What Do Jet Taggers Learn :
A Causal and Explainable AI Dissection

Sanmay Ganguly
Indian Institute of Technology - Kanpur
https://sanmayphy.github.io/

16/09/2026

Anusandhan National Research Foundation

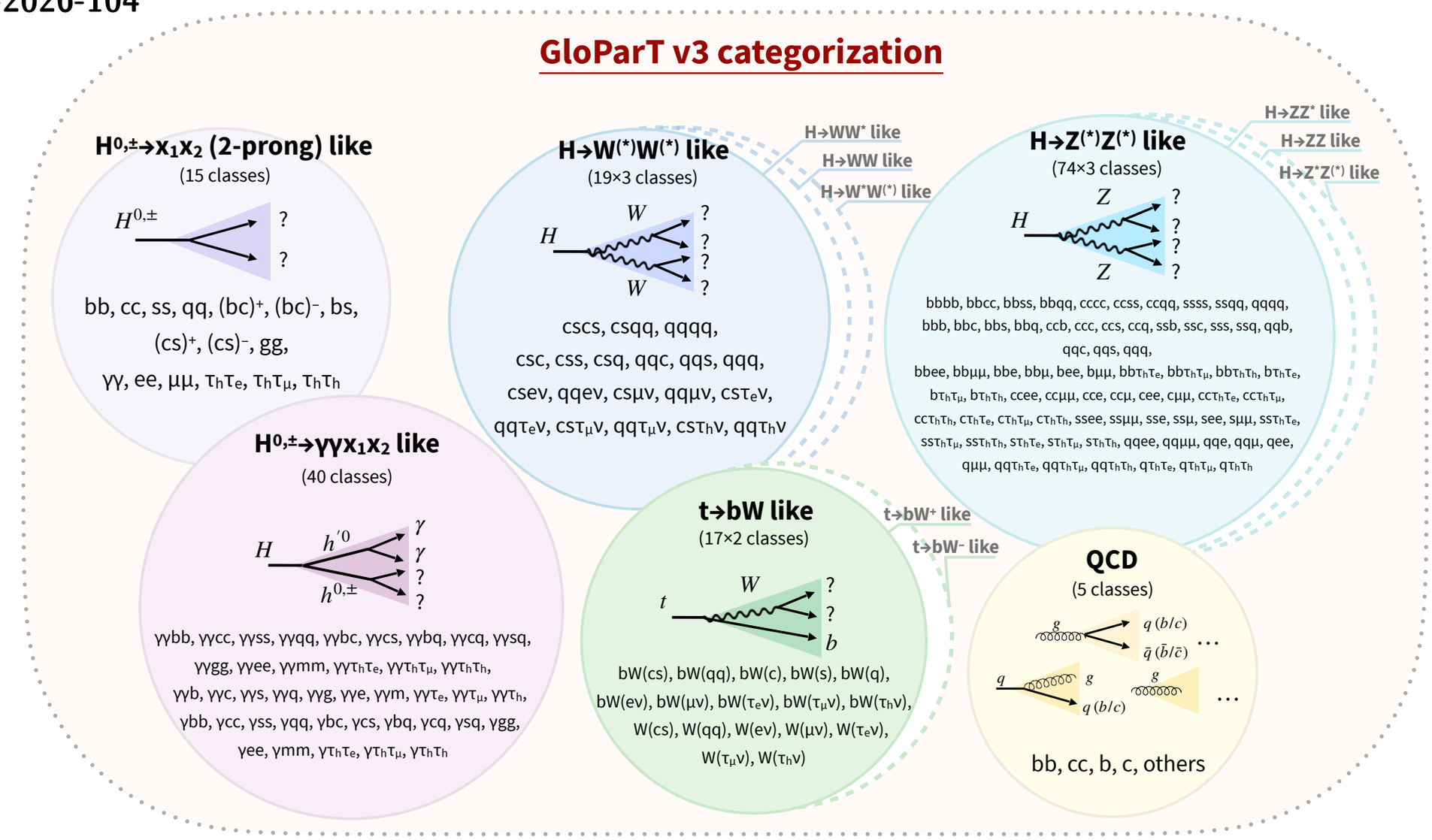
ML4Jets - 2026,Vienna

ArXiv : 2604.25885
Pahal D. Patel
Physics Undergrad @IIT-Kanpur Physics
==> Columbia University MS in AI

