Attention to the strengths of physics interactions

Enhanced Deep Learning Event Classification for Particle Physics Experiments

Polina Moskvitina

Sascha Caron, Clara Nellist, Roberto Ruiz de Austri, Rob Verheyen, Zhongyi Zhang

European AI for Fundamental Physics Conference, 30 April 2024

Our group

Sascha
(Supervisor,
RU/Nikhef)

Clara
(Co-supervisor,
RU/Nikhef)

Polina
(PhD,
RU/Nikhef)

From DarkMachines :

Zhongyi
(Postdoc,
RU/Nikhef)

Rob
(Postdoc,
UCL)

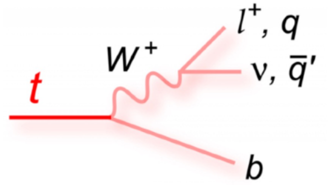
Roberto
(Researcher,
IFIC)

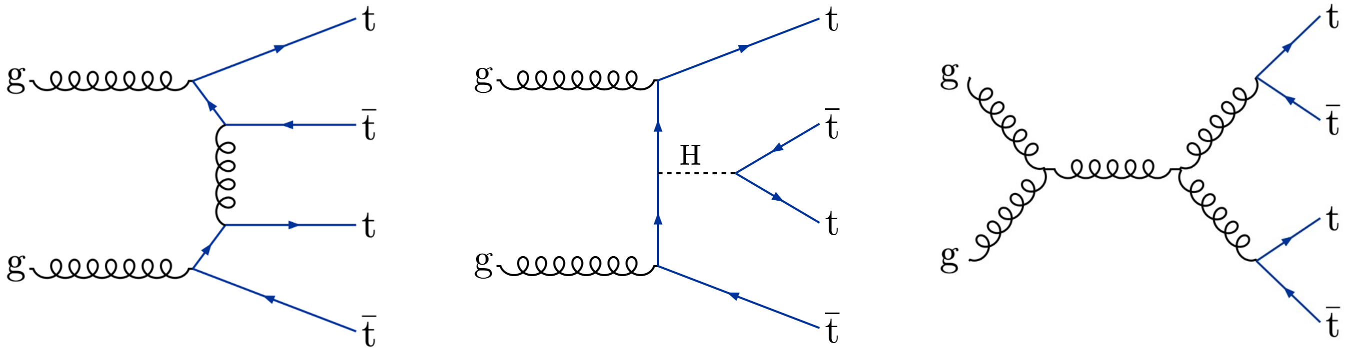
The question was: "**What is *the best event classifier* for the LHC?**"

The four-top-quarks and $t\bar{t}H$ production at LHC

Production of **four top quarks** is very rare

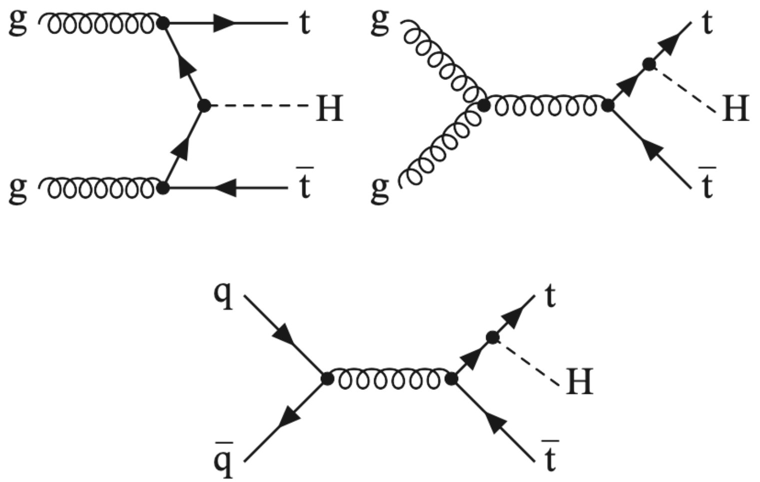
- **NLO QCD:** $\sigma(t\bar{t}t\bar{t}) = 12\ \mathrm{fb} \pm 20\%$ [JHEP02(2018)031]
- **NLO+NLL:** $\sigma(t\bar{t}t\bar{t}) = 13.4\ \mathrm{fb} \pm 11\%$ [arXiv:2212.03259]





Examples of Feynman diagrams for SM $t\bar{t}t\bar{t}$ production at leading order in QCD and via an off-shell Higgs boson mediator

The **Top-top-Higgs**
has a small cross section (1/100 ggF)
$\sigma(t\bar{t}H) \sim 0.507\ \mathrm{pb}$



Example tree-level Feynman diagrams for the pp $\rightarrow t\bar{t}H$

First observation of $t\bar{t}t\bar{t}$ production with an observed (expected) significance of **6.1σ (4.3σ)** with **GNN** by **ATLAS** [Eur. Phys. J. C 83, 496 (2023)]
5.6σ (4.9σ) with **BDT** by **CMS** [Phys. Lett. B 847 (2023) 138290]

Observation of $t\bar{t}H$ production
6.3σ (5.1σ) by **ATLAS** Phys. Lett. B 784 (2018) 173]
5.2σ (4.2σ) by **CMS** [Phys. Rev. Lett. 120, 231801]

The four-top decays and Background composition

Simulated pp Collisions at $\sqrt{S} = 13$ TeV

The most sensitive channel for **four-top** is:

- **Multilepton final state:**
  2 Leptons Same Sign and 3 Leptons (2LSS/3L),
  13% branching ration, highest sensitivity – observation

Signal region:
$\geq$ 6 jets $\geq$ 2b-jets and $\mathbf{H_T}$ $\geq$500 GeV

Signal process:
- $t\bar{t}t\bar{t}$

Physical backgrounds:
- $t\bar{t}Z$, $\boxed{t\bar{t}H,}$ $t\bar{t}W$, $t\bar{t}WW$

$$\text{event ID; process ID; weight; } \not{E}_T;\ \phi_{\not{E}_T};\ \text{obj}_1, E_1, p_{T_1}, \eta_1, \phi_1;\ \text{obj}_2, E_2, p_{T_2}, \eta_2, \phi_2; \ldots$$

- **All other kinematic variables can be calculated from four-vectors**

Later, it is used for a second analysis as a signal (see slide 11)

	jets	b-jets	e^-	e^+	μ^-	μ^+	γ	$N_{\max}$
FCN, BDT	4	4	1	1	1	1		12
CNN, PN, ParT			no limits					18

$N_{\max}$ – the maximum number of objects in an event

Summary of ML model details

arXiv:2211.05143

Variables per particle
E, p_T, η, ϕ, jet$_{\text{tag}}$, b-jet$_{\text{tag}}$, e^-_{tag}, e^+_{tag}, μ^-_{tag}, μ^+_{tag}, γ_{tag}

