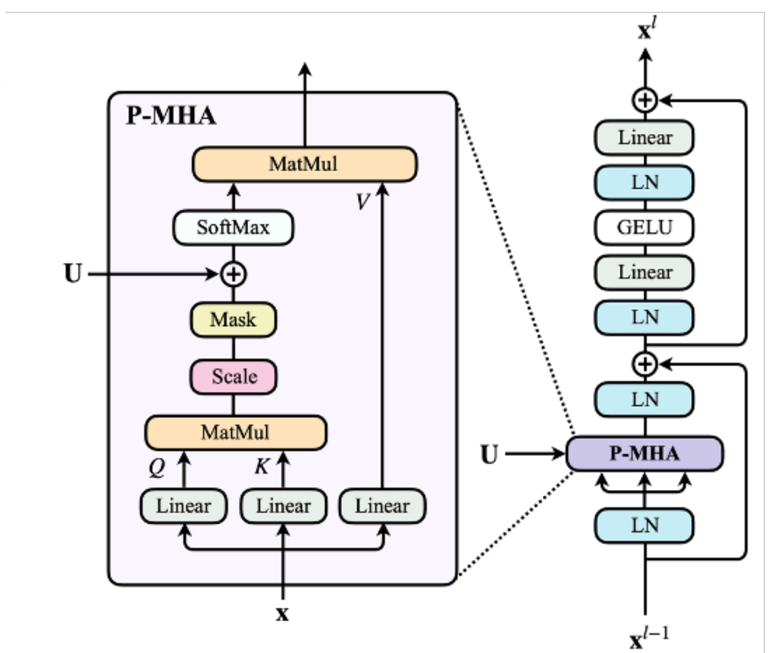
- Input embedding: Not only inject single particle information, but also include pair-wise features
- Particle Attention Block: Multi-Head Attention (MHA)
  - Pair-wise feature are introduced as the attention mask (P-MHA)
- Class Attention Block: Multi-Head Attention
  - Class token is used for the MHA calculation

PART 101

- Particle Transformer: **ArXiv: 2202.03772**
  - the transformer designed for particle physics



(a) Particle Transformer



(b) Particle Attention Block

$$\text{P-MHA}(Q,K,V) = \text{SoftMax}(QK^T/\sqrt{d_k} + \mathbf{U})V$$

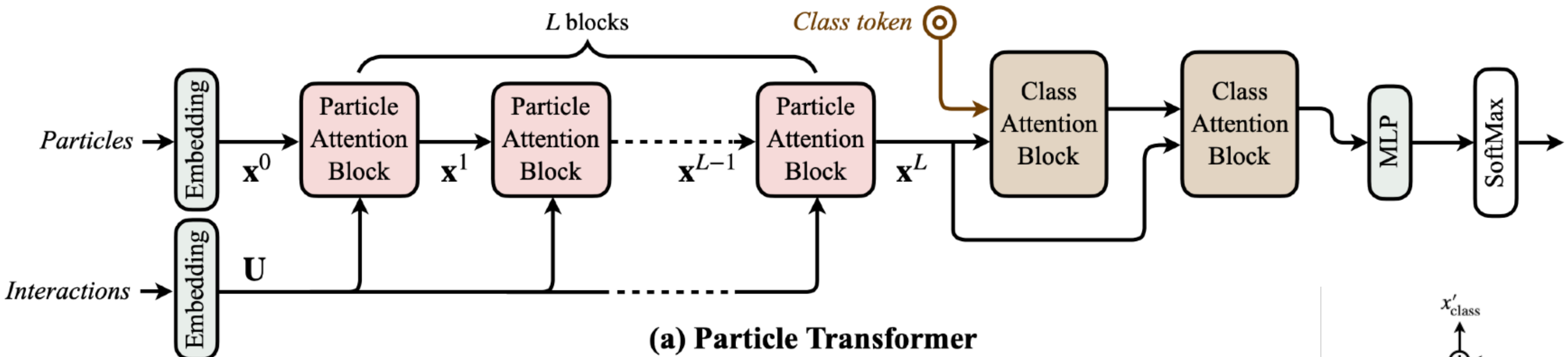
d_k: dimension of K

Choice of the pair-wise features: from LundNet

$$\begin{aligned}
\Delta &= \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2} \\
k_T &= \min(p_{T,a}, p_{T,b}) \cdot \Delta \\
z &= \min(p_{T,a}, p_{T,b})/(p_{T,a} + p_{T,b}) \\
m^2 &= (E_a + E_b)^2 - \|\mathbf{p}_a + \mathbf{p}_b\|^2
\end{aligned}$$

