ParT

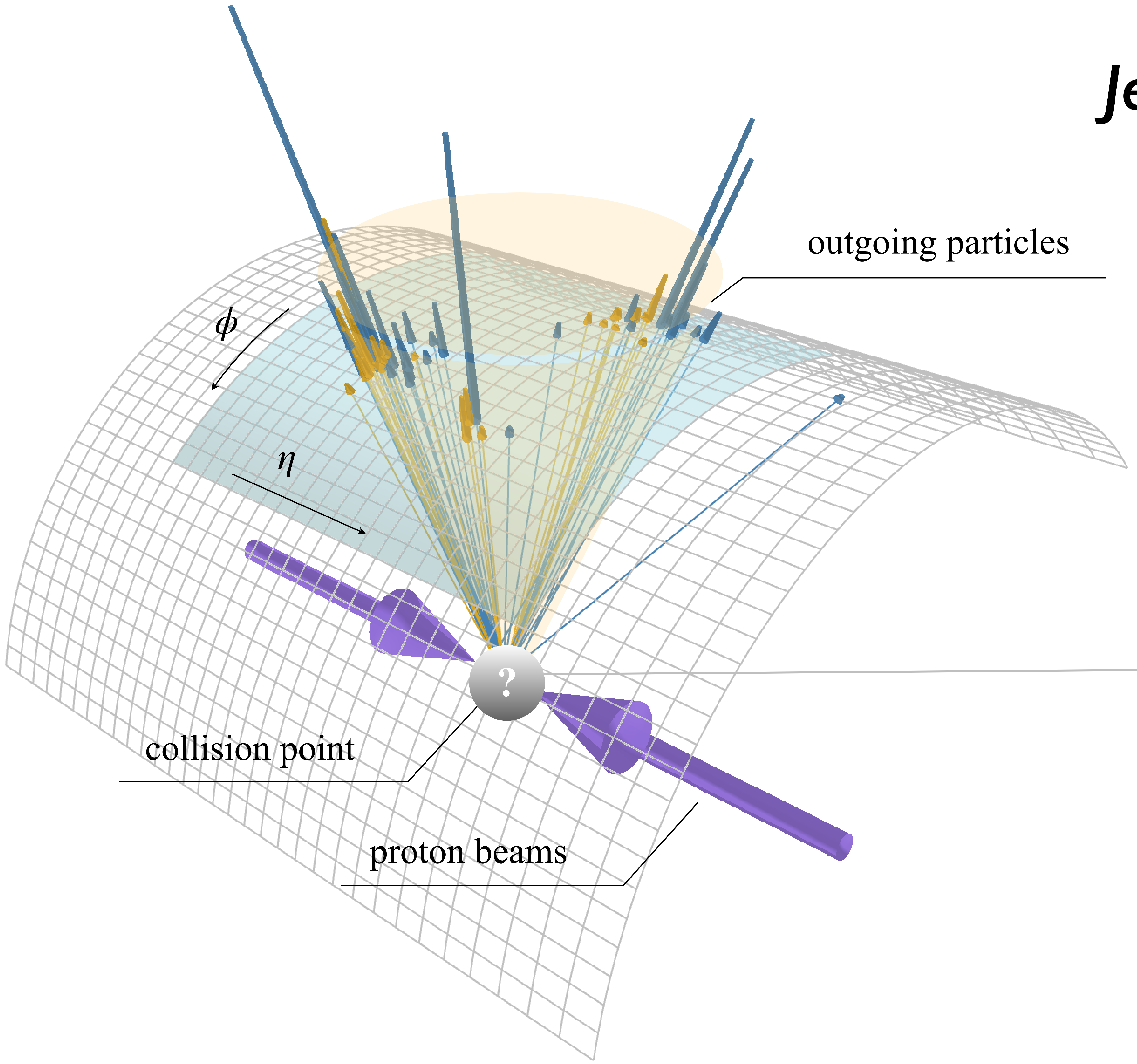
Particle Transformer for Jet Tagging

Congqiao Li (PKU)