ArXiv : 2605.09881
Saurabh Rai
Physics PhD @IIT-Kanpur Physics

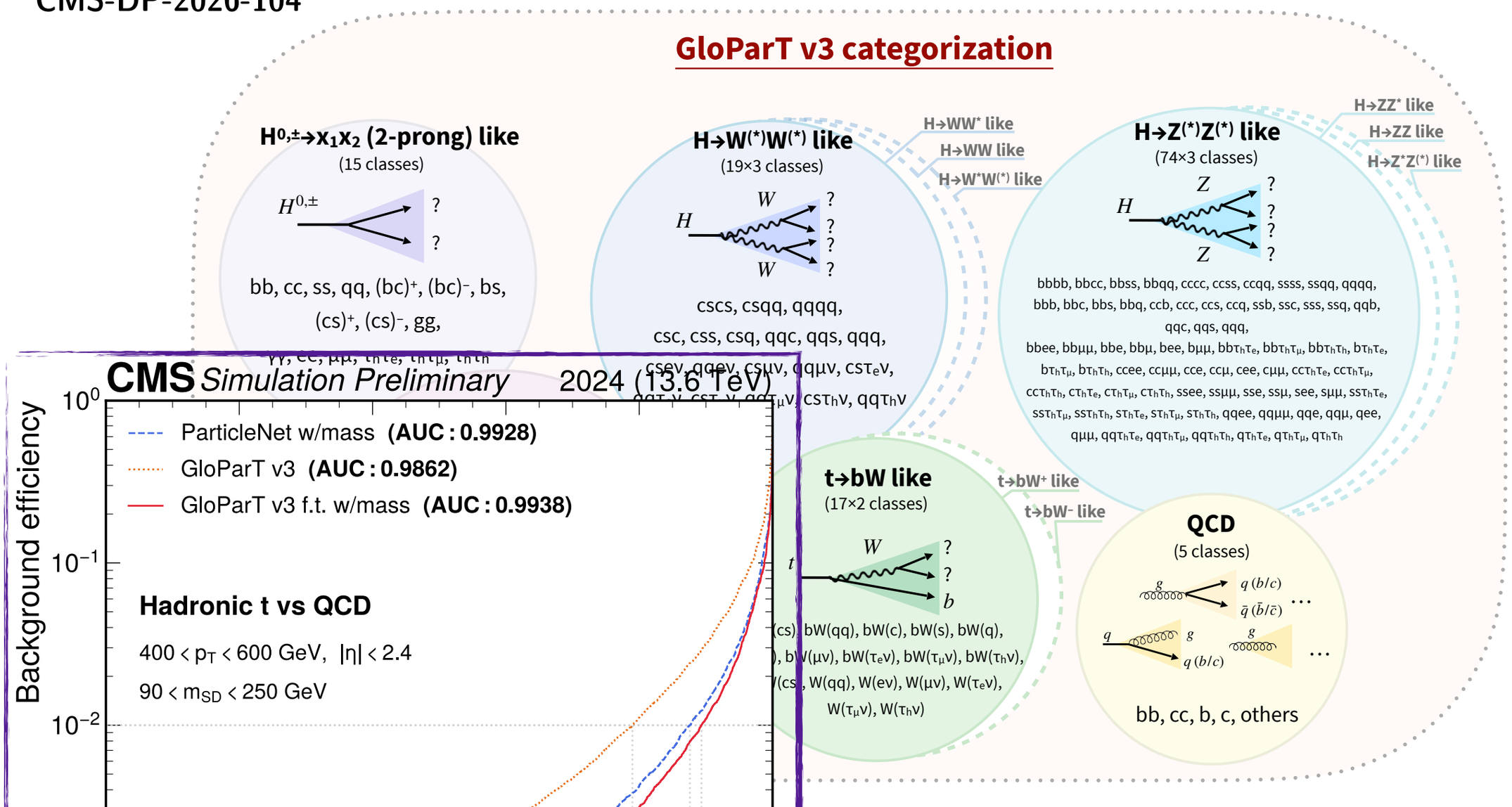
ML Based Decision Making : The Default

CMS-DP-2026-104



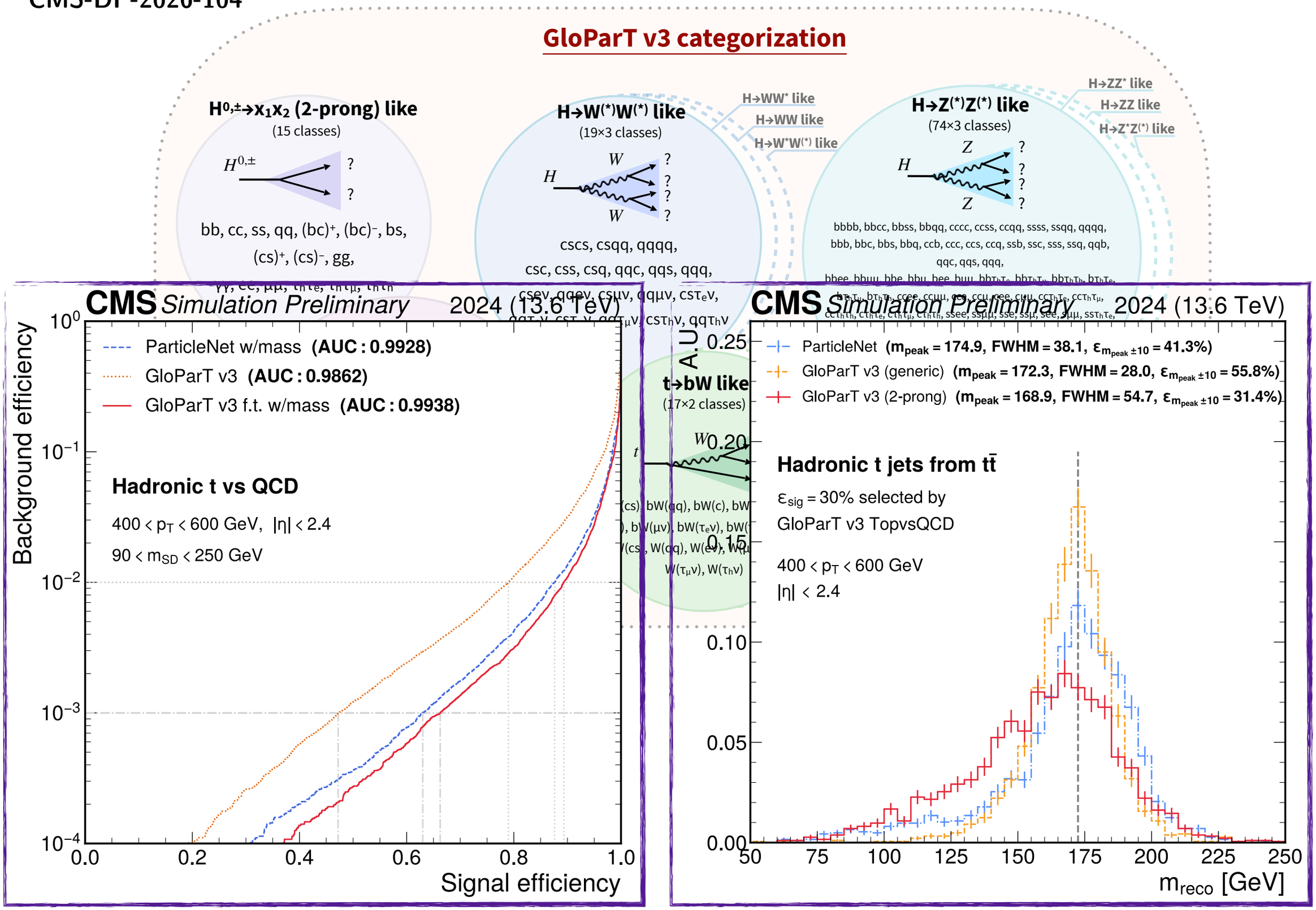
ML Based Decision Making : The Default

CMS-DP-2026-104

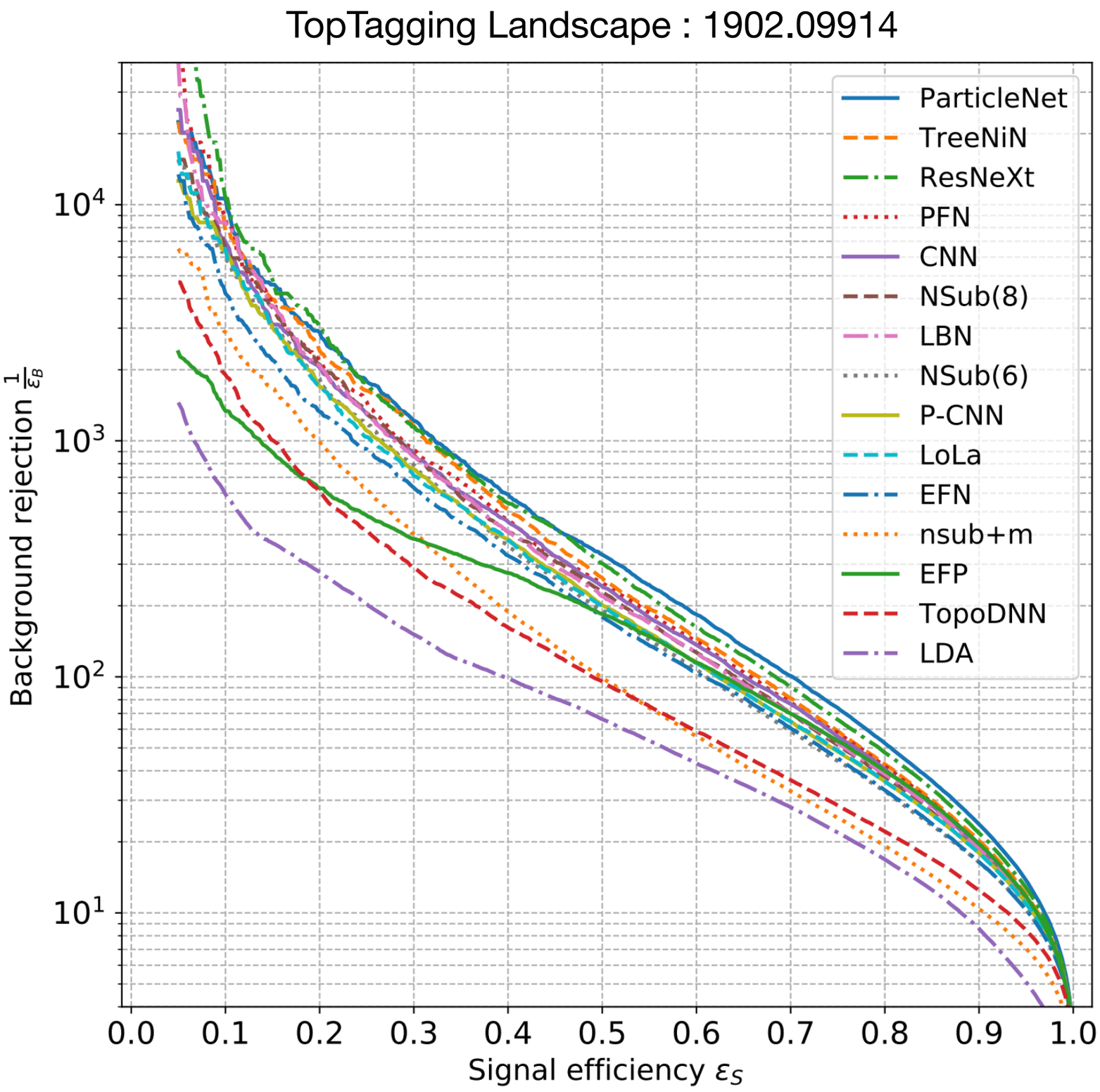


ML Based Decision Making : The Default

CMS-DP-2026-104

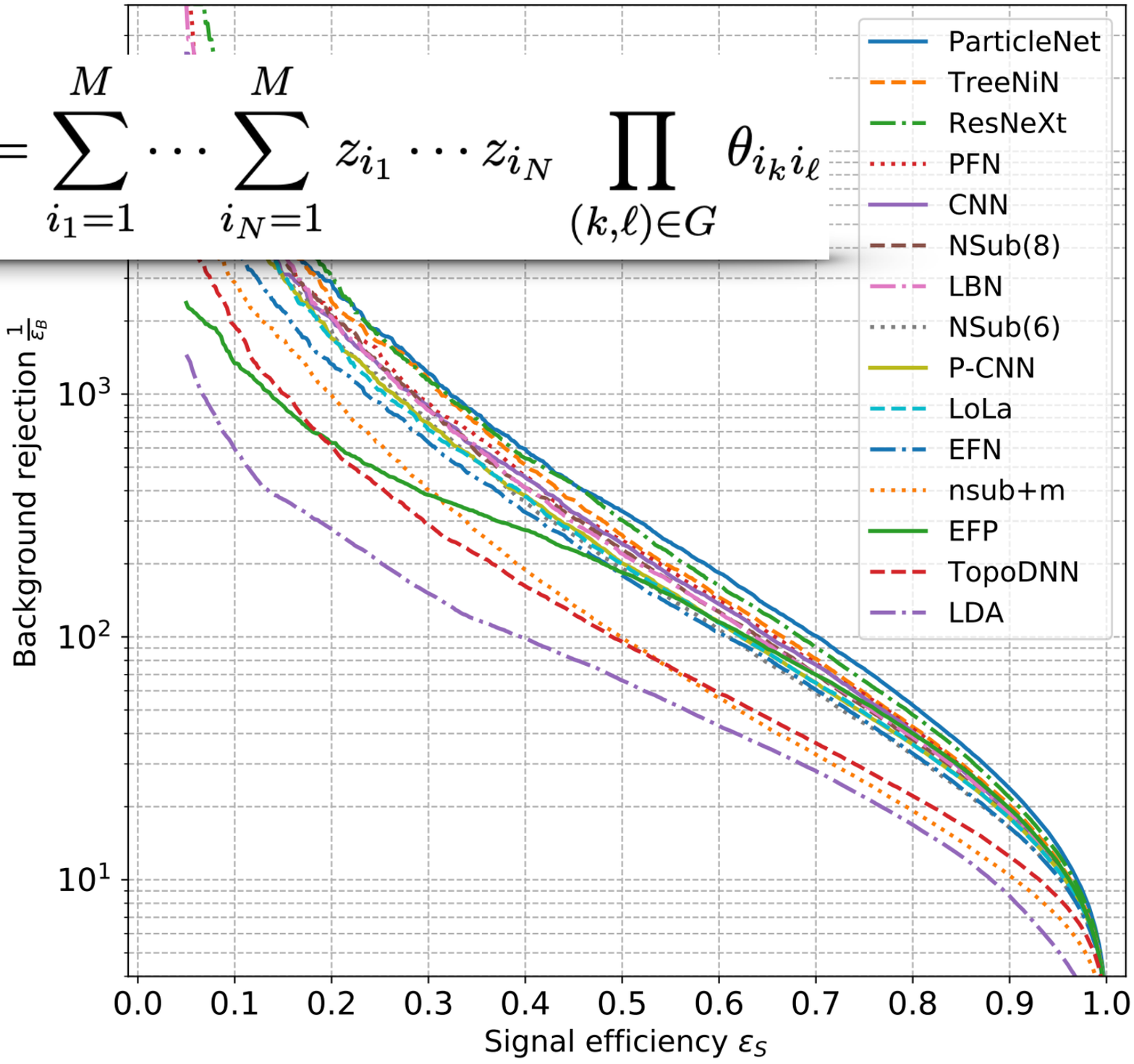


Surpassing Analytic Observables



Surpassing Analytic Observables

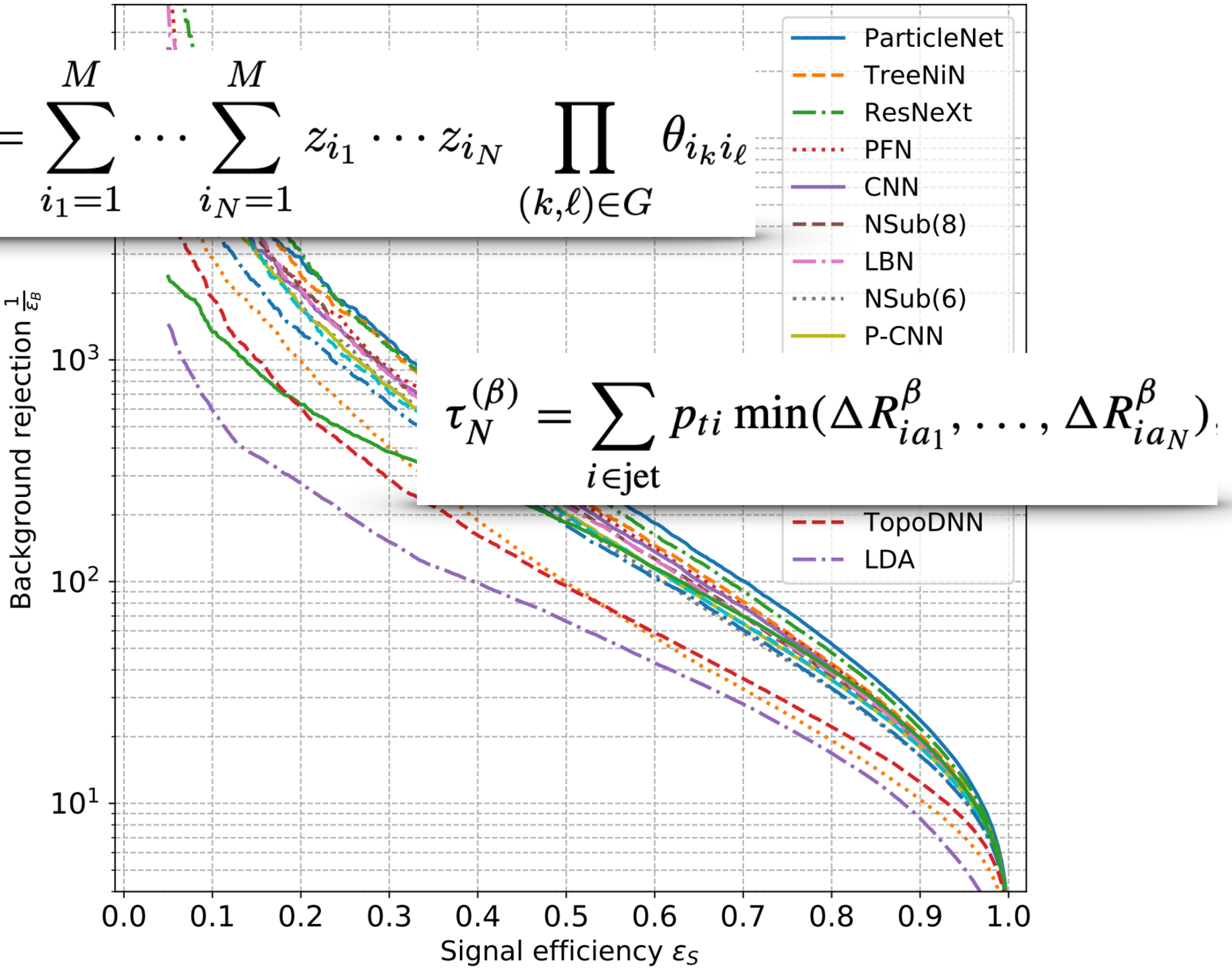
$$\mathrm{EFP}_G = \sum_{i_1=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} \cdots z_{i_N} \prod_{(k,\ell)\in G} \theta_{i_k i_\ell}$$