Also receives $\not{E}_T$; $\phi_{\not{E}_T}$

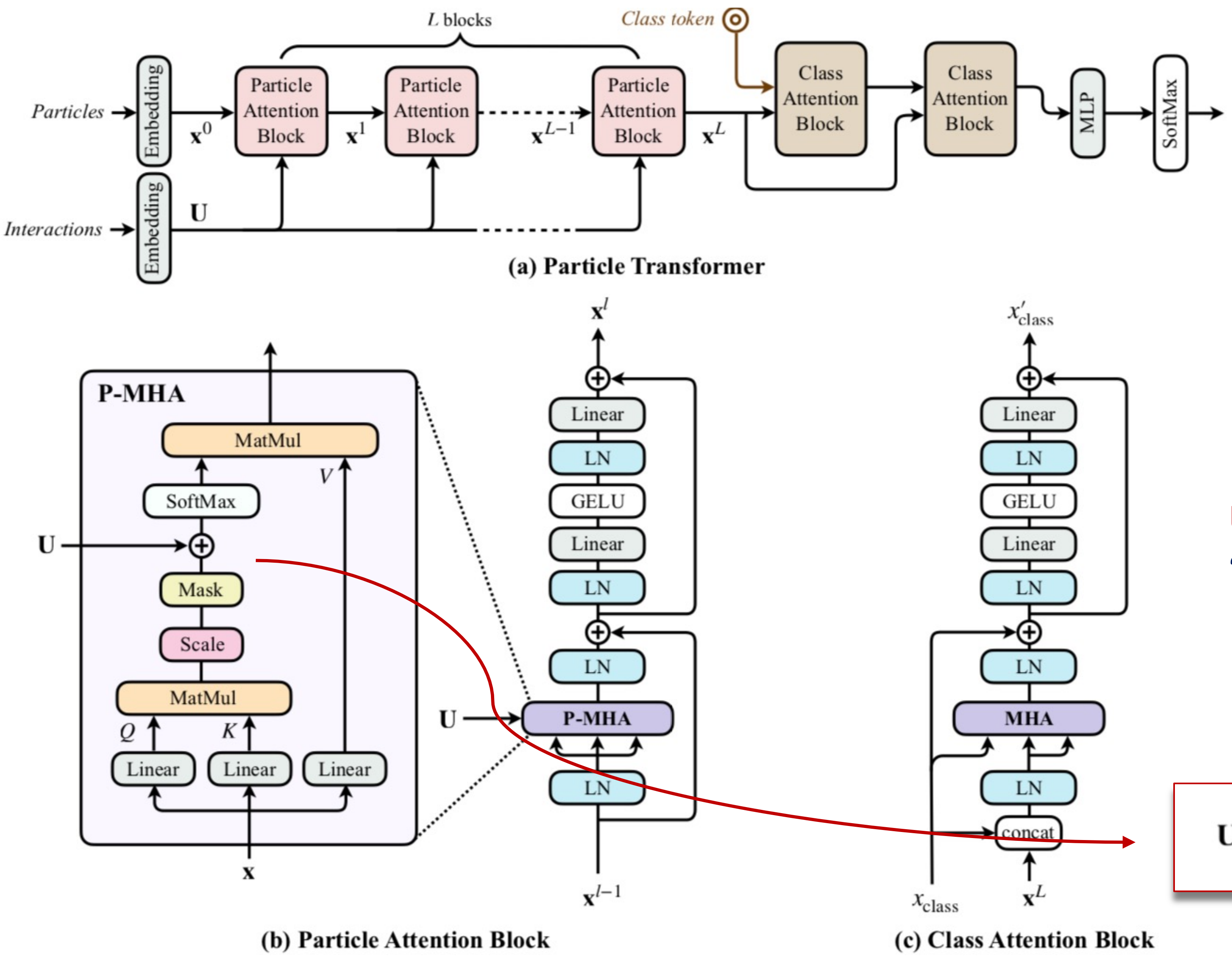
NN structure	Pairwise kinematic features	Loss function
BDT		Cross-entropy
BDT$_{\text{int.}}$	m_{ij}, ΔR_{ij}	
FCN		
CNN		
PN		
PN$_{\text{int.}}$	m_{ij}, ΔR_{ij}	
PN$_{\text{int. SMids}}$	m_{ij}, ΔR_{ij} + SM matrix[1]	
PN$_{\text{int. SM const}}$	m_{ij}, ΔR_{ij} + SM matrix[2]	
PN$_{\text{int. SM}}$	m_{ij}, ΔR_{ij} + SM matrix[3]	
ParT		
ParT$_{\text{int.}}$	m_{ij}, ΔR_{ij}	
ParT$_{\text{int. SM (FL)}}$	m_{ij}, ΔR_{ij} + SM matrix[3]	Focal [$\alpha = 0.75$, $\gamma = 3$]
ParT$_{\text{int. SMids}}$	m_{ij}, ΔR_{ij} + SM matrix[1]	Cross-entropy
ParT$_{\text{int. SM const}}$	m_{ij}, ΔR_{ij} + SM matrix[2]	
ParT$_{\text{int. SM}}$	m_{ij}, ΔR_{ij} + SM matrix[3]	
SetT$_{\text{int. SM}}$	m_{ij}, ΔR_{ij} + SM matrix[3]	

The **particle input** variables and **pairwise kinematic features** that were used in the **NN structures**, each with their respective **loss function**

16 MODELS IN TOTAL!

Transformers

[arXiv:2202.03772](https://arxiv.org/abs/2202.03772)



The architecture of (a) Particle Transformer (b) Particle Attention Block (c) Class Attention Block

Attention Modules

(scaled dot product attention):

- $Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + \boldsymbol{U}\right)V$
- $Q = queries, K = keys, V = values$
- *Self-attention* $\longrightarrow Q = K = V$

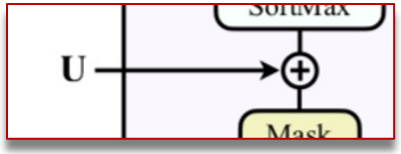
$$\boldsymbol{Q} = \boldsymbol{Q} \times \boldsymbol{W_Q}$$
$$\boldsymbol{K} = \boldsymbol{K} \times \boldsymbol{W_K}$$
$$\boldsymbol{V} = \boldsymbol{V} \times \boldsymbol{W_V}$$

Attention is All You Need!