BOOST 2022
August 17, 2022

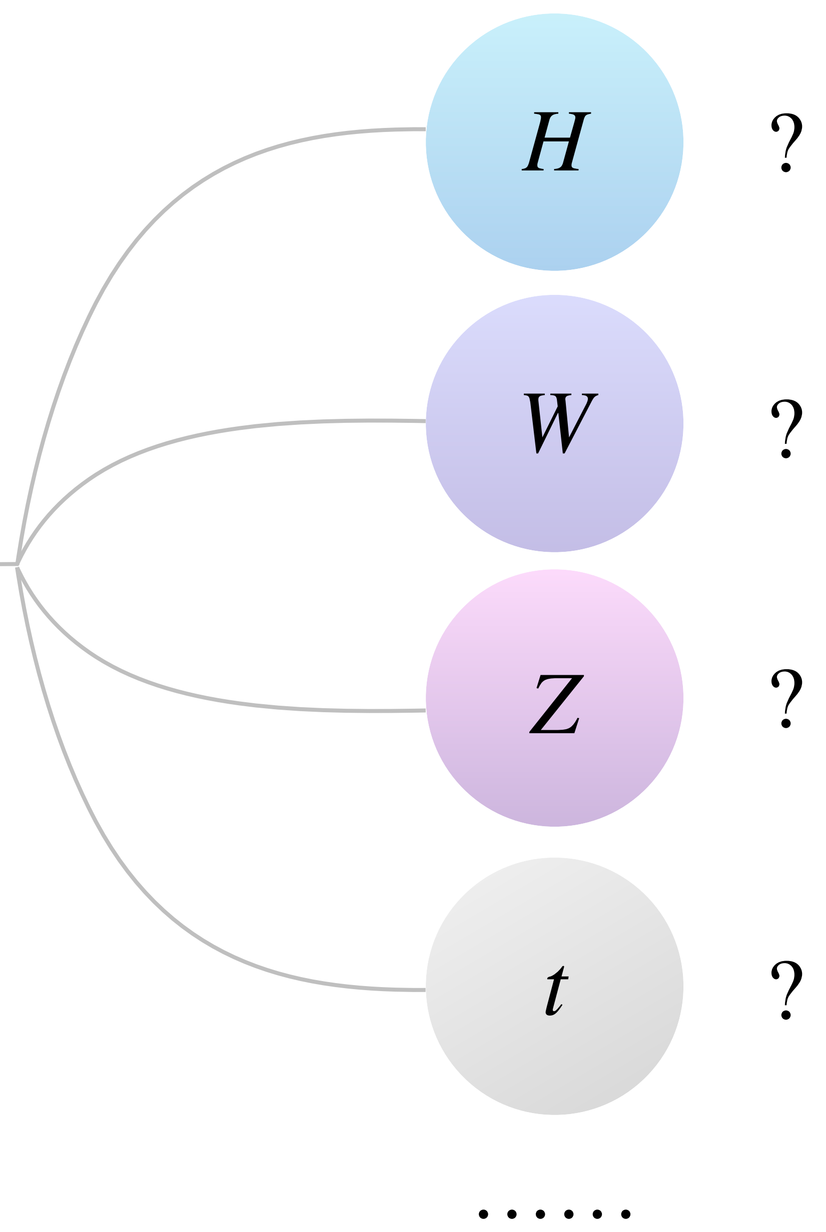
Based on Huilin Qu, Congqiao Li & Sitian Qian, arXiv: 2202.03772

Introduction



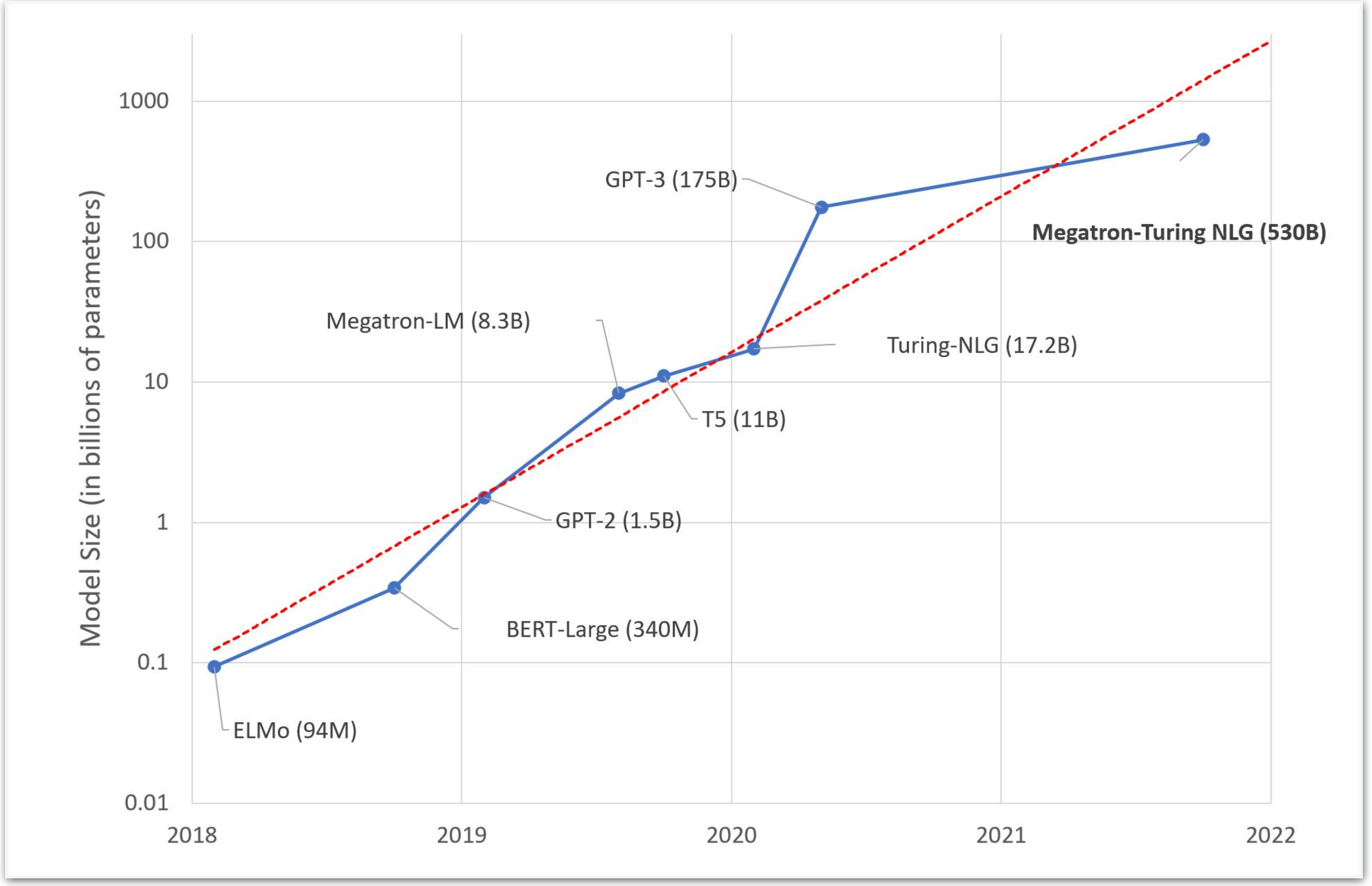
Jet Tagging

- *Powerful handle to search for new phenomena*
- *Significant progress thanks to advanced ML*
- ***But, can we do even better?***