Surpassing Analytic Observables

$$\mathrm{EFP}_G = \sum_{i_1=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} \cdots z_{i_N} \prod_{(k,\ell)\in G} \theta_{i_k i_\ell}$$

$$\tau_N^{(\beta)} = \sum_{i \in \text{jet}} p_{ti} \min(\Delta R_{ia_1}^{\beta}, \ldots, \Delta R_{ia_N}^{\beta})$$



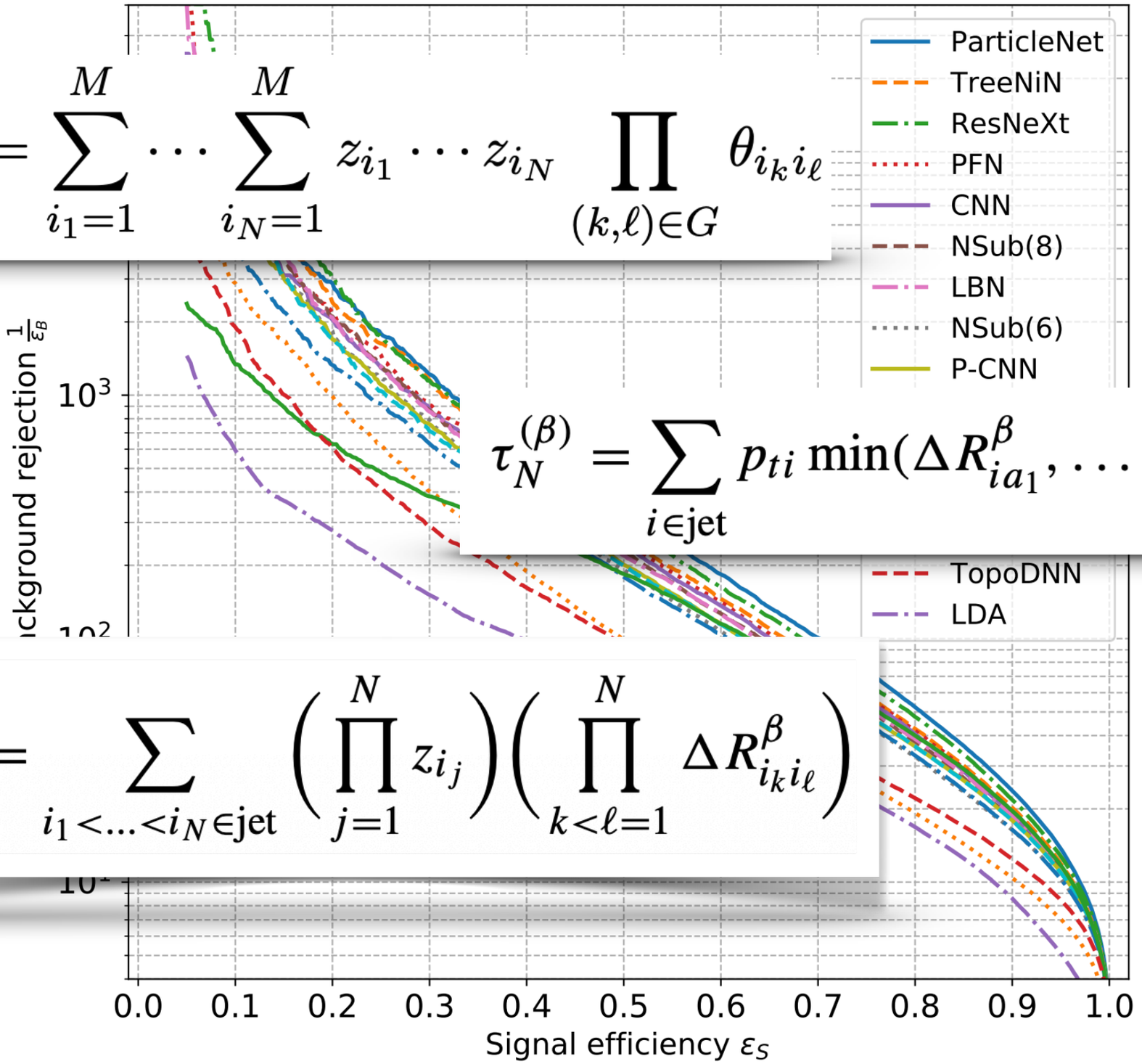
Surpassing Analytic Observables

TopTagging Landscape : 1902.09914

$$\mathrm{EFP}_G = \sum_{i_1=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} \cdots z_{i_N} \prod_{(k,\ell)\in G} \theta_{i_k i_\ell}$$

$$\tau_N^{(\beta)} = \sum_{i \in \text{jet}} p_{ti} \min(\Delta R_{ia_1}^{\beta}, \ldots, \Delta R_{ia_N}^{\beta})$$

$$e_N^{(\beta)} = \sum_{i_1<\ldots<i_N \in \text{jet}} \left(\prod_{j=1}^{N} z_{i_j} \right) \left(\prod_{k<\ell=1}^{N} \Delta R_{i_k i_\ell}^{\beta} \right)$$



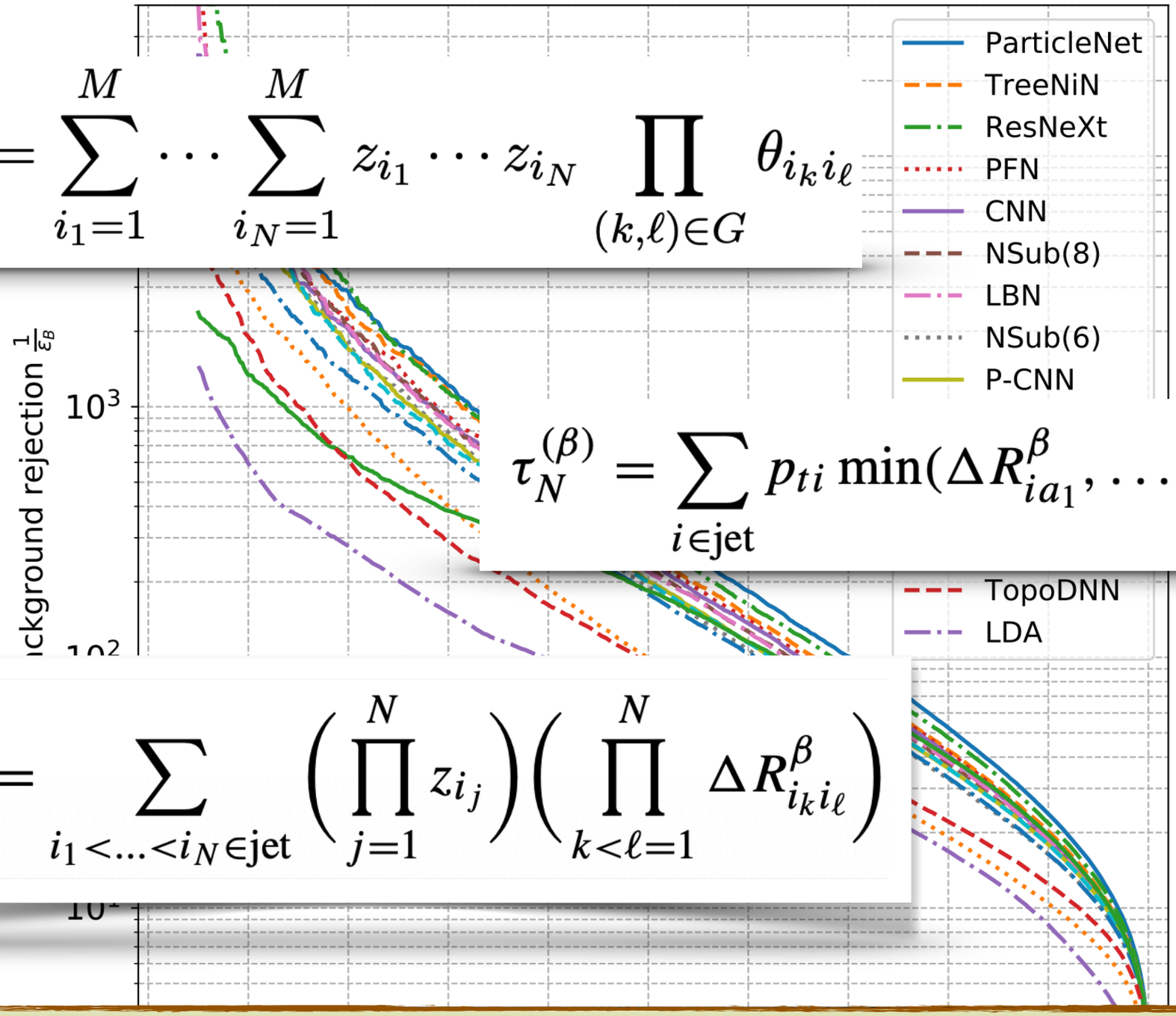
Surpassing Analytic Observables

TopTagging Landscape : 1902.09914

$$\mathrm{EFP}_G = \sum_{i_1=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} \cdots z_{i_N} \prod_{(k,\ell)\in G} \theta_{i_k i_\ell}$$

$$\tau_N^{(\beta)} = \sum_{i \in \text{jet}} p_{ti} \min(\Delta R_{ia_1}^{\beta}, \ldots, \Delta R_{ia_N}^{\beta})$$

$$e_N^{(\beta)} = \sum_{i_1<\ldots<i_N \in \text{jet}} \left(\prod_{j=1}^{N} z_{i_j} \right) \left(\prod_{k<\ell=1}^{N} \Delta R_{i_k i_\ell}^{\beta} \right)$$