- Particle Transformer:
  - the transformer designed for particle physics

ArXiv: 2202.03772



(a) Particle Transformer

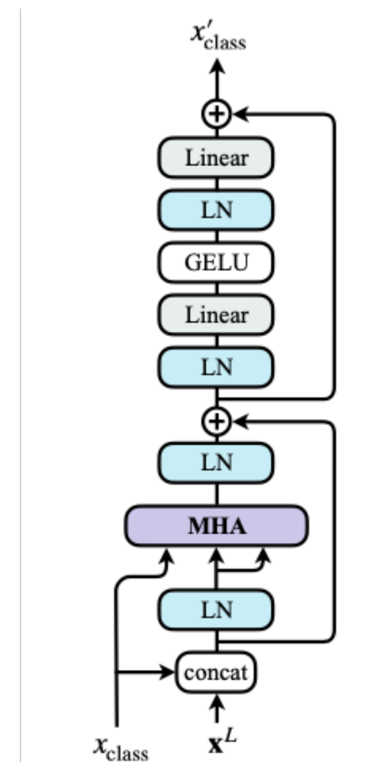
$$\mathrm{MHA_C}(Q_\mathrm{C}, K_\mathrm{C}, V_\mathrm{C}) = \mathrm{SoftMax}(Q_\mathrm{C} K_\mathrm{C}^T / \sqrt{d_{k\mathrm{C}}}) V_\mathrm{C}$$

$Q_\mathrm{C} = W_{q\mathrm{C}} x_\mathrm{class} + b_{q\mathrm{C}}$

$$\underbrace{K_\mathrm{C} = W_{k\mathrm{C}} \mathbf{z} + b_{k\mathrm{C}} \qquad V_\mathrm{C} = W_{v\mathrm{C}} \mathbf{z} + b_{v\mathrm{C}}}_{\mathbf{z} = [x_\mathrm{class}, \mathbf{x}^L]}$$

$d_{k\mathrm{C}}$: dimension of K_C

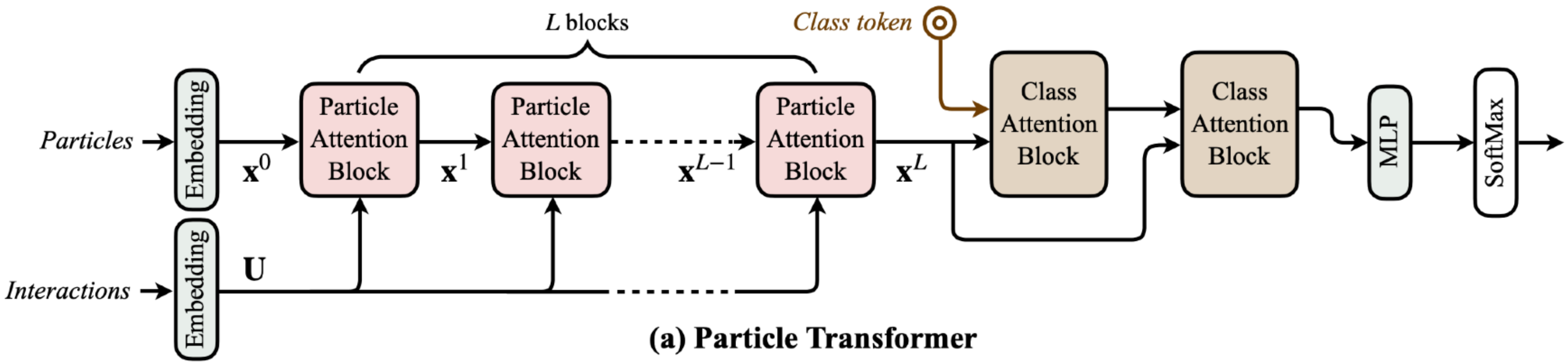
Concatenate class information and particle embedding



(c) Class Attention Block

ParT @ CMS for HF Tagging

- The <u>debut</u> of ParT in CMS: Run3 HF tagging   <u>ArXiv: 2202.03772</u>



(a) Particle Transformer

- ParT Architect for CMS HF tagging:
  - 3 Particle Attention Blocks + 1 Class Attention Block
  - GELU as activation for attention blocks, ReLU as attention for initial embedding + final MLP
  - Number of "Heads" in MHA: 8
  - Number of feature: 128 (for $\mathbf{x}^l$s)
    - For initial embedding the MLPs are (128-512-128) and (64-64-64-8) for *Particles* and *Interactions* respectively.