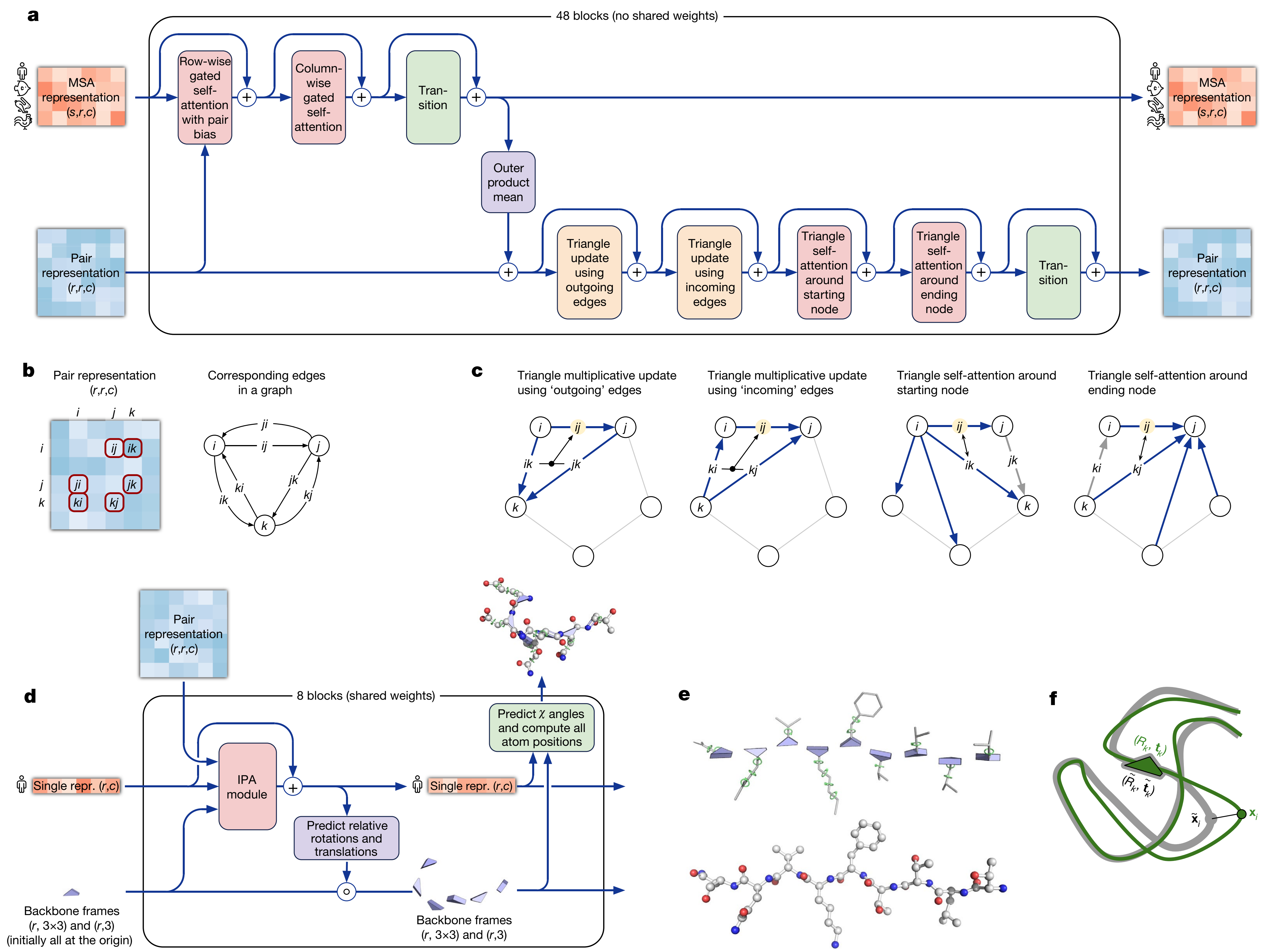


Transformers in Action

- **Attention mechanism and Transformers**: the new state-of-the-art architecture in ML
  - Large Language Models: BERT, GPT-3, …
  - Computer Vision: ViT, Swin-T, …
  - AlphaFold2 for protein structure prediction



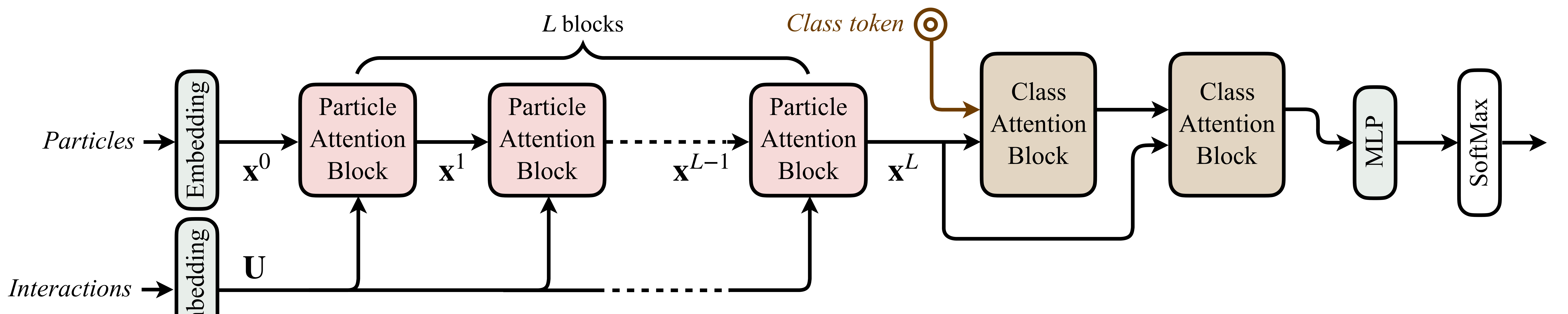
Large Language Models: A New Moore's Law?