[arXiv:1706.03762](https://arxiv.org/abs/1706.03762)

U —> Attention matrix —> correlation of "data sequence with data sequence"

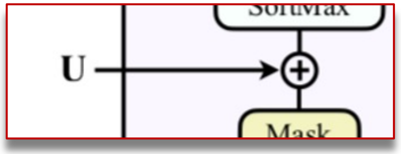
***U* ->** **Pairwise Features + SM interaction matrix** (attention matrix)



To show the interaction strength based on the SM coupling constants

Adding Pairwise features

Include pairwise features in **Par**ticle **T**ransformer through a trainable embedding $\boldsymbol{U_{ij}}$ for particles i and j



Pairwise Features +
SM interaction matrix
(attention matrix)

Attention Modules

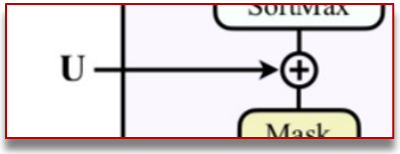
$$Attention(Q,K,T) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + \boldsymbol{U}\right)V$$

Features from the paper [arXiv:2202.03772]

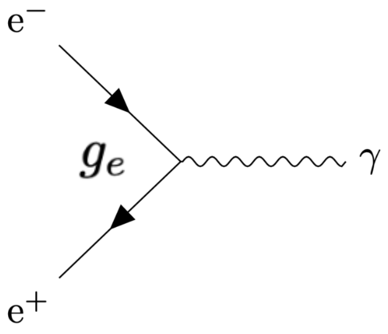
- **ParT** uses high level features for better performance
  1. $\Delta = \sqrt{(y_a - y_b)^2 + (\phi_a + \phi_b)^2}$
  2. $k_t = min(p_{T,a}, p_{T,b})\Delta$
  3. $z = min(p_{T,a}, p_{T,b})/(p_{T,a}, p_{T,b})$
  4. $m^2 = (E_a + E_b)^2 - \|p_a + p_b\|^2$
- These were also tested in **LightGBM**

What we end up using :
$\boldsymbol{m_{ij}}, \Delta\, \boldsymbol{R_{ij}}$ and dynamically calculated **coupling constants** of interaction terms (i.e. a feature that is coupling constant when i and j are components of a **SM** current, and 0 otherwise)

Interaction Matrices



Pairwise Features +
SM interaction matrix
(attention matrix)



Matrix [1] – SM ids

```
#   -   j   jb  e-  e+  m-  m+  g
([[0,  0,  0,  0,  0,  0,  0,  0],   # -
  [0,  1,  1,  0,  0,  0,  0,  1],   # j
  [0,  1,  1,  0,  0,  0,  0,  1],   # jb
  [0,  0,  0,  0,  1,  0,  0,  1],   # e-
  [0,  0,  0,  1,  0,  0,  0,  1],   # e+
  [0,  0,  0,  0,  0,  0,  1,  1],   # m-
  [0,  0,  0,  0,  0,  1,  0,  1],   # m+
  [0,  1,  1,  1,  1,  1,  1,  0]])  # g
```

Matrix [2] – SM const

```
#  -     j       bjet    e-    e+    m-    m+    g(photon)
([[0,    0,      0,      0,    0,    0,    0,    0     ],  # -
  [0,    g_s,    g_s,    0,    0,    0,    0,    g_e/2],   # j
  [0,    g_s,    g_s,    0,    0,    0,    0,    g_e/3],   # bjet
  [0,    0,      0,      0,    g_z,  0,    0,    g_e   ],  # e-
  [0,    0,      0,      g_z,  0,    0,    0,    g_e   ],  # e+
  [0,    0,      0,      0,    0,    0,    g_z,  g_e   ],  # m-
  [0,    0,      0,      0,    0,    g_z,  0,    g_e   ],  # m+
  [0,    g_e/2,  g_e/3,  g_e,  g_e,  g_e,  g_e,  0]    ])  # g
```

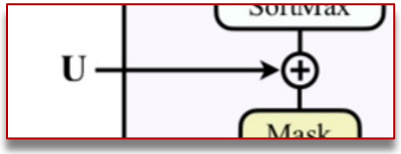
- '1' indicates an interaction possible at **LO** in the **SM**
- '0' indicates interactions that only appear at higher orders

- $g_Z = 0.758$ for the weak force for leptons
- $g_s = 1.22$ for the strong force in jet interactions
- $g_e = 0.31$ for the electromagnetic force in photon interactions

The energy dependence of the coupling constants

Matrix [3] – SM

```
#  -      j       bjet     e-     e+     m-     m+     g(photon)
([[0,     0,      0,       0,     0,     0,     0,     0      ],  # -
  [0,     g_s,    g_s,     0,     0,     0,     0,     g_e/2],  # j
  [0,     g_s,    g_s,     0,     0,     0,     0,     g_e/3],  # bjet
  [0,     0,      0,       0,     g_z,   0,     0,     g_e  ],  # e-
  [0,     0,      0,       g_z,   0,     0,     0,     g_e  ],  # e+
  [0,     0,      0,       0,     0,     0,     g_z,   g_e  ],  # m-
  [0,     0,      0,       0,     0,     g_z,   0,     g_e  ],  # m+
  [0,     g_e/2,  g_e/3,   g_e,   g_e,   g_e,   g_e,   0]   ])  # g
```



Pairwise Features + SM interaction matrix (attention matrix)

Dynamically calculated **coupling constants** of interaction terms !

α is the running coupling constant

$$(*)\quad \alpha(Q^2) = \frac{\alpha(\mu_0^2)}{1 - \frac{n\alpha(\mu_0^2)}{3\pi} \cdot \ln\left(\frac{Q^2}{\mu_0^2}\right)},$$

$$g_e = \sqrt{4\pi\alpha}$$

$$\alpha_s(Q^2) = \frac{\alpha_s(\mu_0^2)}{1 + \frac{\alpha_s(\mu_0^2)(33-2n_f)}{12\pi} \ln\left(\frac{Q^2}{\mu_0^2}\right)},$$

$$g_s = \sqrt{4\pi\alpha_s}$$

n_f – number of quark flavors that are active

Where $\mu_0 = 91.1876$ GeV, $\alpha(\mu_0) = \frac{1}{127.5}$, $\alpha_s(\mu_0) = 0.118$, $n_f = 6$

$$Q^2 = \bar{p}_t^2 = \left(\frac{p_t^i + p_t^j}{2}\right)^2$$

energy scale

(*) Considered only leptons

$\boldsymbol{n = 3}$ – approximates the contribution of the different particles in the loop