ParT @ CMS for HF Tagging

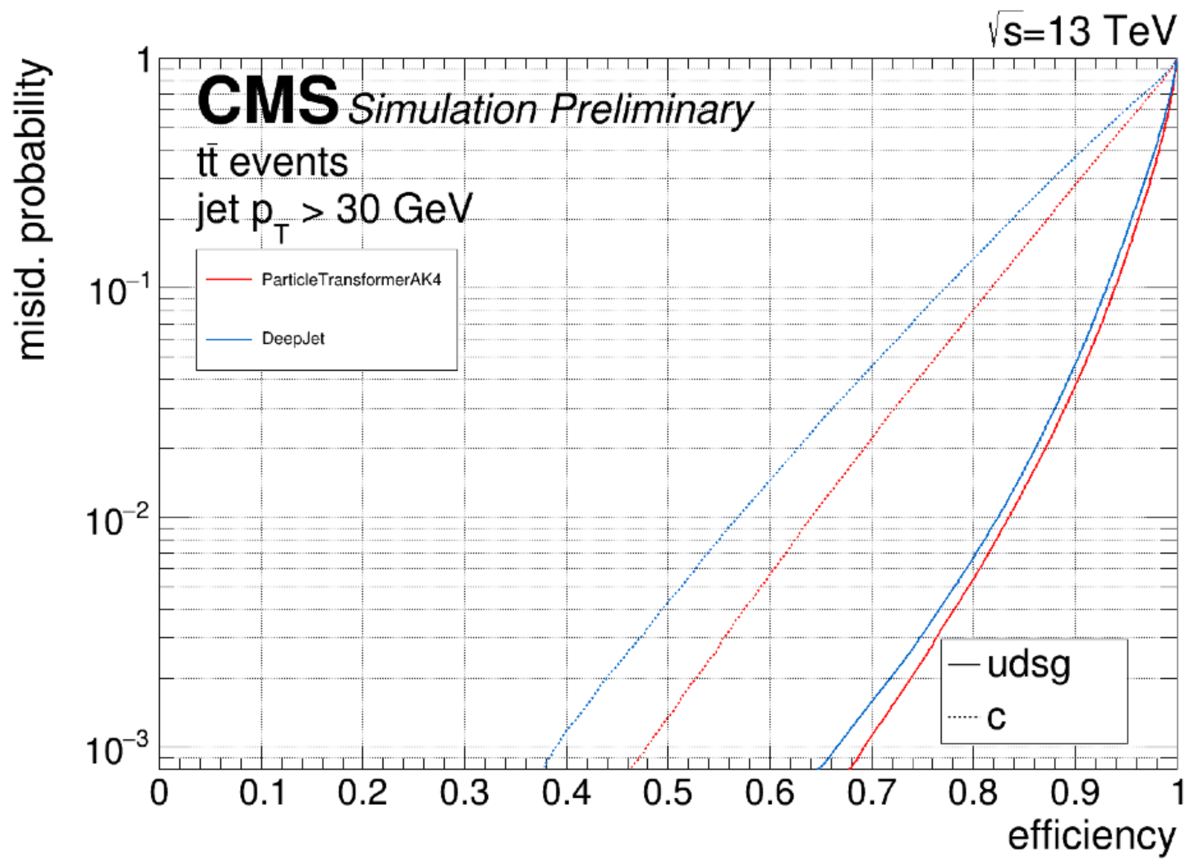
- ParT @ CMS:
  - The input features are almost identical with DeepJet inputs
    - (+): 4 momenta of the constituents; (-): global features.
  - FLOPs: ~117M, #Params: ~1.4M
    - Inference time is similar with ParticleNet's
- Training details:
  - Dataset: ~65M jet from $t\bar{t}$ and QCD multi-jets
  - Ranger optimizer
  - LR initially set to 1e-3 and will decay linearly after 70% of training to reach 1e-5 in the end
  - Batch size: 512
  - Loss function: cross entropy
  - Model with best loss performance on validation set will be kept