AlphaFold2: predicting protein structures with atomic accuracy

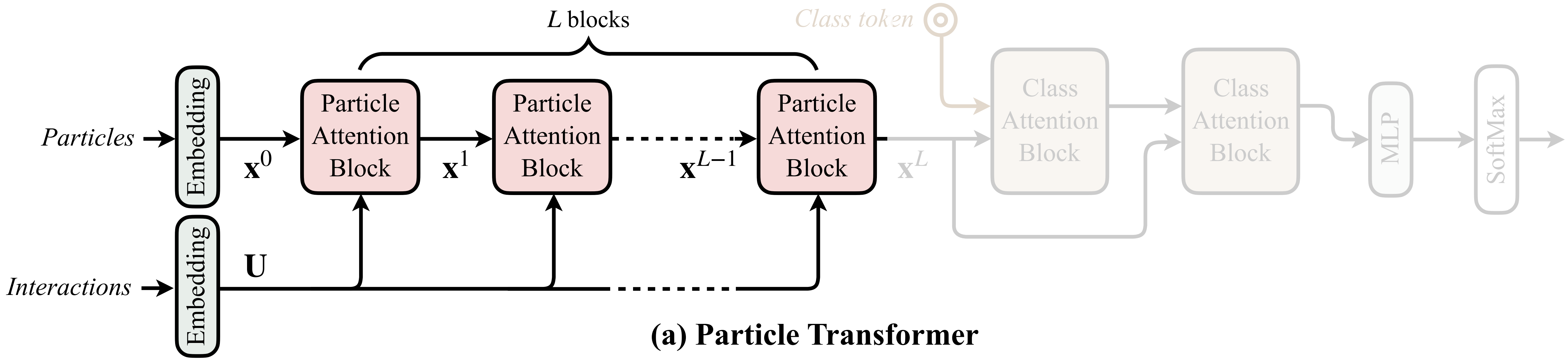
Transformer Meets Jet Tagging

- **Particle Transformer** (ParT)
    - Transformer architecture tailored for particle physics
    - capable of processing not only single particle information, but also **pairwise information**