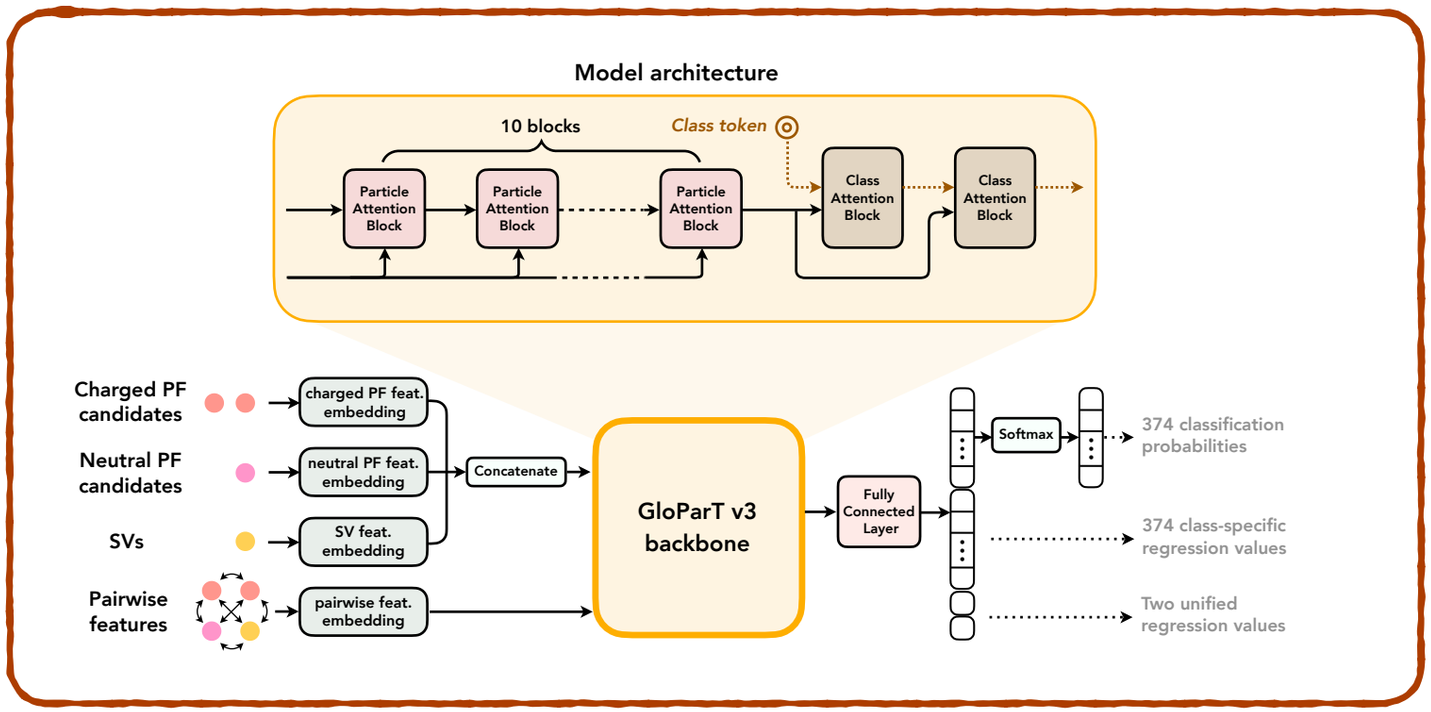
First principle understanding : what makes things work?

How Do We Read ML's Mind?

Are Are there strong correlation in between the hidden layers? Do each part of a network tries to learn different features?

Is it an artifact of the training fine-tuning ?

Do all of these aspects have mutual roles to play?



Are the higher order showering / hadronization effects being captured better by the raw inputs which are mis-modeled in fixed order MC ?

How do we map back to all of these to our physics driven reasoning?

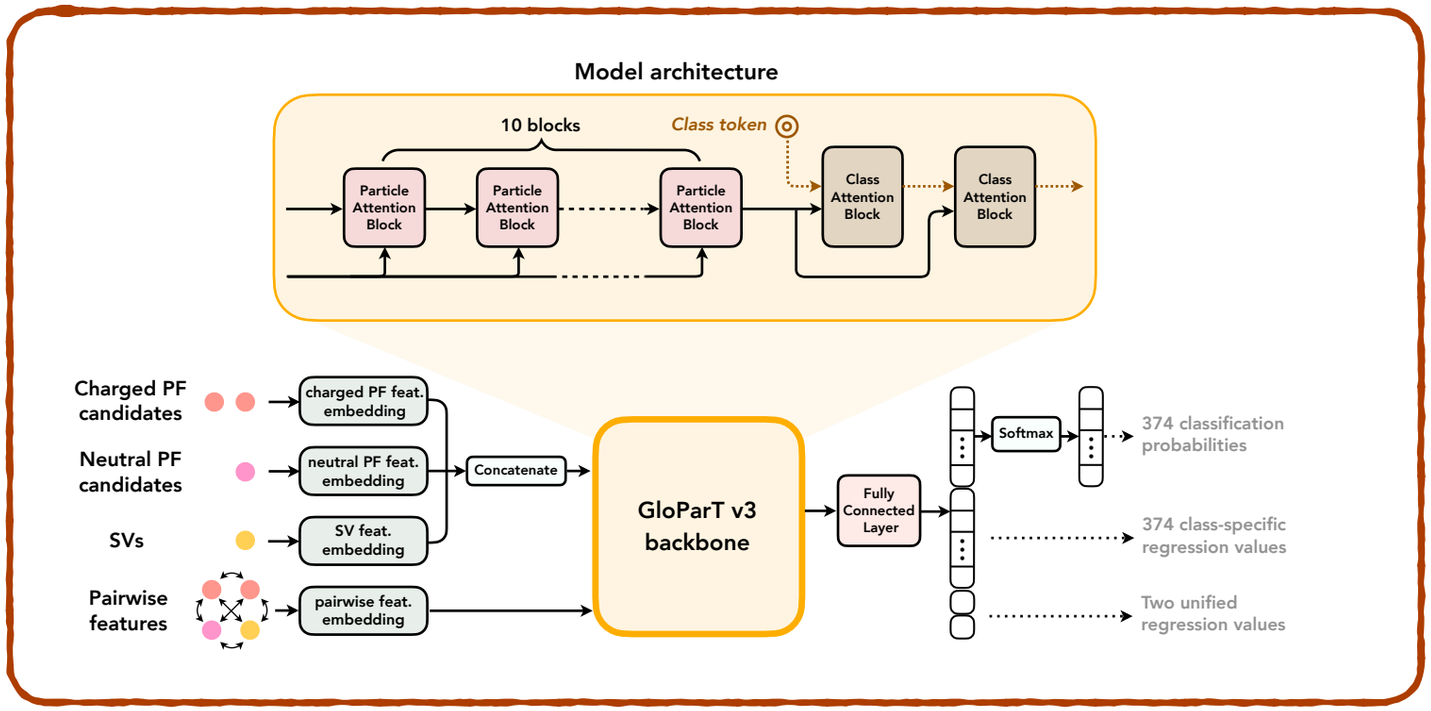
Is it absolutely task specific?

How Do We Read ML's Mind?

Are Are there strong correlation in between the hidden layers? Do each part of a network tries to learn different features?

Is it an artifact of the training fine-tuning ?

Do all of these aspects have mutual roles to play?



Are the higher order showering / hadronization effects being captured better by the raw inputs which are mis-modeled in fixed order MC ?

How do we map back to all of these to our physics driven reasoning?

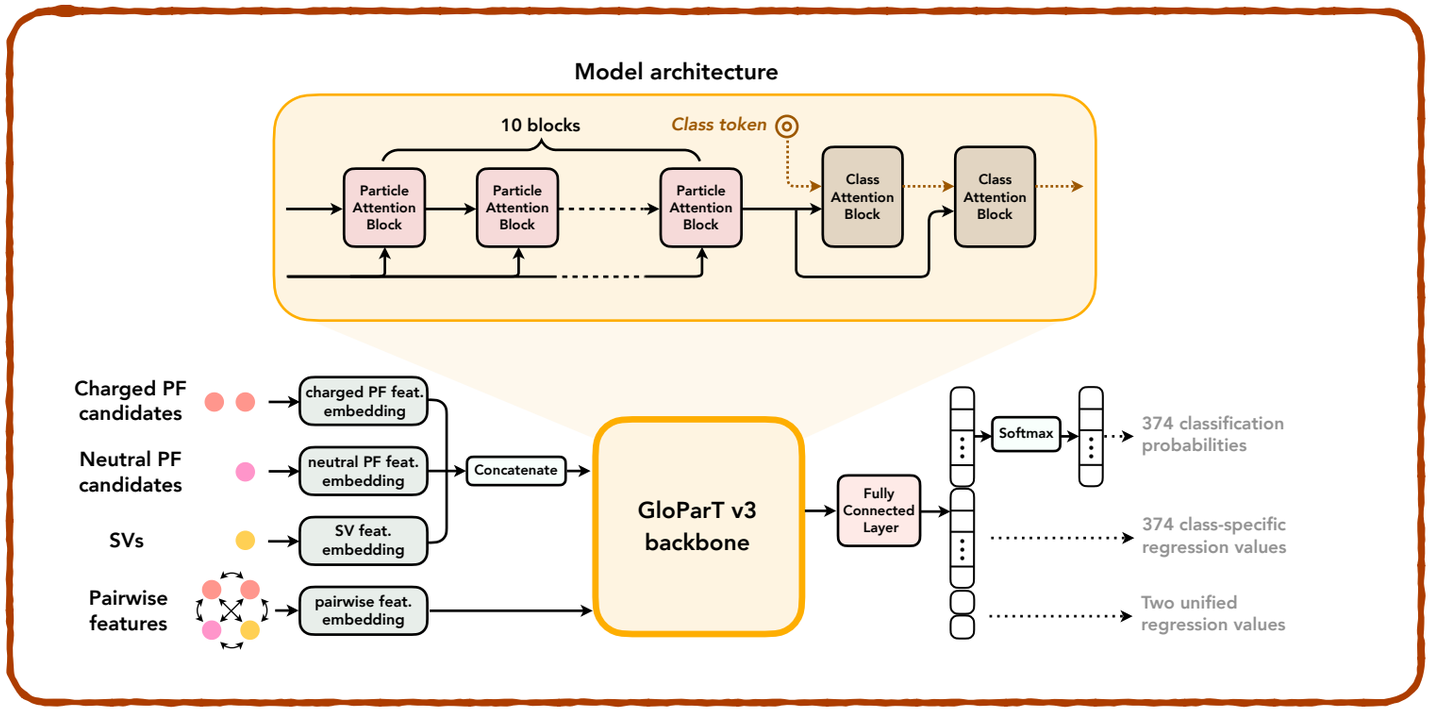
Is it absolutely task specific?

How Do We Read ML's Mind?

Are Are there strong correlation in between the hidden layers? Do each part of a network tries to learn different features?

Is it an artifact of the training fine-tuning ?

Do all of these aspects have mutual roles to play?



Are the higher order showering / hadronization effects being captured better by the raw inputs which are mis-modeled in fixed order MC ?

How do we map back to all of these to our physics driven reasoning?