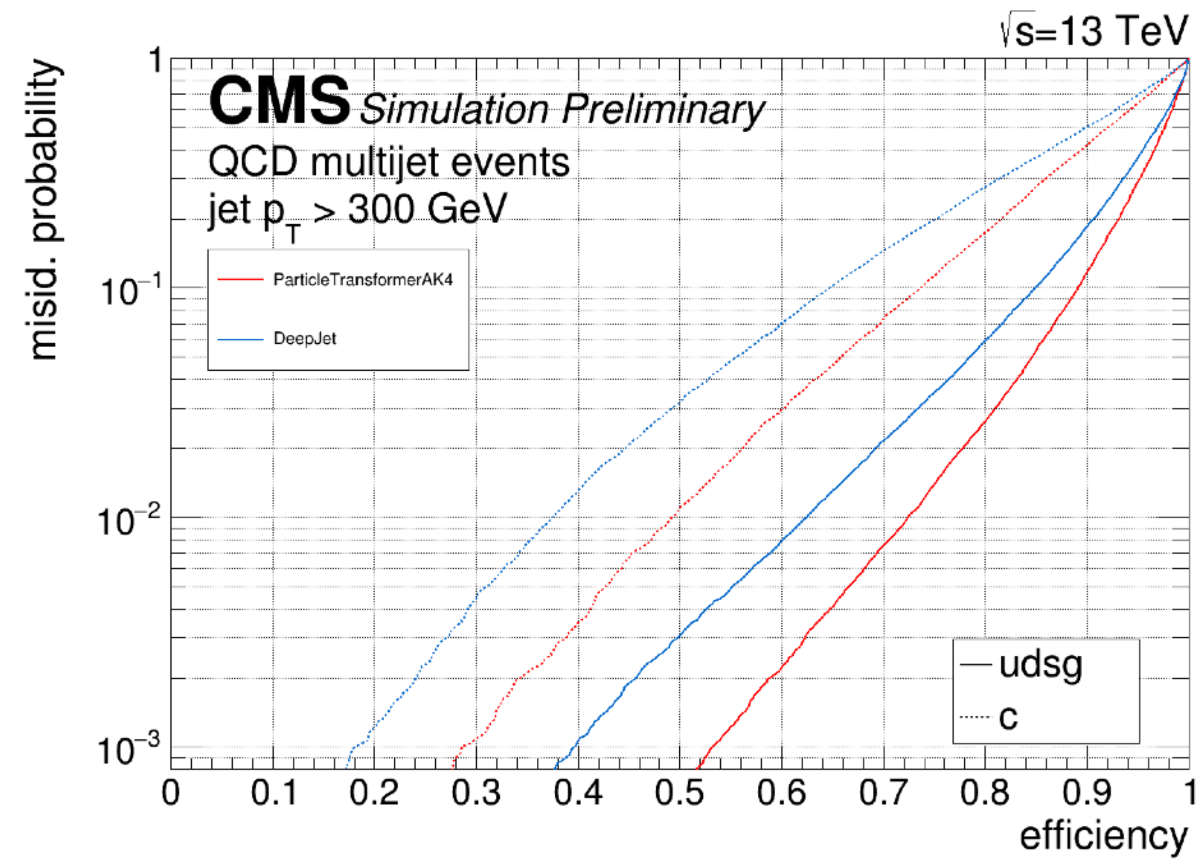
PART @ CMS FOR HF TAGGING

- Performance: b tagging

DP-2022/050

Left: $t\bar{t}$ events, Right: QCD multijets





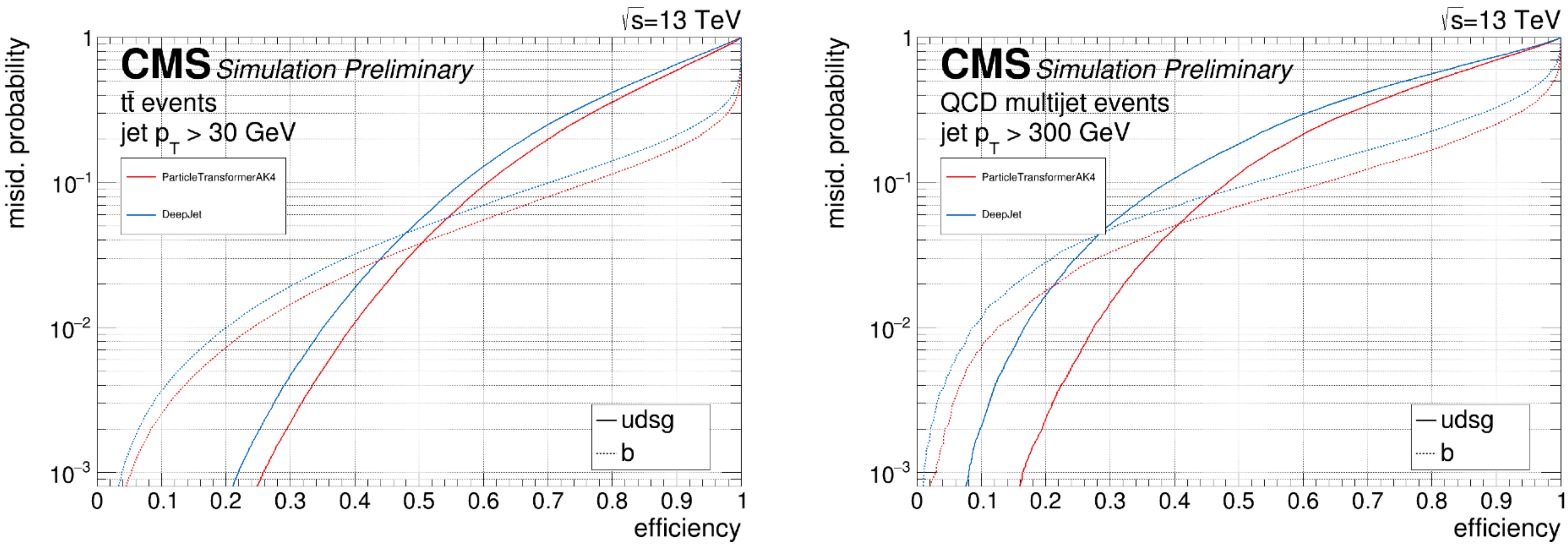
Performance on b tagging of ParT @ CMS:
Solid(dashed) ROC curves indicate the mistagging rates for "udsg"("c") jet at given b tagging signal efficiency

ParT @ CMS for HF Tagging

- Performance: c tagging

DP-2022/050

Left: $t\bar{t}$ events, Right: QCD multijets



Performance on c tagging of ParT @ CMS:
Solid(dashed) ROC curves indicate the mistagging rates for "udsg"("b") jet at given c tagging signal efficiency

ParT @ CMS for HF Tagging

- Summary:
  - Jet tagging is an important yet challenging task in HEP, among which flavor tagging is of vital importance
  - CMS has kept investigating HF tagging with more and more advanced models
  - For Run3, Particle Transformer (ParT) is considered as the official HF tagging algorithm
  - Significant improvement on HF tagging performance is observed from the comparison between ParT and previous CMS State-Of-The-Art model: DeepJet

BACK UP

ParT vs ParticleNet on AK8 Jet

Table 4. Number of trainable parameters and FLOPs.

	Accuracy	# params	FLOPs
PFN	0.772	86.1 k	4.62 M
P-CNN	0.809	354 k	15.5 M