$$g_Z = 0.758$$

Results for the $t\bar{t}t\bar{t}$ signal

arXiv:2211.05143

		BDT	$\text{BDT}_{\text{int.}}$	FCN	CNN
$t\bar{t}+h$	AUC	0.825(0)	0.831(0)	0.821(2)	0.778(6)
	$\epsilon_B(\epsilon_S=0.7)$	0.206(0)	0.192(0)	0.203(1)	0.272(11)
	$\epsilon_B(\epsilon_S=0.3)$	0.026(1)	0.026(0)	0.026(1)	0.037(1)
$t\bar{t}+W$	AUC	0.891(0)	0.895(0)	0.887(0)	0.867(5)
	$\epsilon_B(\epsilon_S=0.7)$	0.099(0)	0.092(0)	0.103(1)	0.125(8)
	$\epsilon_B(\epsilon_S=0.3)$	0.011(0)	0.011(0)	0.010(0)	0.011(1)
$t\bar{t}+WW$	AUC	0.740(0)	0.746(0)	0.737(1)	0.745(2)
	$\epsilon_B(\epsilon_S=0.7)$	0.347(0)	0.339(0)	0.342(5)	0.335(3)
	$\epsilon_B(\epsilon_S=0.3)$	0.050(0)	0.051(0)	0.054(0)	0.051(0)
$t\bar{t}+Z$	AUC	0.833(0)	0.856(0)	0.836(0)	0.839(1)
	$\epsilon_B(\epsilon_S=0.7)$	0.191(0)	0.163(0)	0.192(0)	0.190(4)
	$\epsilon_B(\epsilon_S=0.3)$	0.026(0)	0.019(0)	0.023(0)	0.021(1)
		PN	$\text{PN}_{\text{int.}}$	$\text{PN}_{\text{int. SM}}$	$\text{ParT}_{\text{int. SM (FL)}}$
$t\bar{t}+h$	AUC	0.824(0)	0.842(1)	**0.846(1)**	0.844(1)
	$\epsilon_B(\epsilon_S=0.7)$	0.199(0)	0.176(3)	**0.171(2)**	0.176(2)
	$\epsilon_B(\epsilon_S=0.3)$	0.025(0)	**0.019(1)**	0.020(1)	0.020(1)
$t\bar{t}+W$	AUC	0.887(0)	0.895(2)	0.900(1)	**0.902(4)**
	$\epsilon_B(\epsilon_S=0.7)$	0.102(1)	0.097(1)	**0.091(1)**	**0.091(5)**
	$\epsilon_B(\epsilon_S=0.3)$	0.011(0)	0.011(0)	**0.010(0)**	0.011(0)
$t\bar{t}+WW$	AUC	0.742(0)	0.760(1)	0.765(0)	0.768(3)
	$\epsilon_B(\epsilon_S=0.7)$	0.335(2)	0.311(1)	0.297(2)	0.294(7)
	$\epsilon_B(\epsilon_S=0.3)$	0.051(0)	0.044(1)	**0.044(1)**	0.044(1)
$t\bar{t}+Z$	AUC	0.851(0)	0.879(1)	**0.887(1)**	0.892(0)
	$\epsilon_B(\epsilon_S=0.7)$	0.168(4)	0.136(1)	**0.126(2)**	0.119(4)
	$\epsilon_B(\epsilon_S=0.3)$	0.020(0)	0.016(1)	0.016(0)	0.016(0)
		ParT	$\text{ParT}_{\text{int.}}$	$\text{ParT}_{\text{int. SM}}$	$\text{SetT}_{\text{int. SM}}$
$t\bar{t}+h$	AUC	0.824(0)	0.837(2)	**0.846(1)**	0.845(1)
	$\epsilon_B(\epsilon_S=0.7)$	0.197(3)	0.179(6)	0.174(1)	0.176(3)
	$\epsilon_B(\epsilon_S=0.3)$	0.023(0)	0.020(0)	0.020(0)	0.020(0)
$t\bar{t}+W$	AUC	0.896(1)	0.899(1)	**0.905(2)**	0.898(1)
	$\epsilon_B(\epsilon_S=0.7)$	0.097(2)	0.090(1)	**0.089(3)**	0.094(2)
	$\epsilon_B(\epsilon_S=0.3)$	0.010(0)	0.010(0)	**0.009(0)**	0.011(0)
$t\bar{t}+WW$	AUC	0.737(0)	0.767(1)	**0.769(0)**	0.763(1)
	$\epsilon_B(\epsilon_S=0.7)$	0.354(3)	0.295(5)	**0.288(2)**	0.301(5)
	$\epsilon_B(\epsilon_S=0.3)$	0.050(1)	0.040(0)	**0.042(0)**	0.047(1)
$t\bar{t}+Z$	AUC	0.839(1)	0.885(0)	**0.891(1)**	0.886(2)
	$\epsilon_B(\epsilon_S=0.7)$	0.182(2)	0.130(1)	**0.119(3)**	0.129(4)
	$\epsilon_B(\epsilon_S=0.3)$	0.021(1)	0.016(0)	**0.015(0)**	**0.014(0)**

The areas under the ROC curve and the **background efficiencies**, at **signal efficiencies of** 70% and 30% respectively

- Quoted uncertainties are extracted from three independent runs for each network architecture
- Numbers in bold indicate the best performance

Let's zoom in →

Results for the $t\bar{t}t\bar{t}$ and $t\bar{t}H$ signals

arXiv:2211.05143

The AUC for both 4 top and top-top-Higgs signal detection

		PN	$\mathrm{PN}_{\mathrm{int.}}$	$\mathrm{PN}_{\mathrm{int.\,SMids}}$	$\mathrm{PN}_{\mathrm{int.\,SM\,const}}$	$\mathrm{PN}_{\mathrm{int.\,SM}}$
$t\bar{t}t\bar{t}$	AUC	0.8471(1)	0.8729(0)	0.8725(0)	0.8727(0)	**0.8739(0)**
	$\epsilon_B(\epsilon_S = 0.7)$	0.1758(3)	0.1387(1)	0.1377(0)	0.1384(0)	**0.1369(1)**
	$\epsilon_B(\epsilon_S = 0.3)$	0.0207(0)	0.0182(0)	0.0178(0)	0.0178(0)	**0.0176(0)**
		ParT	$\mathrm{ParT}_{\mathrm{int.}}$	$\mathrm{ParT}_{\mathrm{int.\,SMids}}$	$\mathrm{ParT}_{\mathrm{int.\,SM\,const}}$	$\mathrm{ParT}_{\mathrm{int.\,SM}}$
$t\bar{t}t\bar{t}$	AUC	0.8404(0)	0.8708(0)	0.8715(0)	0.8717(0)	**0.8732(0)**
	$\epsilon_B(\epsilon_S = 0.7)$	0.1842(3)	0.1394(0)	0.1389(2)	0.1372(1)	**0.1366(0)**
	$\epsilon_B(\epsilon_S = 0.3)$	0.0230(0)	0.0172(0)	0.0180(0)	**0.0167(0)**	0.0169(0)

The models containing both the **pairwise features** and the **SM interaction matrix** performs **best**

		PN	$\mathrm{PN}_{\mathrm{int.}}$	$\mathrm{PN}_{\mathrm{int.\,SMids}}$	$\mathrm{PN}_{\mathrm{int.\,SM\,const}}$	$\mathrm{PN}_{\mathrm{int.\,SM}}$
$t\bar{t}+h$	AUC	0.8146(2)	0.8505(0)	0.8489(1)	0.8505(0)	**0.8523(0)**
	$\epsilon_B(\epsilon_S = 0.7)$	0.2292(1)	0.1787(0)	0.1785(1)	0.1764(3)	**0.1733(1)**
	$\epsilon_B(\epsilon_S = 0.3)$	0.0471(1)	0.0345(0)	0.0343(1)	0.0350(0)	**0.0340(0)**
		ParT	$\mathrm{ParT}_{\mathrm{int.}}$	$\mathrm{ParT}_{\mathrm{int.\,SMids}}$	$\mathrm{ParT}_{\mathrm{int.\,SM\,const}}$	$\mathrm{ParT}_{\mathrm{int.\,SM}}$
$t\bar{t}+h$	AUC	0.8058(1)	0.8507(0)	0.8473(0)	0.8497(0)	**0.8532(0)**
	$\epsilon_B(\epsilon_S = 0.7)$	0.2399(2)	0.1794(1)	0.1836(3)	0.1801(1)	**0.1748(1)**
	$\epsilon_B(\epsilon_S = 0.3)$	0.0502(0)	0.0357(0)	0.0355(1)	0.0367(0)	**0.0351(0)**