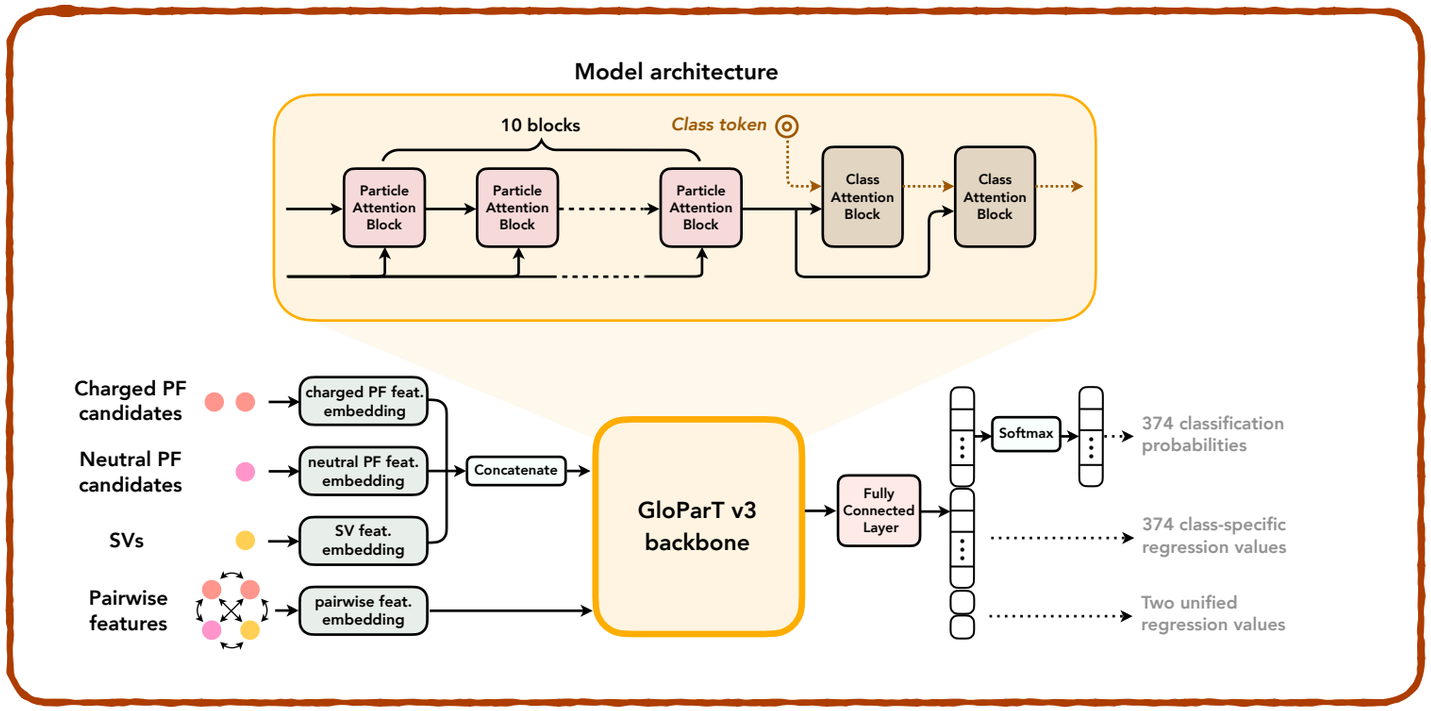
Is it absolutely task specific?

How Do We Read ML's Mind?

Are Are there strong correlation in between the hidden layers? Do each part of a network tries to learn different features?

Is it an artifact of the training fine-tuning ?

Do all of these aspects have mutual roles to play?



Are the higher order showering / hadronization effects being captured better by the raw inputs which are mis-modeled in fixed order MC ?

How do we map back to all of these to our physics driven reasoning?

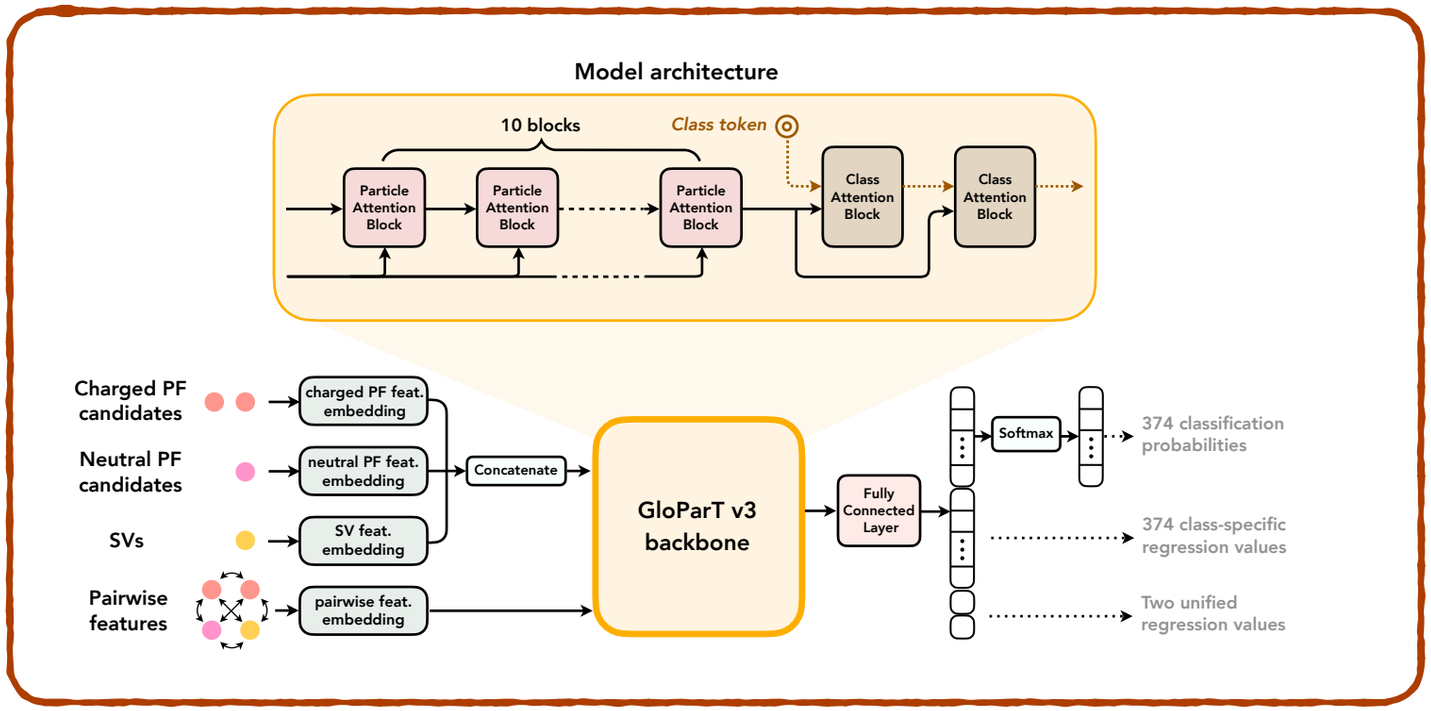
Is it absolutely task specific?

How Do We Read ML's Mind?

Are Are there strong correlation in between the hidden layers? Do each part of a network tries to learn different features?

Is it an artifact of the training fine-tuning ?

Do all of these aspects have mutual roles to play?

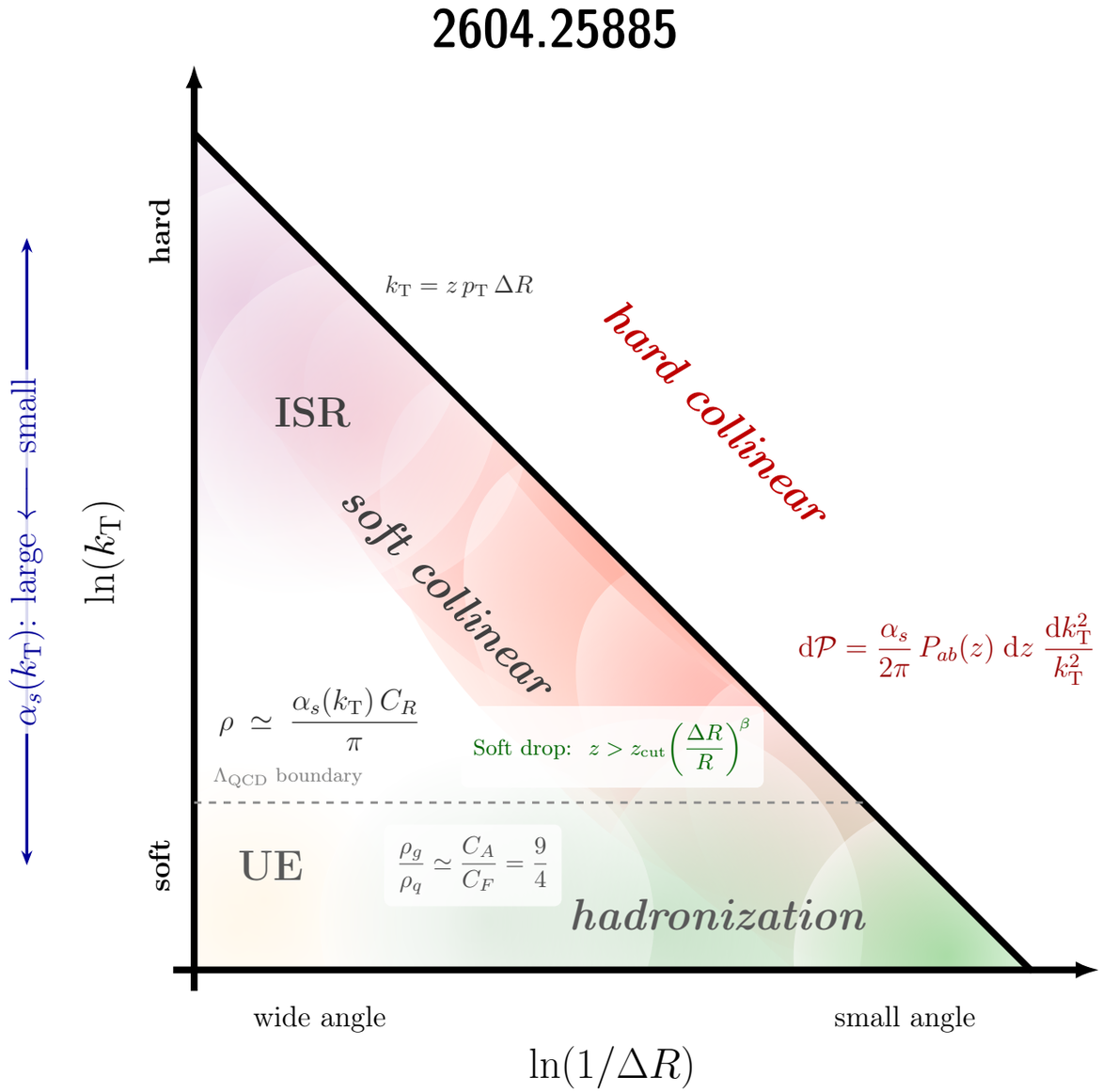


Are the higher order showering / hadronization effects being captured better by the raw inputs which are mis-modeled in fixed order MC ?