ParticleNet	0.844	370 k	540 M
ParT	**0.861**	2.14 M	340 M
ParT (plain)	0.849	2.13 M	260 M

Table 1. Jet tagging performance on the JetClass dataset. ParT is compared to PFN (Komiske et al., 2019b), P-CNN (CMS Collaboration, 2020b) and the state-of-the-art ParticleNet (Qu & Gouskos, 2020). For all the metrics, a higher value indicates better performance. The ParT architecture using plain MHAs instead of P-MHAs, labelled as ParT (plain), is also shown for comparison.

	All classes		$H \to b\bar{b}$	$H \to c\bar{c}$	$H \to gg$	$H \to 4q$	$H \to \ell\nu qq'$	$t \to bqq'$	$t \to b\ell\nu$	$W \to qq'$	$Z \to q\bar{q}$
	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99.5\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$
PFN	0.772	0.9714	2924	841	75	198	265	797	721	189	159
P-CNN	0.809	0.9789	4890	1276	88	474	947	2907	2304	241	204
ParticleNet	0.844	0.9849	7634	2475	104	954	3339	10526	11173	347	283
ParT	**0.861**	**0.9877**	**10638**	**4149**	**123**	**1864**	**5479**	**32787**	**15873**	**543**	**402**
ParT (plain)	0.849	0.9859	9569	2911	112	1185	3868	17699	12987	384	311

ParT vs ParticleNet on AK8 Jet

Table 3. Impacts of the training dataset size. Entries in bold correspond to the training using the full 100 M training dataset.