The **background** can be significantly **reduced** by about **30%** compared to a **PN (GNN)**

Significance

arXiv:2211.05143

Highlights the **enhanced performance** of **ParT int. SM** models over baseline **PN (GNN)** (neglecting sys err) for 4top signal

$$\sigma = \frac{s}{\sqrt{b}} \qquad \sigma_{\delta_{sys}=0.2} = \frac{s}{\sqrt{b_{sys}}}$$

$$\mathrm{b}_{sys} = b + (b \cdot \delta_{sys})^2$$

- At $\epsilon_S = 0.7$: significance boost from **2.21** to **2.98** $\boldsymbol{\sigma}$ with **ParT int. SM** => **PN** requires **82%** more luminosity !
- At $\epsilon_S = 0.3$: significance boost from **8.29** to **9.88** $\boldsymbol{\sigma}$ with **ParT int. SM** => **PN** needs **42%** more luminosity !
- At $\epsilon_S = 0.3$: significance boost from **8.29** to **10.48** $\boldsymbol{\sigma}$ with **ParT int. SM (FL)** => **PN** needs **60%** more luminosity !

Significance table (calculations assume L = 100 fb^{-1})

		σ	$\sigma_{\delta sys} = 0.2$
BDT	$\epsilon_S = 0.3$	20.77	6.79
	$\epsilon_S = 0.7$	16.82	2.01
$\mathrm{BDT_{int.}}$	$\epsilon_S = 0.3$	21.93	7.53
	$\epsilon_S = 0.7$	17.51	2.17
FCN	$\epsilon_S = 0.3$	20.31	6.51
	$\epsilon_S = 0.7$	16.67	1.97
CNN	$\epsilon_S = 0.3$	20.88	6.86
	$\epsilon_S = 0.7$	16.73	1.98
PN	$\epsilon_S = 0.3$	23.09	8.29
	$\epsilon_S = 0.7$	17.68	2.21
$\mathrm{PN_{int.}}$	$\epsilon_S = 0.3$	25.30	9.83
	$\epsilon_S = 0.7$	**20.51**	**2.97**
$\mathrm{PN_{int.\,SM}}$	$\epsilon_S = 0.3$	**25.65**	**10.09**
	$\epsilon_S = 0.7$	**20.50**	**2.97**
ParT	$\epsilon_S = 0.3$	22.37	7.82
	$\epsilon_S = 0.7$	17.72	2.23
$\mathrm{ParT_{int.}}$	$\epsilon_S = 0.3$	24.54	9.29
	$\epsilon_S = 0.7$	20.21	2.89
$\mathrm{ParT_{int.\,SM}}$	$\epsilon_S = 0.3$	25.36	9.88
	$\epsilon_S = 0.7$	**20.53**	**2.98**
$\mathrm{ParT_{int.\,SM\,(FL)}}$	$\epsilon_S = 0.3$	**26.19**	**10.48**
	$\epsilon_S = 0.7$	20.28	2.91
$\mathrm{SetT_{int.\,SM}}$	$\epsilon_S = 0.3$	**25.58**	**10.03**
	$\epsilon_S = 0.7$	20.18	2.88

Results

arXiv:2211.05143

We asked the question: → "**Do the models saturate ?**"

PN and ParT Models (with the pairwise features + the SM coupling constants)

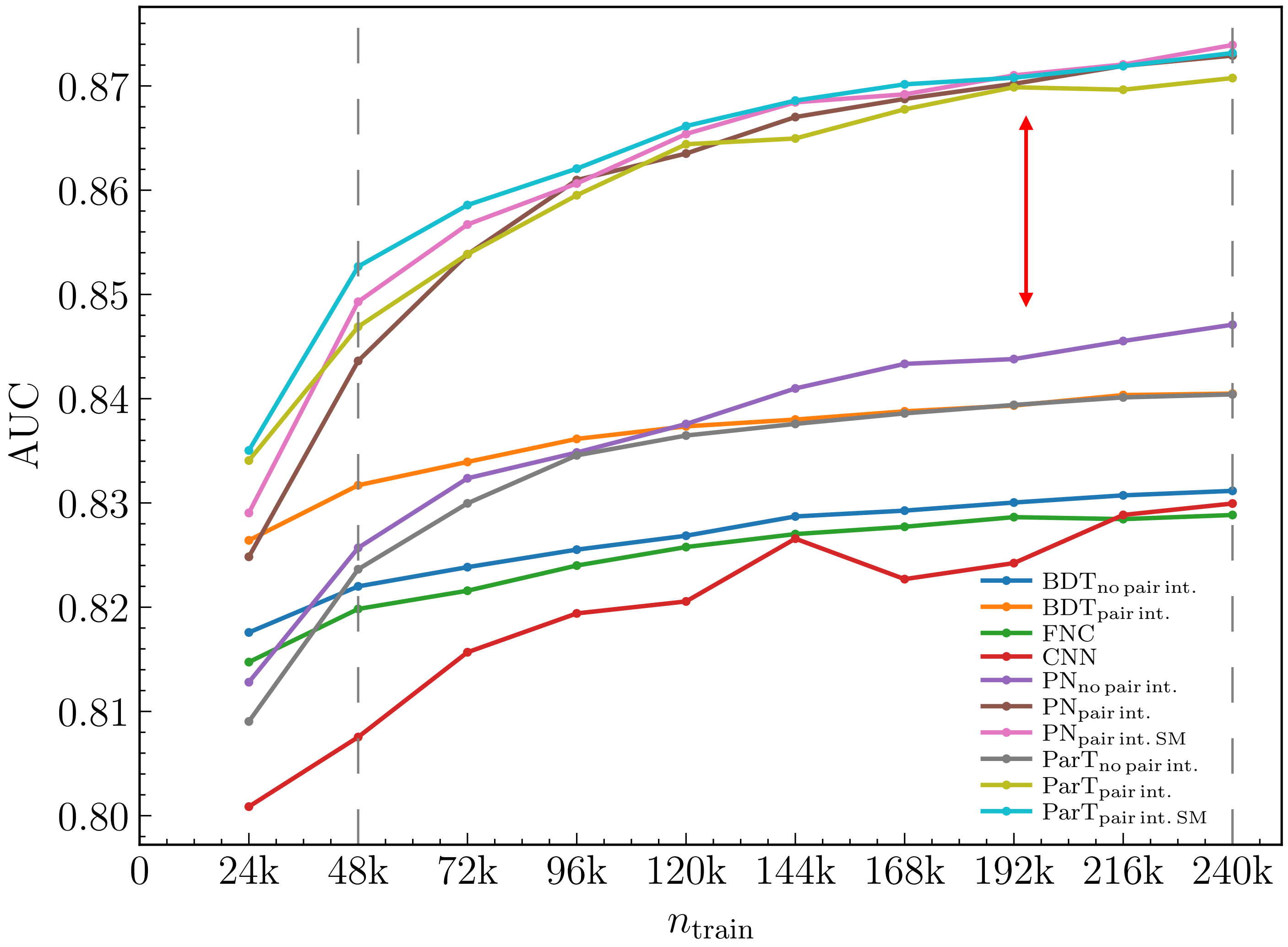
- Shows a steeper increase in **AUC** with fewer data
- Indicate higher data efficiency → less data needed for strong performance