(a) Particle Transformer

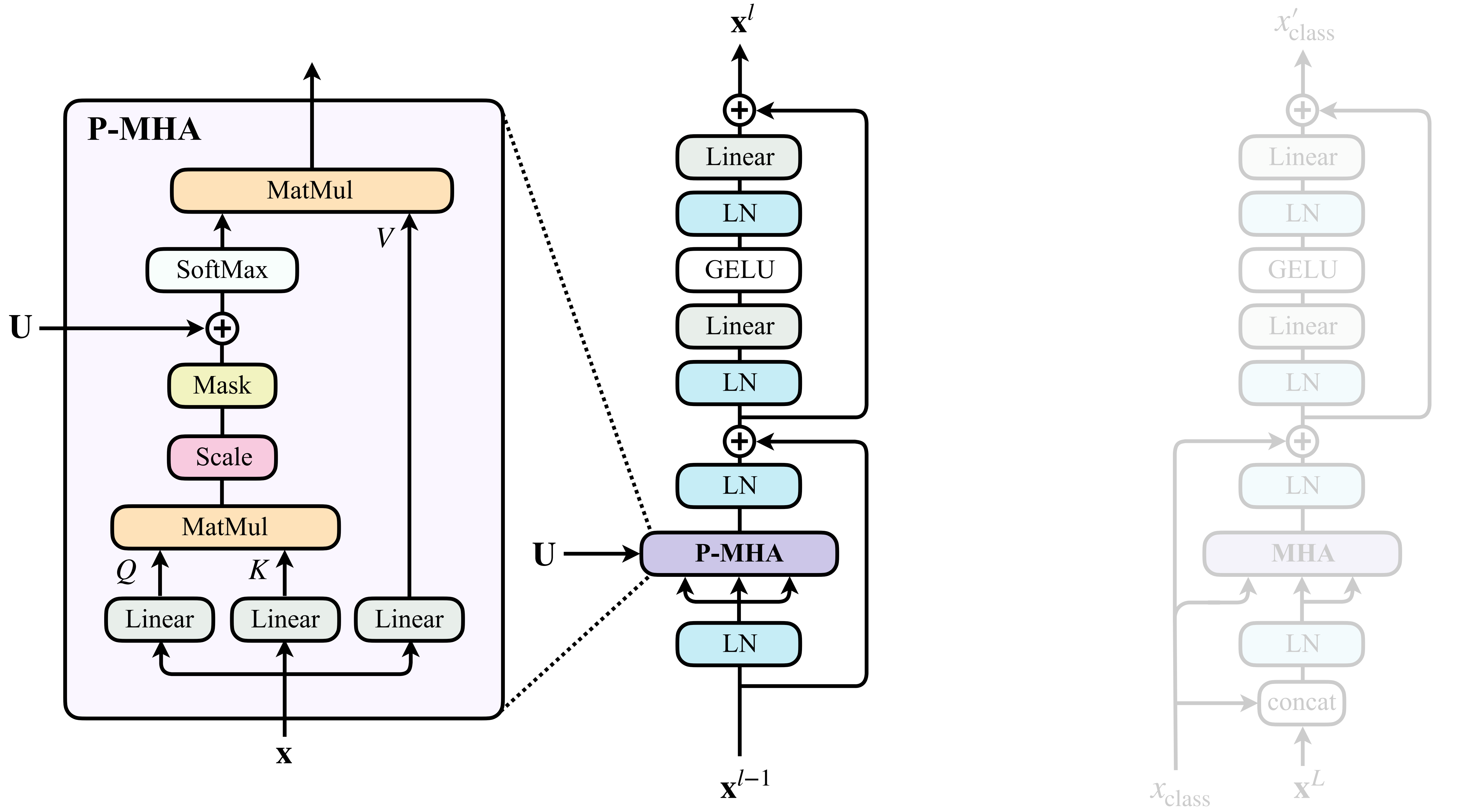
Particle Attention Block



(a) Particle Transformer

Particle Attention Blocks

Fully exploit the correlations between particles



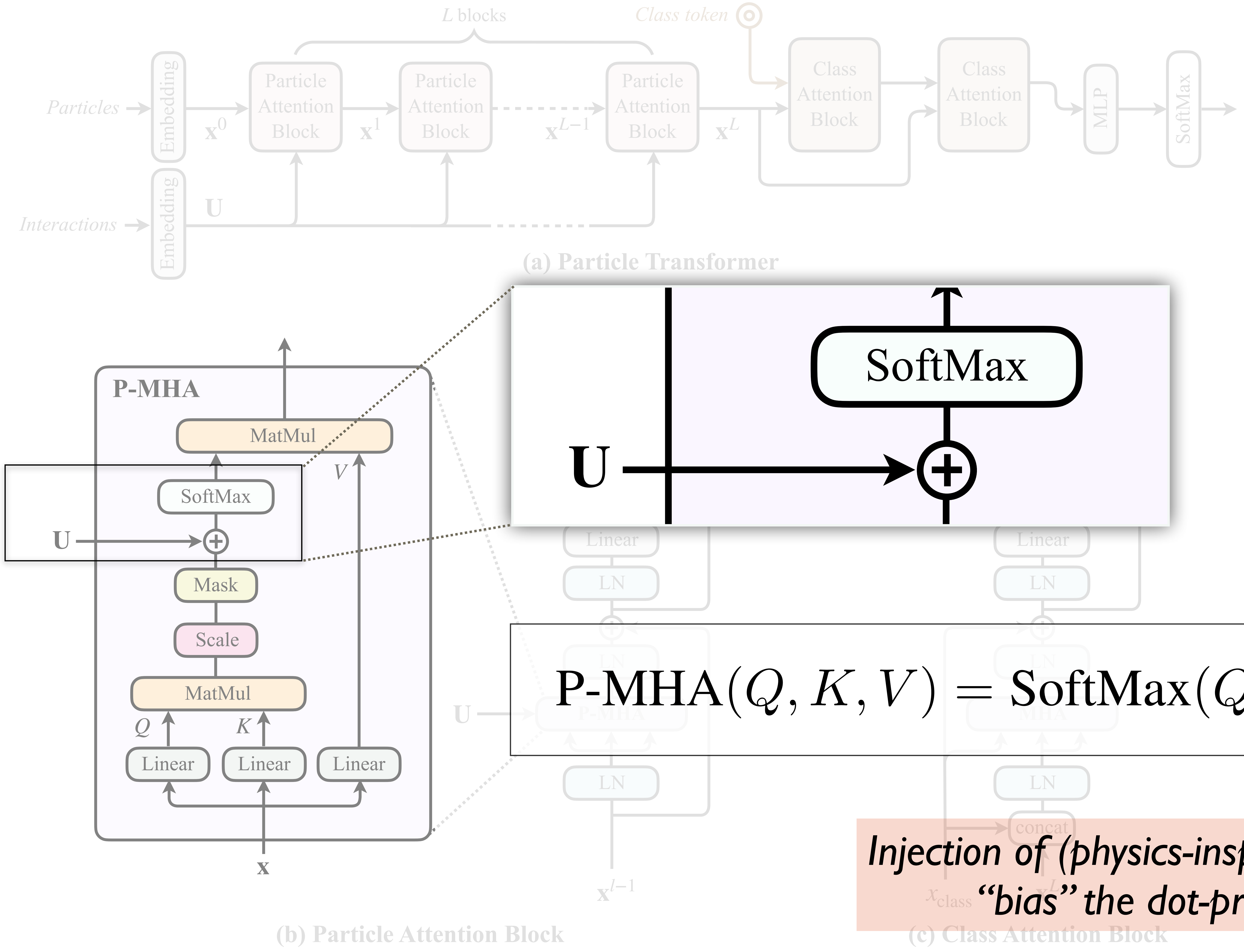
(b) Particle Attention Block

(c) Class Attention Block

Particle Attention Block

Particle Attention Blocks

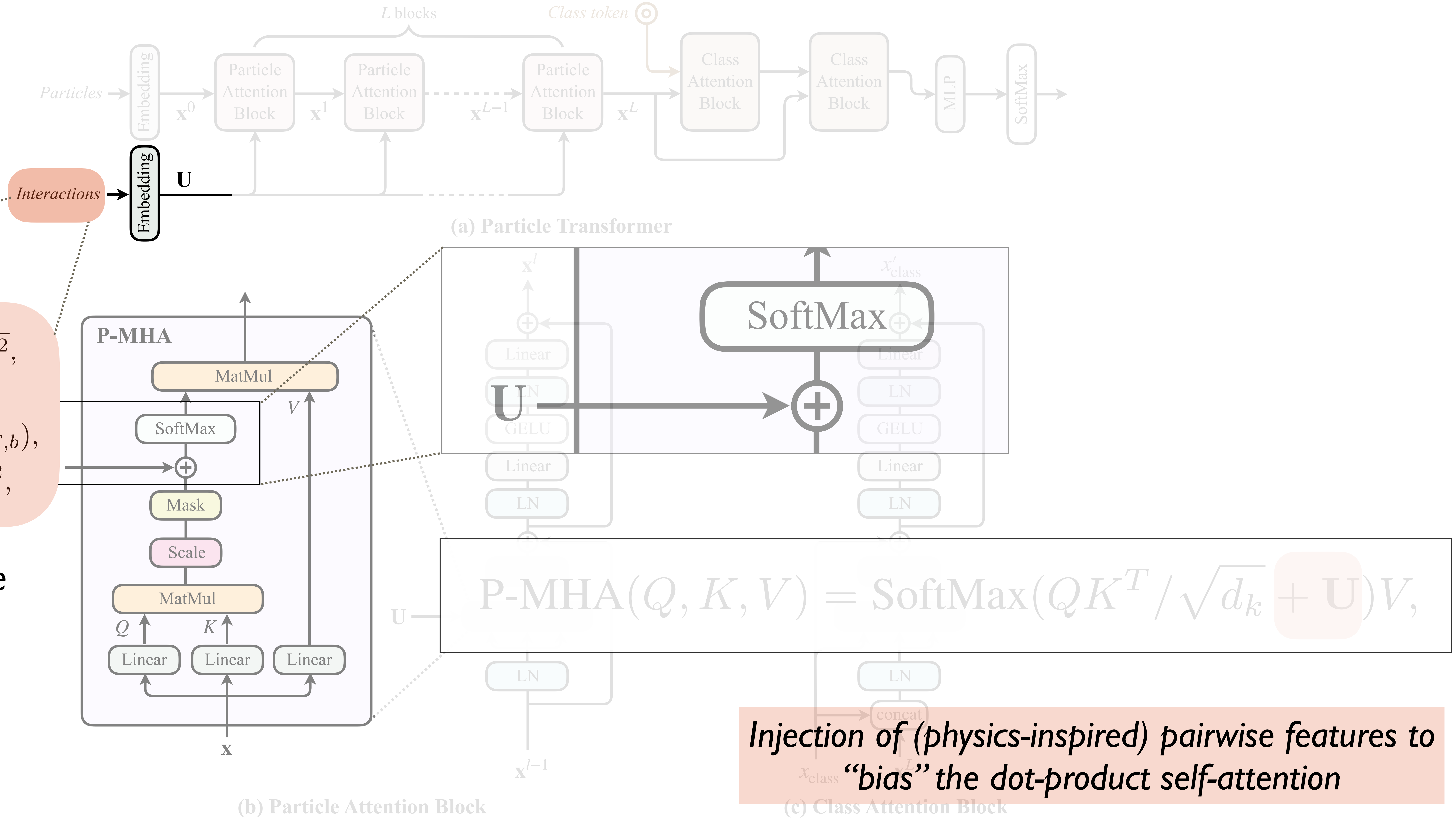