	All classes		$H \to b\bar{b}$	$H \to c\bar{c}$	$H \to gg$	$H \to 4q$	$H \to \ell\nu qq'$	$t \to bqq'$	$t \to b\ell\nu$	$W \to qq'$	$Z \to q\bar{q}$
	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99.5\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$
ParticleNet (2 M)	0.828	0.9820	5540	1681	90	662	1654	4049	4673	260	215
ParticleNet (10 M)	0.837	0.9837	5848	2070	96	770	2350	5495	6803	307	253
ParticleNet (100 M)	0.844	0.9849	7634	2475	104	954	3339	10526	11173	347	283
ParT (2 M)	0.836	0.9834	5587	1982	93	761	1609	6061	4474	307	236
ParT (10 M)	0.850	0.9860	8734	3040	110	1274	3257	12579	8969	431	324
ParT (100 M)	0.861	0.9877	10638	4149	123	1864	5479	32787	15873	543	402

Table 6. Comparison between ParT and existing models on the quark-gluon tagging dataset. ParT refers to the model trained from scratch on this dataset. ParticleNet-f.t. and ParT-f.t. denote the corresponding models pre-trained on JetClass and fine-tuned on this dataset. Results for other models are quoted from their published results: P-CNN and ParticleNet (Qu & Gouskos, 2020), PFN (Komiske et al., 2019b), ABCNet (Mikuni & Canelli, 2020), PCT (Mikuni & Canelli, 2021), rPCN (Shimmin, 2021), and LorentzNet (Gong et al., 2022). The subscript "exp" and "full" distinguish models using partial or full particle identification information.

	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{30\%}$
$\text{P-CNN}_{\text{exp}}$	0.827	0.9002	34.7	91.0
PFN_{exp}	—	0.9005	34.7 ± 0.4	—
$\text{ParticleNet}_{\text{exp}}$	0.840	0.9116	39.8 ± 0.2	98.6 ± 1.3
rPCN_{exp}	—	0.9081	38.6 ± 0.5	—
ParT_{exp}	0.840	0.9121	41.3 ± 0.3	101.2 ± 1.1
$\text{ParticleNet-f.t.}_{\text{exp}}$	0.839	0.9115	40.1 ± 0.2	100.3 ± 1.0
$\textbf{ParT-f.t.}_{\textbf{exp}}$	**0.843**	**0.9151**	$\mathbf{42.4 \pm 0.2}$	$\mathbf{107.9 \pm 0.5}$
PFN_{full}	—	0.9052	37.4 ± 0.7	—
$\text{ABCNet}_{\text{full}}$	0.840	0.9126	42.6 ± 0.4	118.4 ± 1.5
PCT_{full}	0.841	0.9140	43.2 ± 0.7	118.0 ± 2.2
$\text{LorentzNet}_{\text{full}}$	0.844	0.9156	42.4 ± 0.4	110.2 ± 1.3
$\text{ParT}_{\text{full}}$	0.849	0.9203	47.9 ± 0.5	129.5 ± 0.9
$\textbf{ParT-f.t.}_{\textbf{full}}$	**0.852**	**0.9230**	$\mathbf{50.6 \pm 0.2}$	$\mathbf{138.7 \pm 1.3}$

Table 5. Comparison between ParT and existing models on the top quark tagging dataset. ParT refers to the model trained from scratch on this dataset. ParticleNet-f.t. and ParT-f.t. denote the corresponding models pre-trained on JetClass and fine-tuned on this dataset. Results for other models are quoted from their published results: P-CNN and ParticleNet (Qu & Gouskos, 2020), PFN (Komiske et al., 2019b), JEDI-net (Moreno et al., 2020), PCT (Mikuni & Canelli, 2021), LGN (Bogatskiy et al., 2020), rPCN (Shimmin, 2021), and LorentzNet (Gong et al., 2022).

	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{30\%}$
P-CNN	0.930	0.9803	201 ± 4	759 ± 24
PFN	—	0.9819	247 ± 3	888 ± 17
ParticleNet	0.940	0.9858	397 ± 7	1615 ± 93
JEDI-net (w/ $\sum O$)	0.930	0.9807	—	774.6
PCT	0.940	0.9855	392 ± 7	1533 ± 101
LGN	0.929	0.964	—	435 ± 95
rPCN	—	0.9845	364 ± 9	1642 ± 93
LorentzNet	0.942	0.9868	498 ± 18	2195 ± 173
ParT	0.940	0.9858	413 ± 16	1602 ± 81
ParticleNet-f.t.	0.942	0.9866	487 ± 9	1771 ± 80
ParT-f.t.	**0.944**	**0.9877**	$\mathbf{691 \pm 15}$	$\mathbf{2766 \pm 130}$