Other Models

- **AUC** scores improve more gradually
- Suggest a requirement for larger datasets to match **PN** and **ParT** performance

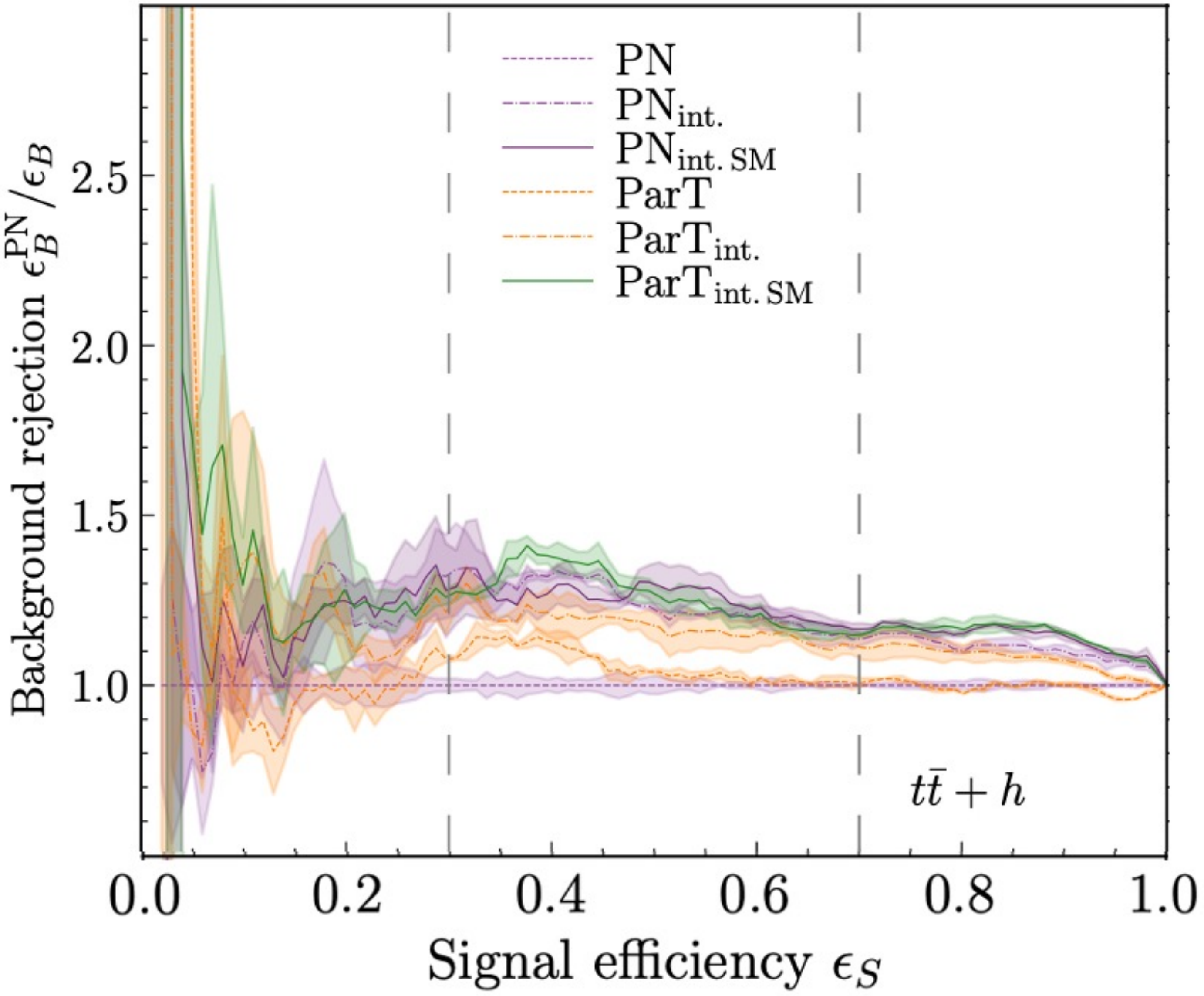
PN and ParT models could be preferable in data-scarce cases

The AUC scores as a function of training size



Results

arXiv:2211.05143



A plot with signal efficiency VS background rejection

*compared to the **ParticleNet (GNN)**

Models with integrated **pairwise features + SM interactions** exhibit up to a 40% higher background rejection

Demonstrates the strength of **SM interaction matrix** as a powerful inductive bias in learning

X-axis – the signal efficiency
Y-axis – the background rejection

Summary

⭐ **Integration of energy-dependent SM interactions into ML models**

- Embedding pairwise features and energy-dependent **SM** interactions into **ML** architectures significantly boosts event classification accuracy and efficiency:

  - Enhanced background suppression by **10-40%** compared to baseline **PN** (**GNN**) models

  - Approximately **10%** of this improvement is due to the **SM interaction matrix**

  - ML models show up to **30%** increase in significance vs. baseline

  - Achieving similar significance via increased luminosity would require **~70%** more data (compare to the baseline model)

Transformers, when supplemented with **pairwise features** and **SM couplings**, show potential as a **powerful tool** for collider physics challenges!

CAN WE TURN CLASSIFIERS INTO ANOMALY DETECTORS ?

⭐ It has been observed that, ***on average***,
the top-performing classifiers serve as best anomaly detectors
(**with the SM interactions**)

For further details, refer to Adrian's [talk]

Thank you for your attention!

Back up

Math Behind the Attention Mechanism

Attention Modules
(scaled dot product attention):

- $Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + \boldsymbol{U}\right)V$
- $Q = queries, K = keys, V = values$
- *Self-attention* $\longrightarrow Q = K = V$

How Particles Inform Each Other ?