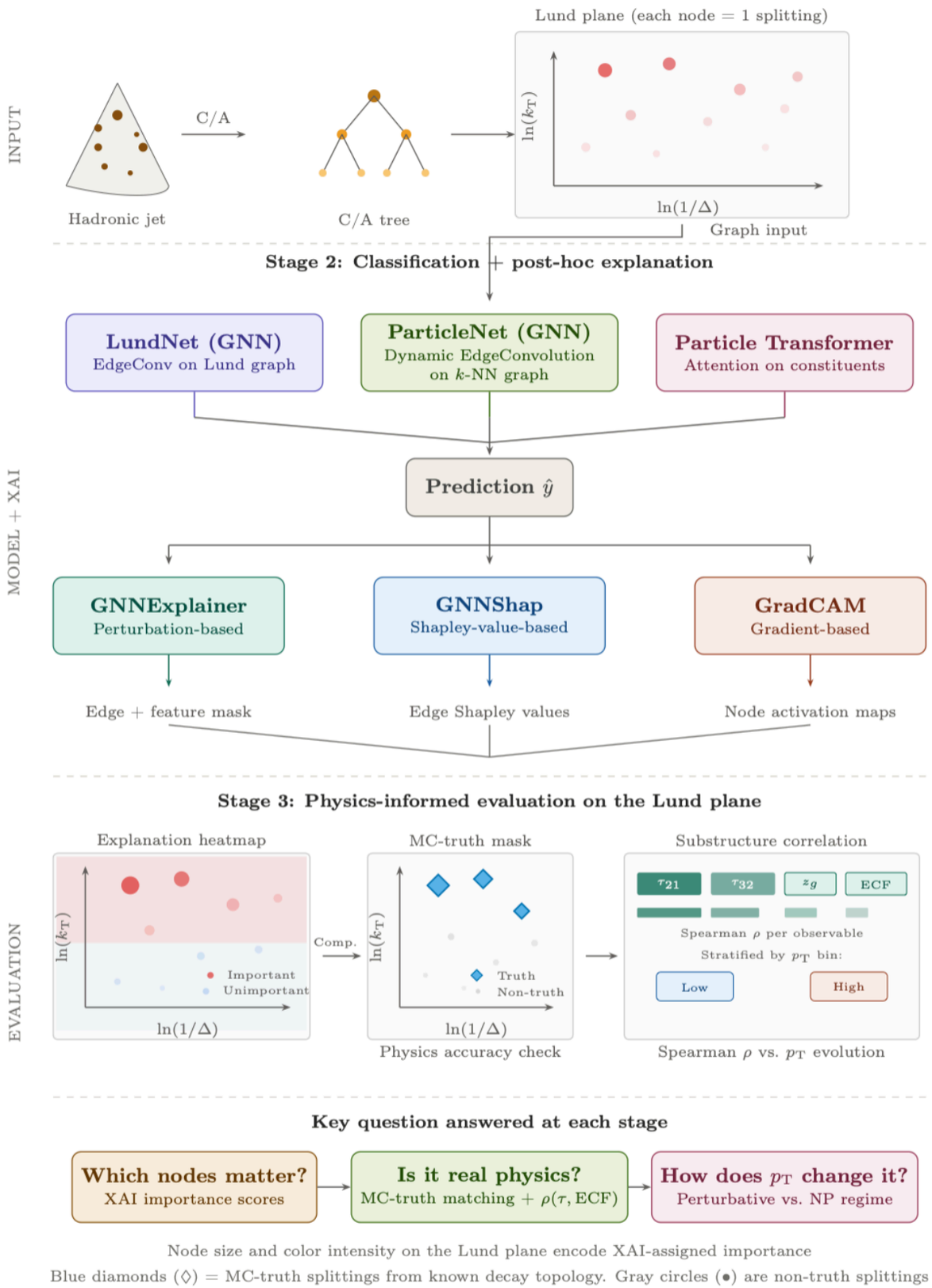
How do we map back to all of these to our physics driven reasoning?

Is it absolutely task specific?

Downstream Tasks on Jets : The Inputs



{LundNet, ParticleNet, Particle Transformer} × {GNNExplainer, GNNShap, GradCAM}



Working of the explainers

GNNExplainer (1903.03894)

$$\max_{\mathcal{G}_{\text{expl}}} I(Y, \mathcal{G}_{\text{expl}}) = H(Y) - H(Y \mid \mathcal{G}_{\text{expl}})$$

$$\mathcal{G}_{\text{expl}} = \big(V,\ E \odot \sigma(\mathbf{M}),\ \mathbf{X} \odot \sigma(\mathbf{F})\big)$$

GNNShap (2401.04829)

Shapley value for the i-th edge, ϕ_i is computed by the formula

$$\phi_i = \sum_{S \subseteq E_c \setminus \{e_i\}} \frac{|S|!(|E_c| - |S| - 1)!}{|E_c|!} (v(S \cup \{e_i\}) - v(S))$$

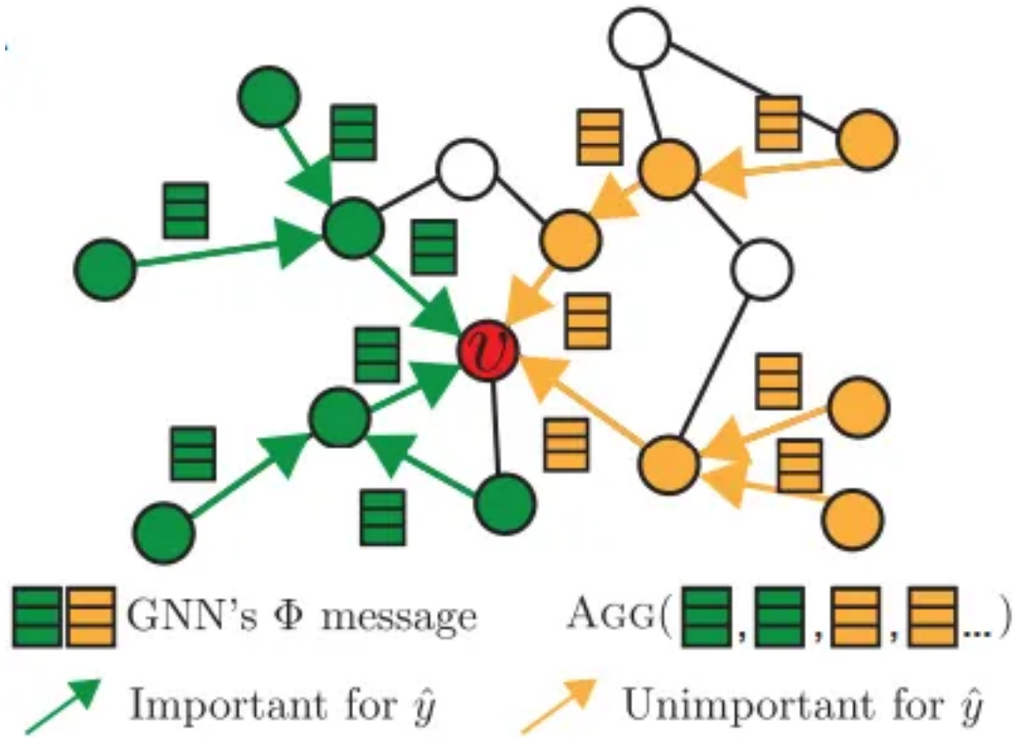
$$w_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} e_{ij}$$

GradCAM (2111.00722)

For class c, GradCAM computes the importance weight for channel k at layer ℓ as:

$$\alpha_k^{c,(\ell)} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial y^c}{\partial H_{ik}^{(\ell)}}$$

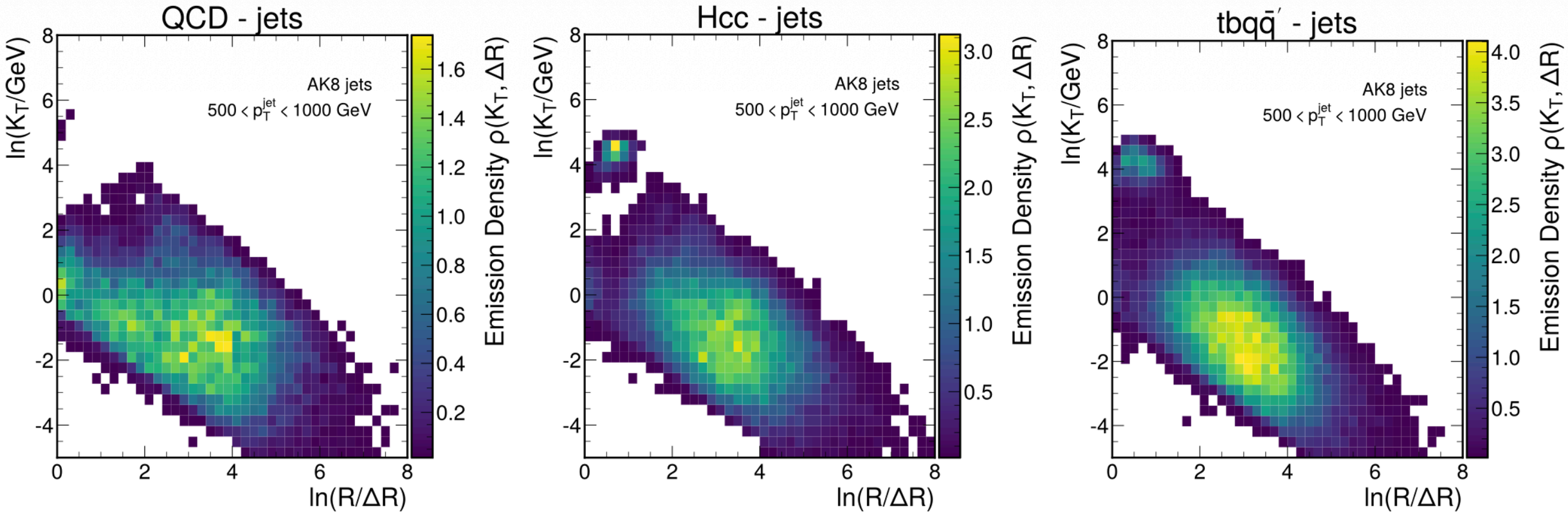
$$w_i^{c,(\ell)} = \text{ReLU}\Big(\sum_{k=1}^{K_\ell} \alpha_k^{c,(\ell)}\, H_{ik}^{(\ell)}\Big)$$