Fully exploit the correlations between particles



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

Injection of (physics-inspired) pairwise features to "bias" the dot-product self-attention

Particle Attention Block



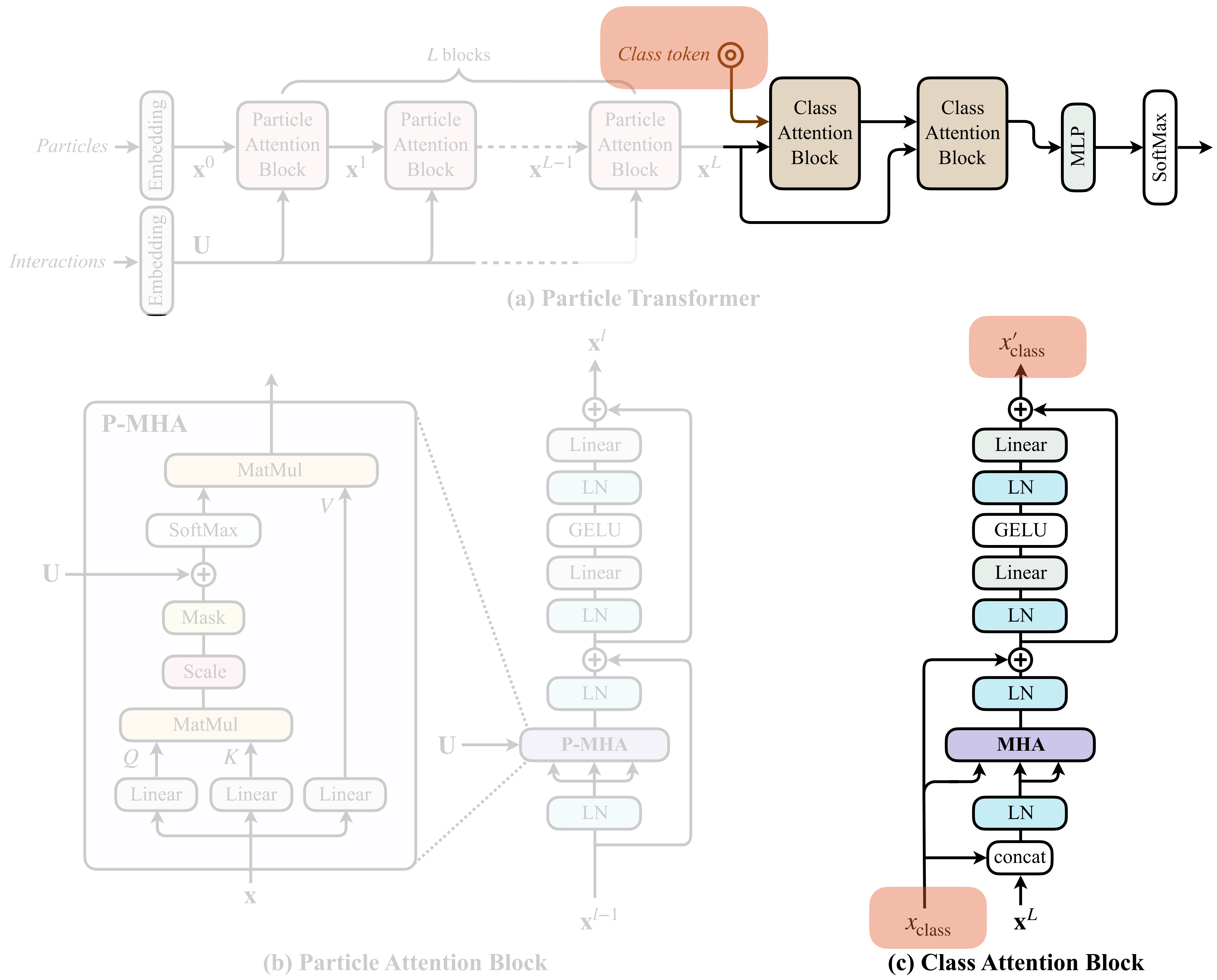
$$\text{P-MHA}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d_k} + \mathbf{U})V,$$