- **Calculating Interaction Scores:**
  - **Attention Score** $(\boldsymbol{Q}, \boldsymbol{K}) = \frac{\boldsymbol{QK^T}}{\sqrt{\boldsymbol{d}}}$

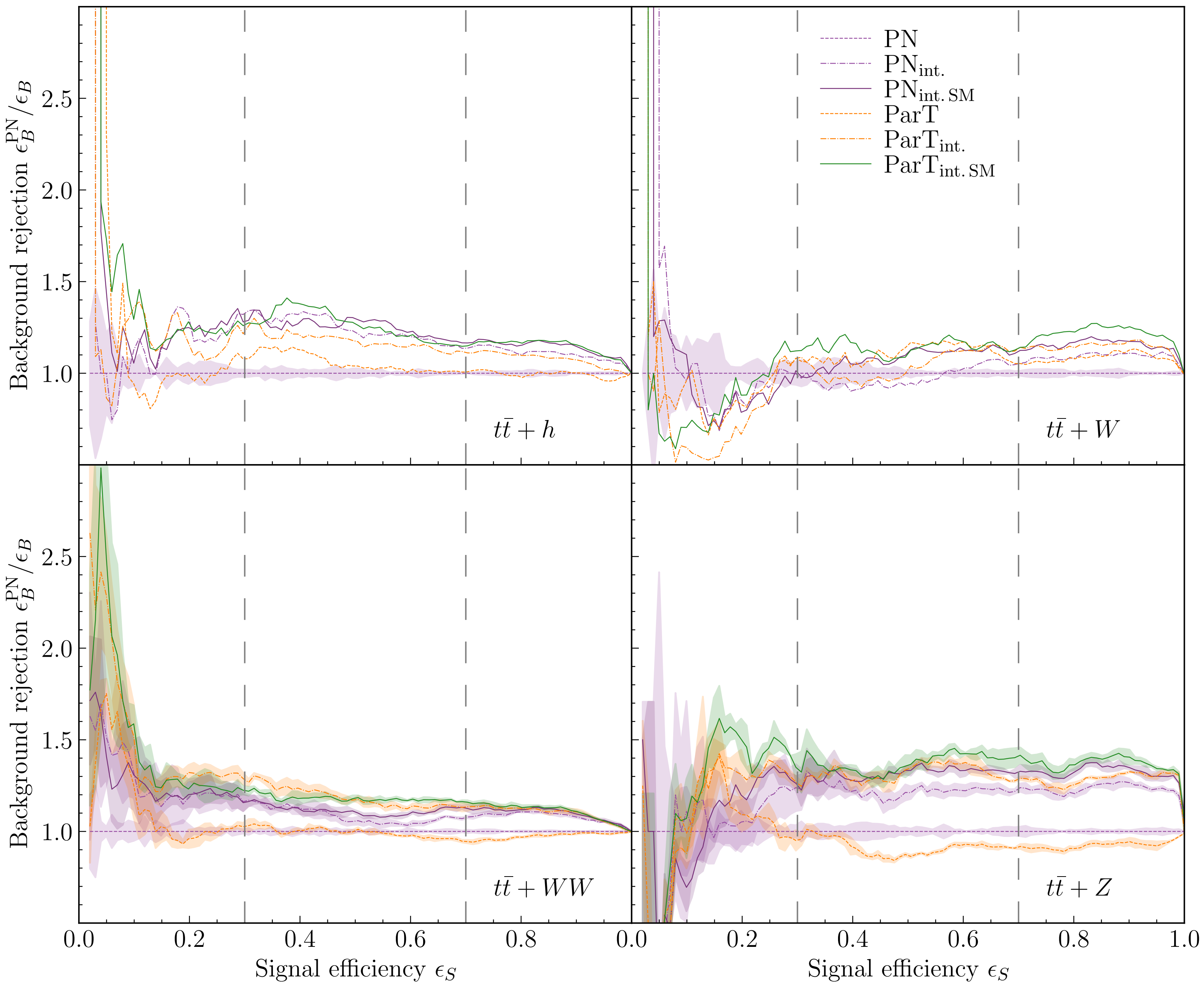
    where $\sqrt{d}$ is the dimentionality of the key vector, used to scale the dot product
- **Normalizing Scores to Probabilities:**
  - **Attention Weighs** $=$ $\boldsymbol{SoftMax}$***(Attention Score)***

    normalizes the scores to ensure they sum up to 1, acting as probabilities
- **Particle Representation:**
  - **Output** $=$ ***Attention Wights*** $*\ \boldsymbol{V}$

    each particle's output is a combination of all particles' information, weighted by their computed relevance