Pic credit : S. Khurana's Medium Article

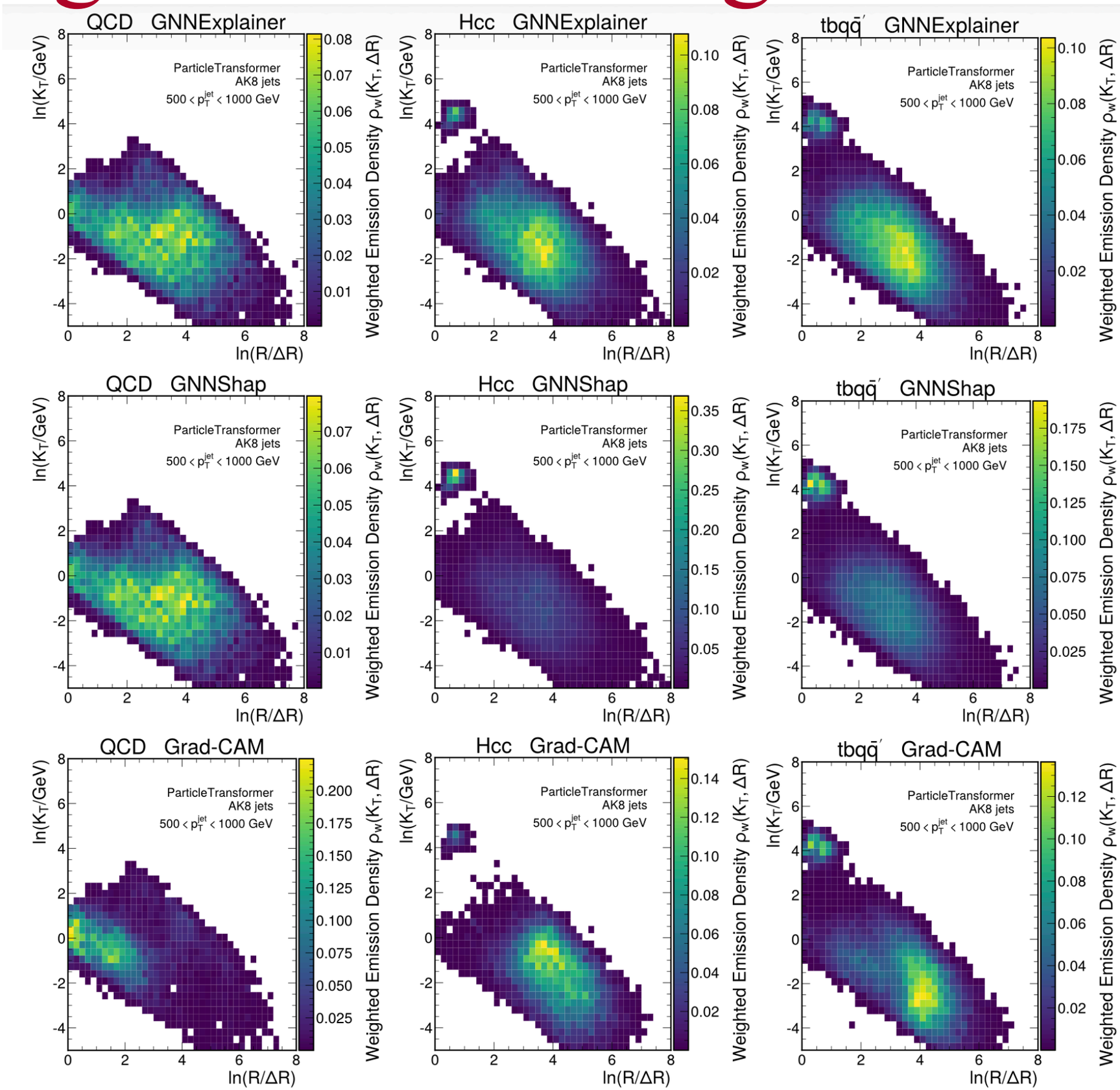
The Truth Lund-Plane Weight

$$\rho(k_{\mathrm{T}}\,,\Delta\mathrm{R}) = \frac{1}{N_{jets}}\frac{d^2 N_{emissions}}{d\ ln(k_T)\ \ d\ ln(\frac{R}{\Delta\mathrm{R}})}$$



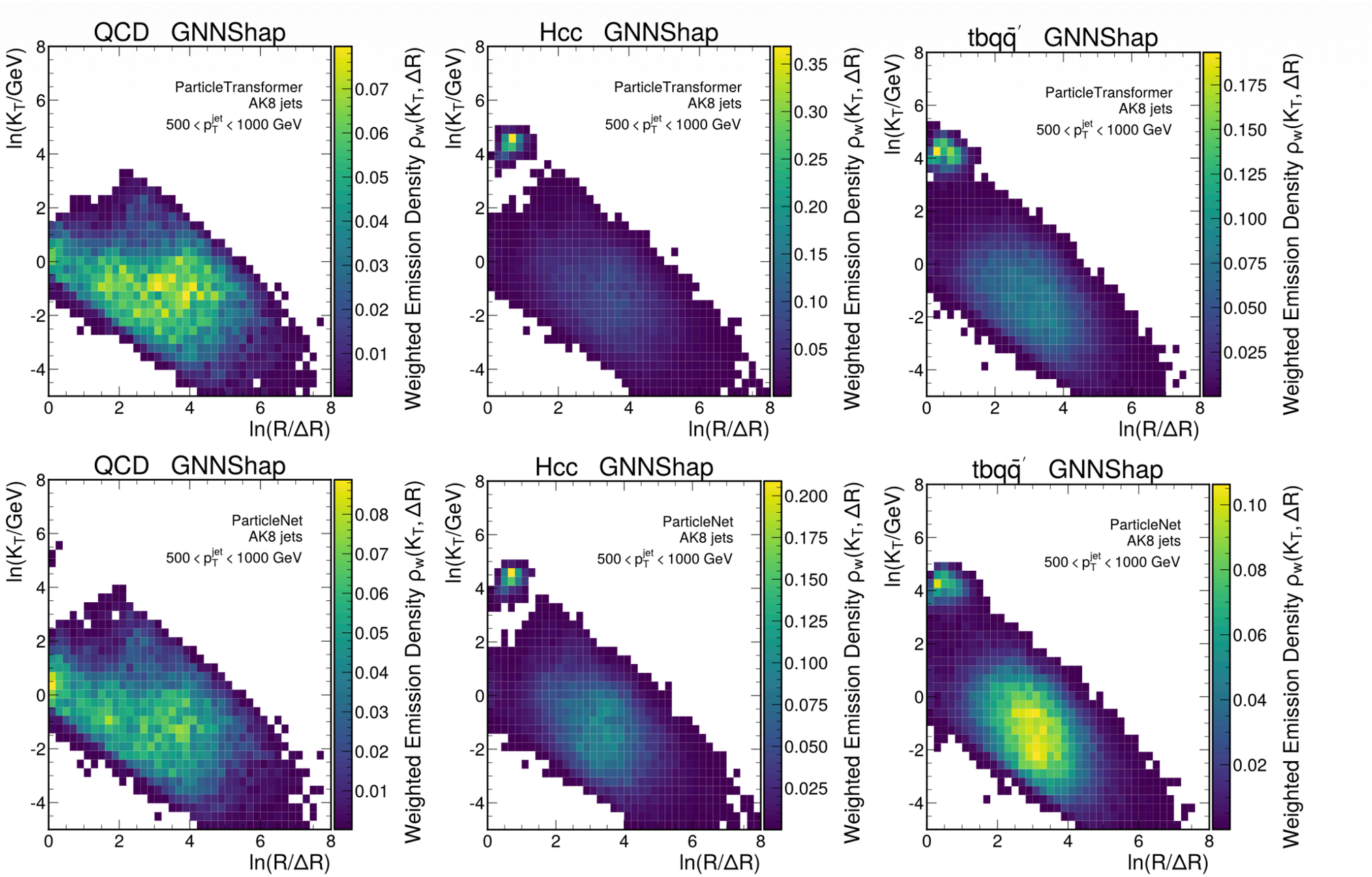
$$\rho_W(k_{\mathrm{T}}\,,\Delta\mathrm{R}) = \frac{1}{N_{jets}}\frac{d^2\left(\sum_{i=1}^{N_{emissions}} w_i\right)}{d\ ln(k_T)\ \ d\ ln(\frac{R}{\Delta\mathrm{R}})}$$

Aggregated Node Weights

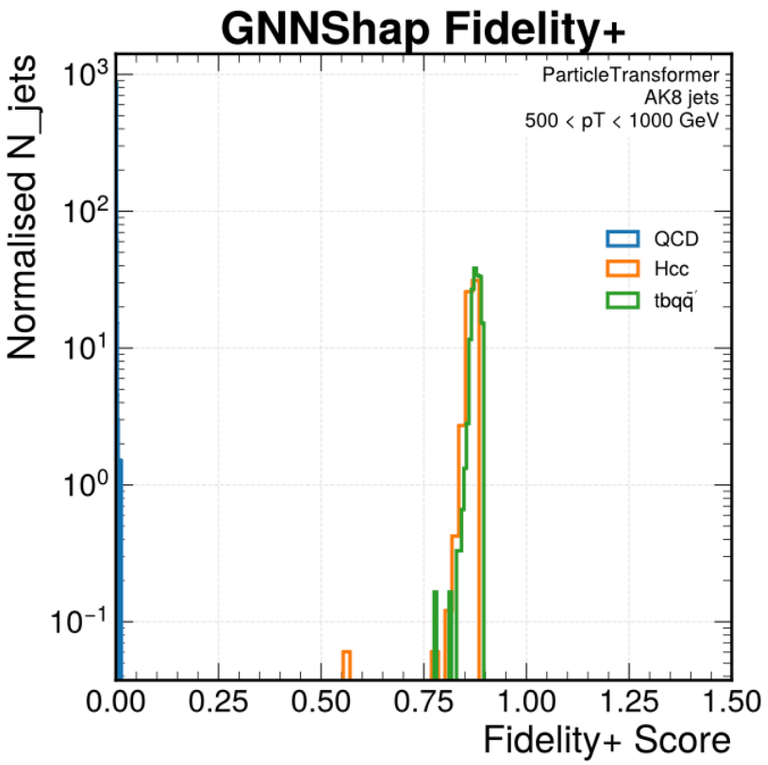


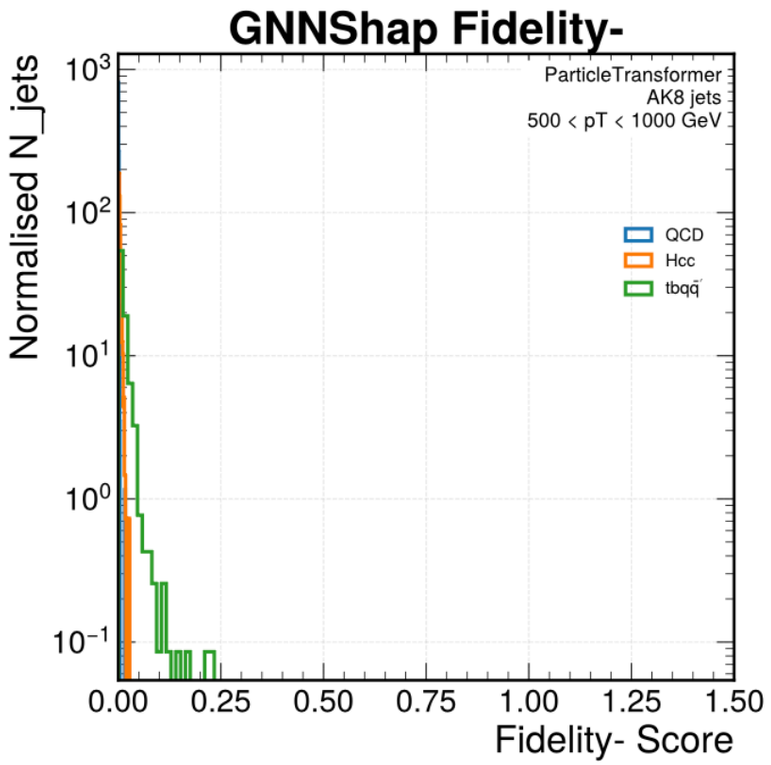
Aggregated Node Weights

Aggregated Node Weights



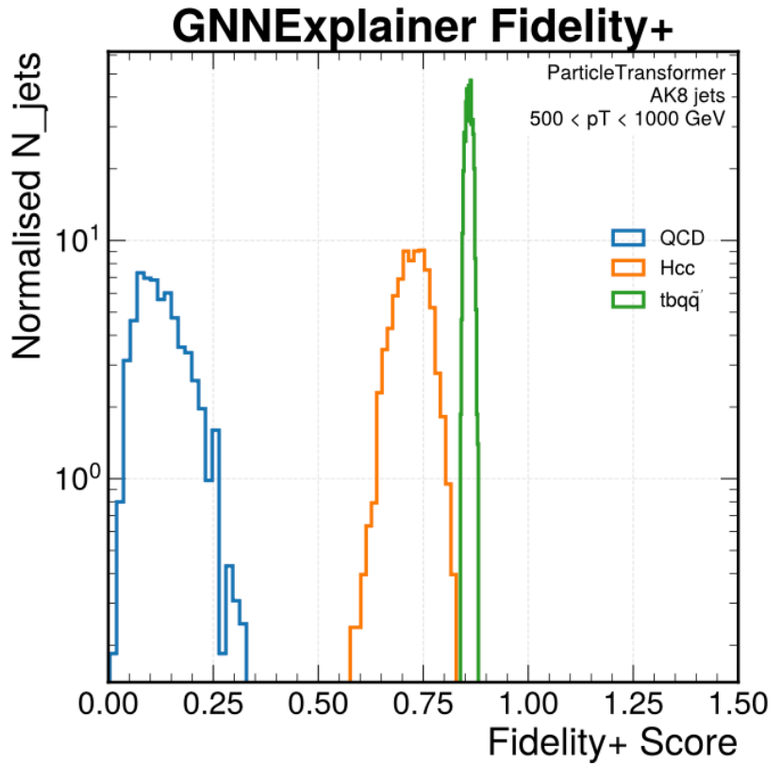
How do we make sense of these?