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

and many other possible pairwise features…

Injection of (physics-inspired) pairwise features to "bias" the dot-product self-attention

Class Attention Block



Class Attention Blocks

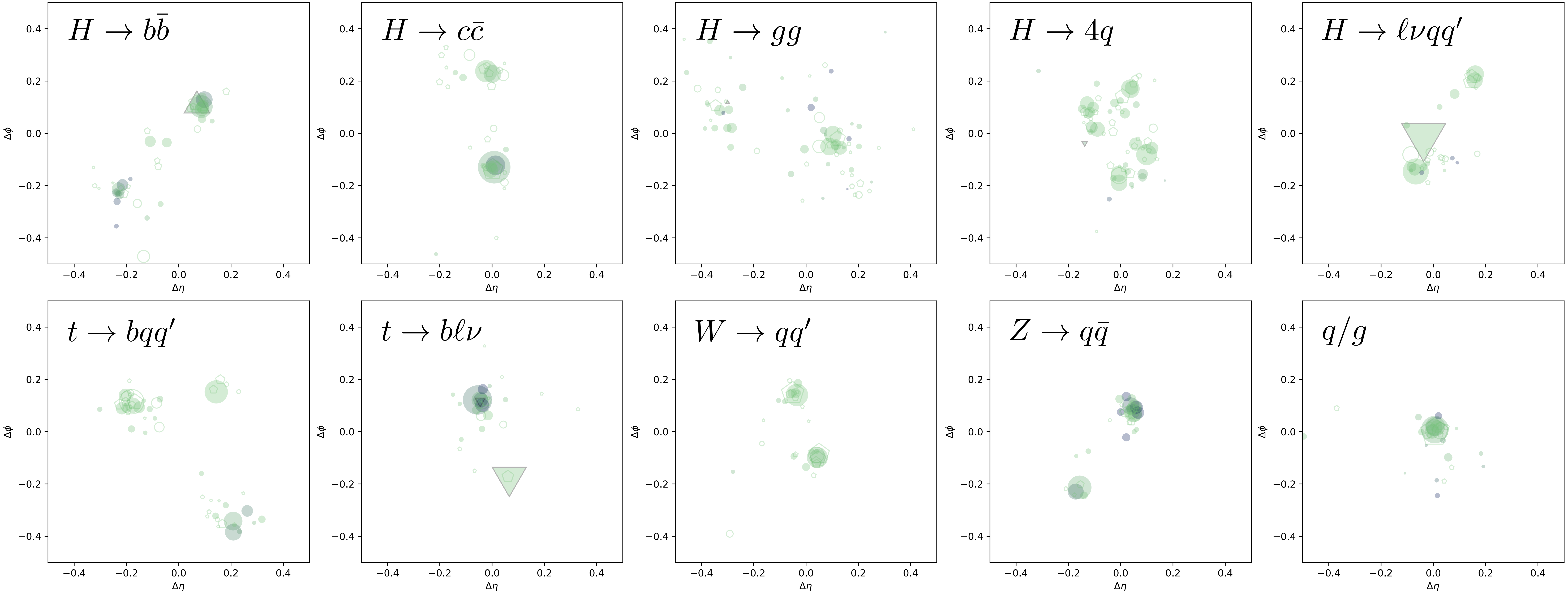
Extract features from each particle via a class token

Class token encodes the overall information — fed into MLP for final classification

Large Model Calls For Larger Dataset

- JETCLASS: a new large and comprehensive jet dataset
  - 100M jets for training: ~two orders of magnitude larger than existing public datasets
  - 10 classes: several unexplored scenarios (e.g., H->WW*->4q, H->WW*->ℓvqq, etc.)
  - a rich set of features for each particle: kinematics + particle identification + track displacement