Result: captures the dynamic interactions between particles

A plot with signal efficiency VS background rejection



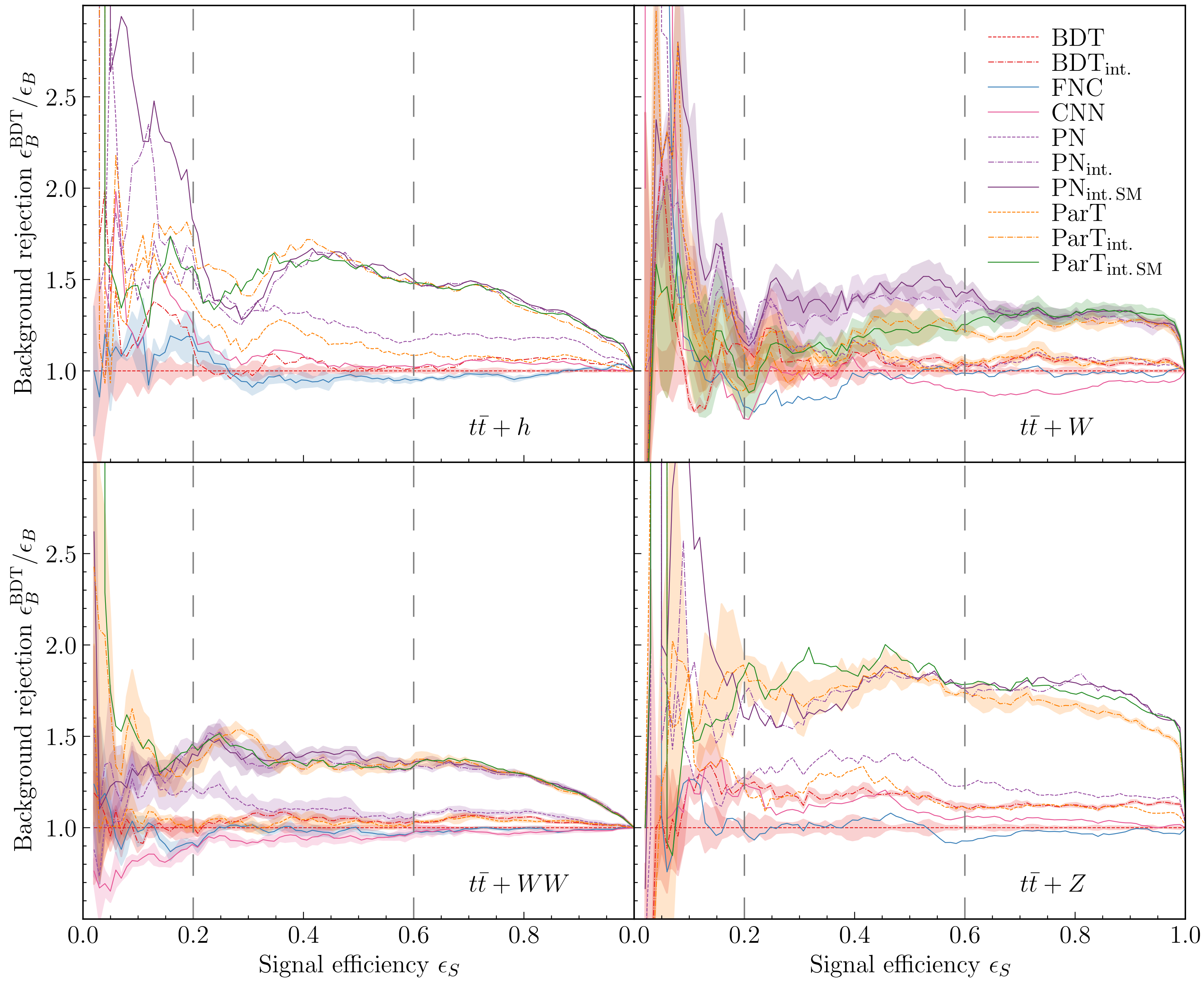
arXiv:2211.05143

compared to the **ParticleNet (GNN)**

We can achieve a **10-40% higher background rejection** for signal efficiencies between 30-90% by switching from **GNN** to **models with the pairwise features + the SM coupling constants**

X-axis – the signal efficiency
Y-axis – the background rejection

A plot with signal efficiency VS background rejection



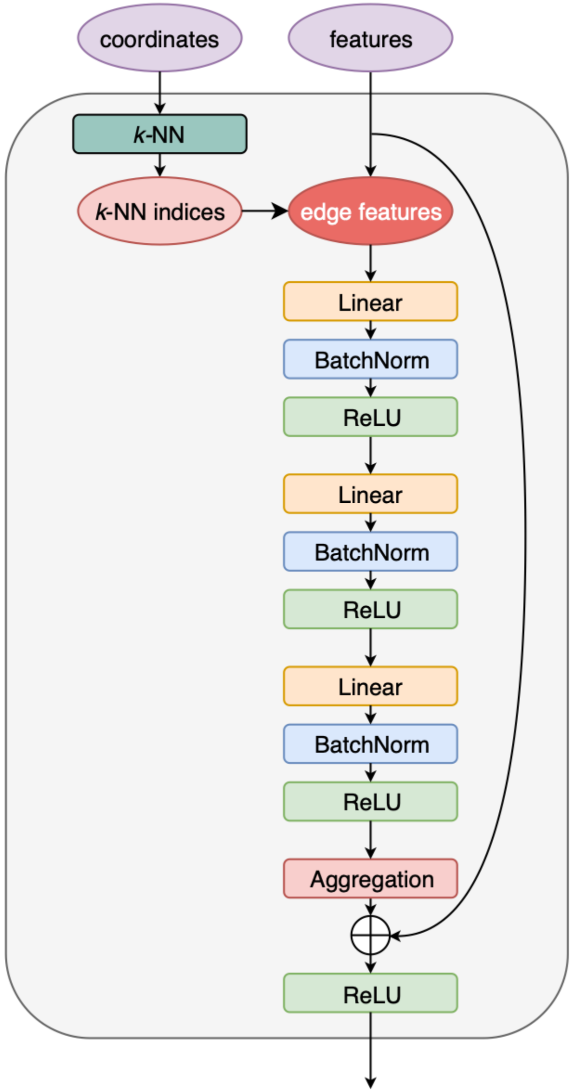
arXiv:2211.05143

Compared to the **BDT** for full size of the dataset

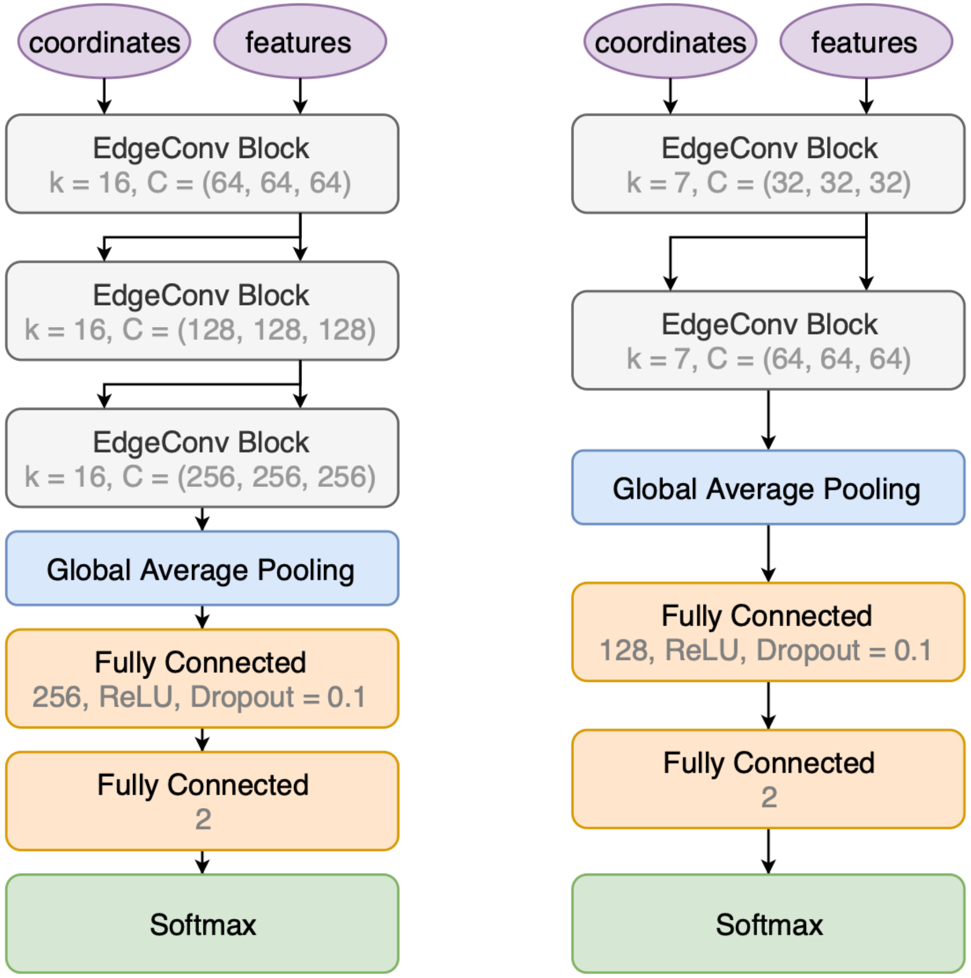
X-axis – the signal efficiency
Y-axis – the background rejection

EdgeConv and ParticleNet

arXiv:1902.08570



The structure of the EdgeConv block



(a) ParticleNet (b) ParticleNet-Lite

The architectures of the ParticleNet and the ParticleNet-Lite networks

1D CNN

arXiv:2011.13466

- Input is a **Particle List**
- **LRP** is a backpropagation method