Simulated w/ MadGraph + Pythia + Delphes



Performance on JetClass Dataset

	All classes Accuracy	AUC	$H \to b\bar{b}$ Rej$_{50\%}$	$H \to c\bar{c}$ Rej$_{50\%}$	$H \to gg$ Rej$_{50\%}$	$H \to 4q$ Rej$_{50\%}$	$H \to \ell\nu qq'$ Rej$_{99\%}$	$t \to bqq'$ Rej$_{50\%}$	$t \to b\ell\nu$ Rej$_{99.5\%}$	$W \to qq'$ Rej$_{50\%}$	$Z \to q\bar{q}$ 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

- Particle Transformer (ParT): significant performance improvement!
  - compared to the existing state-of-the-art, ParticleNet
    - 1.7% increase in accuracy
    - **up to 3x increase in background rejection** (Rej$_{X\%}$)

$$\boxed{\text{Rej}_{X\%} = 1/\epsilon_B @ \epsilon_S = X\%}$$

Performance on JETCLASS Dataset

	All classes Accuracy	AUC	$H \to b\bar{b}$ $\text{Rej}_{50\%}$	$H \to c\bar{c}$ $\text{Rej}_{50\%}$	$H \to gg$ $\text{Rej}_{50\%}$	$H \to 4q$ $\text{Rej}_{50\%}$	$H \to \ell\nu qq'$ $\text{Rej}_{99\%}$	$t \to bqq'$ $\text{Rej}_{50\%}$	$t \to b\ell\nu$ $\text{Rej}_{99.5\%}$	$W \to qq'$ $\text{Rej}_{50\%}$	$Z \to q\bar{q}$ $\text{Rej}_{50\%}$
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

- Particle Transformer (ParT): significant performance improvement!
  - compared to the existing state-of-the-art, ParticleNet
    - 1.7% increase in accuracy
    - **up to 3x increase in background rejection** ($\text{Rej}_{X\%}$)
- ParT (plain): plain Transformer w/o interaction features
  - significant performance drop: barely outperforms ParticleNet
  - **Physics-driven modification of self-attention plays a key role!**

Performance on JETCLASS 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	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99.5\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$
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

- Particle Transformer (ParT): significant performance improvement!
  - compared to the existing state-of-the-art, ParticleNet
    - 1.7% increase in accuracy
    - **up to 3x increase in background rejection** ($\text{Rej}_{X\%}$)
- ParT (plain): plain Transformer w/o interaction features
  - significant performance drop: barely outperforms ParticleNet
  - **Physics-driven modification of self-attention plays a key role!**

Model complexity

	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

- Computation cost still under control.

Performance Vs Dataset Size

	All classes Accuracy	AUC	$H \to b\bar{b}$ $\text{Rej}_{50\%}$	$H \to c\bar{c}$ $\text{Rej}_{50\%}$	$H \to gg$ $\text{Rej}_{50\%}$	$H \to 4q$ $\text{Rej}_{50\%}$	$H \to \ell\nu qq'$ $\text{Rej}_{99\%}$	$t \to bqq'$ $\text{Rej}_{50\%}$	$t \to b\ell\nu$ $\text{Rej}_{99.5\%}$	$W \to qq'$ $\text{Rej}_{50\%}$	$Z \to q\bar{q}$ $\text{Rej}_{50\%}$
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

- Large training dataset: a crucial factor for performance
  - e.g., ParT performance on H->cc tagging
    - 2M->10M: **~50% increase** in background rejection
    - 10M->100M: **~35% further increase**
- Performance of ParT scales better as dataset size increases
  - due to its larger model capacity than ParticleNet

Pre-Training + Fine-Tuning

- The large-scale JetClass dataset enables new training paradigm
  - (supervised) pre-training on JetClass & fine-tuning on downstream tasks
  - significantly outperforms existing models

Top quark tagging benchmark (~2M jets) [SciPost Phys. 7 (2019) 014]

	Accuracy	AUC	$\mathrm{Rej}_{50\%}$	$\mathrm{Rej}_{30\%}$
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**	$\mathbf{691 \pm 15}$	$\mathbf{2766 \pm 130}$

Quark-gluon tagging benchmark (~2M jets) [JHEP 01 (2019) 121]

	Accuracy	AUC	$\mathrm{Rej}_{50\%}$	$\mathrm{Rej}_{30\%}$
$\text{P-CNN}_{\text{exp}}$	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**	$\mathbf{50.6 \pm 0.2}$	$\mathbf{138.7 \pm 1.3}$

Summary

- **JetClass**: large-scale open dataset for deep-learning research in jet physics
- **Particle Transformer**: new architecture for jet tagging with substantially improved performance

JetClass: More possibilities ahead

We invite the community to explore and experiment with this dataset and extend the boundary of deep learning and jet physics even further.

BACKUPS

Input Features

Table 2. Particle input features used for jet tagging on the JETCLASS, the top quark tagging (TOP) and the quark gluon tagging (QG) datasets. For QG, we consider two scenarios: QG_{exp} is restricted to use only the 5-class experimentally realistic particle identification information, while QG_{full} uses the full set of particle identification information in the dataset and further distinguish between different types of charged hadrons and neutral hadrons.

Category	Variable	Definition	JETCLASS	TOP	QG_{exp}	QG_{full}
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	✓	✓	✓	✓
	$\log p_T$	logarithm of the particle's transverse momentum p_T	✓	✓	✓	✓
	$\log E$	logarithm of the particle's energy	✓	✓	✓	✓
	$\log \frac{p_T}{p_T(\text{jet})}$	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`)	✓	—	✓	✓[a]
	`NH`	if the particle is an neutral hadron (`|pid|==130 or 2112 or 0`)	✓	—	✓	✓[b]
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	✓	—	—	—

[a] `(|pid|==211) + (|pid|==321)*0.5 + (|pid|==2212)*0.2`

[b] `(|pid|==130) + (|pid|==2112)*0.2.`