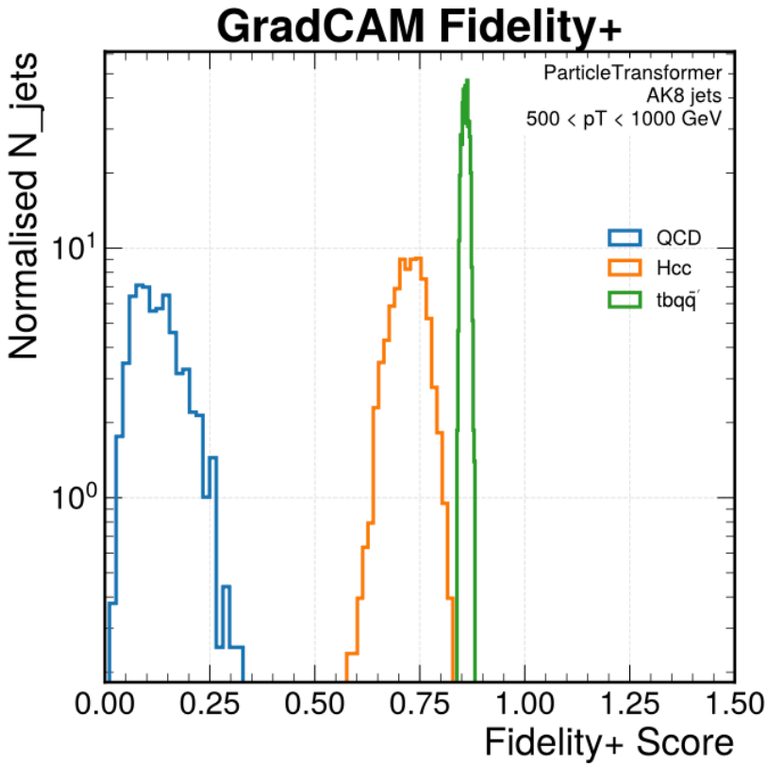




$$\mathrm{Fid}^{-} = f(\mathcal{G}) - f(\mathcal{G}_{\mathrm{expl}})\,,$$
$$\mathrm{Fid}^{+} = f(\mathcal{G}) - f(\mathcal{G} \setminus \mathcal{G}_{\mathrm{expl}})\,,$$

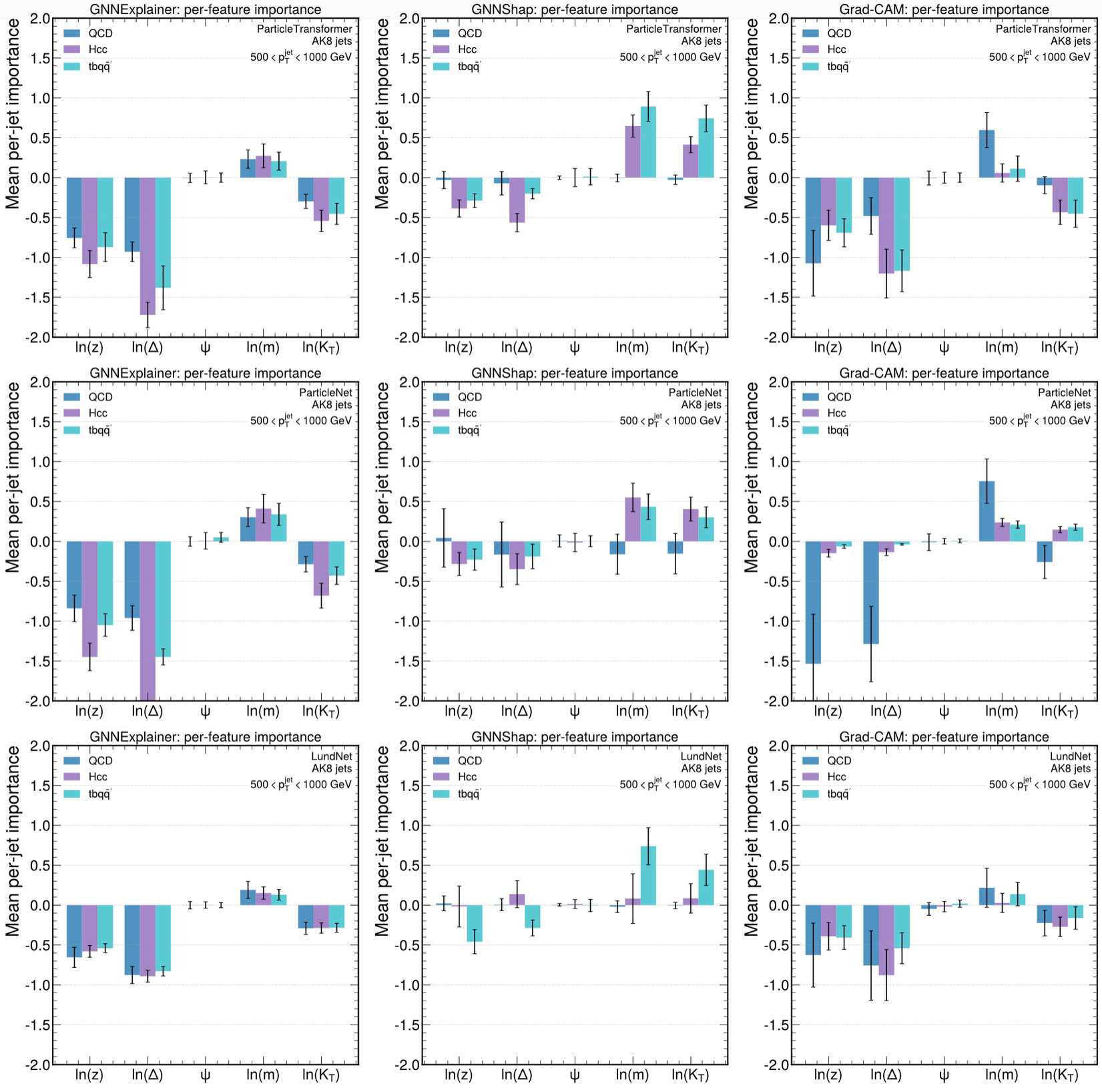
The Fid+ score shift for QCD is counter intuitive.





Follows from LJP distribution of QCD.

The average feature weight



$$\bar{F}_i = \frac{\sum_{j \in V_\mathcal{G}} w_j F_{ij}}{\sum_{j \in V_\mathcal{G}} w_j}$$

$\ln(z)$ & $\ln(\Delta)$ play a

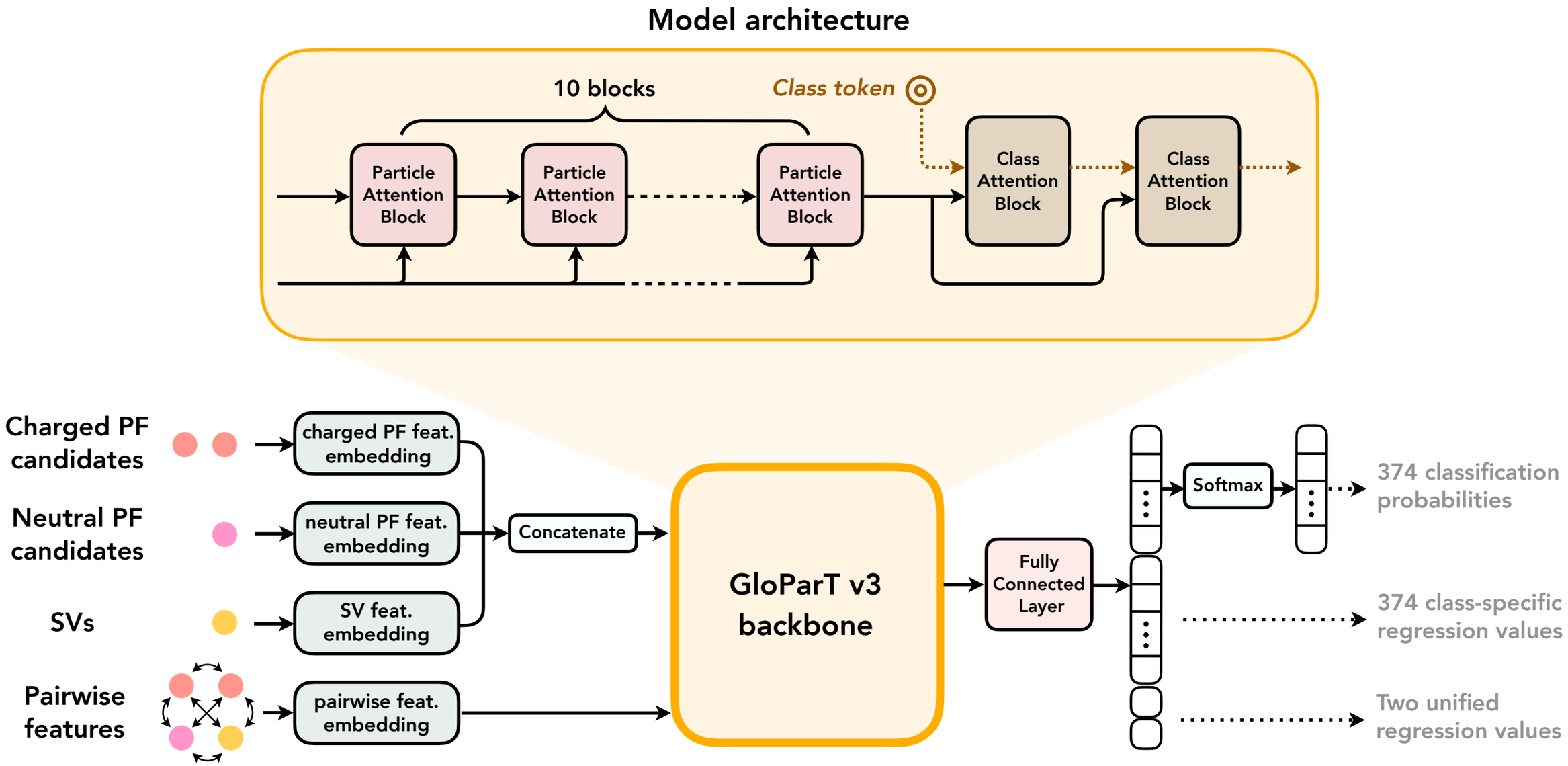
Crucial role towards decision making.

$\ln(m)$ & $\ln(k_T)$ also plays a

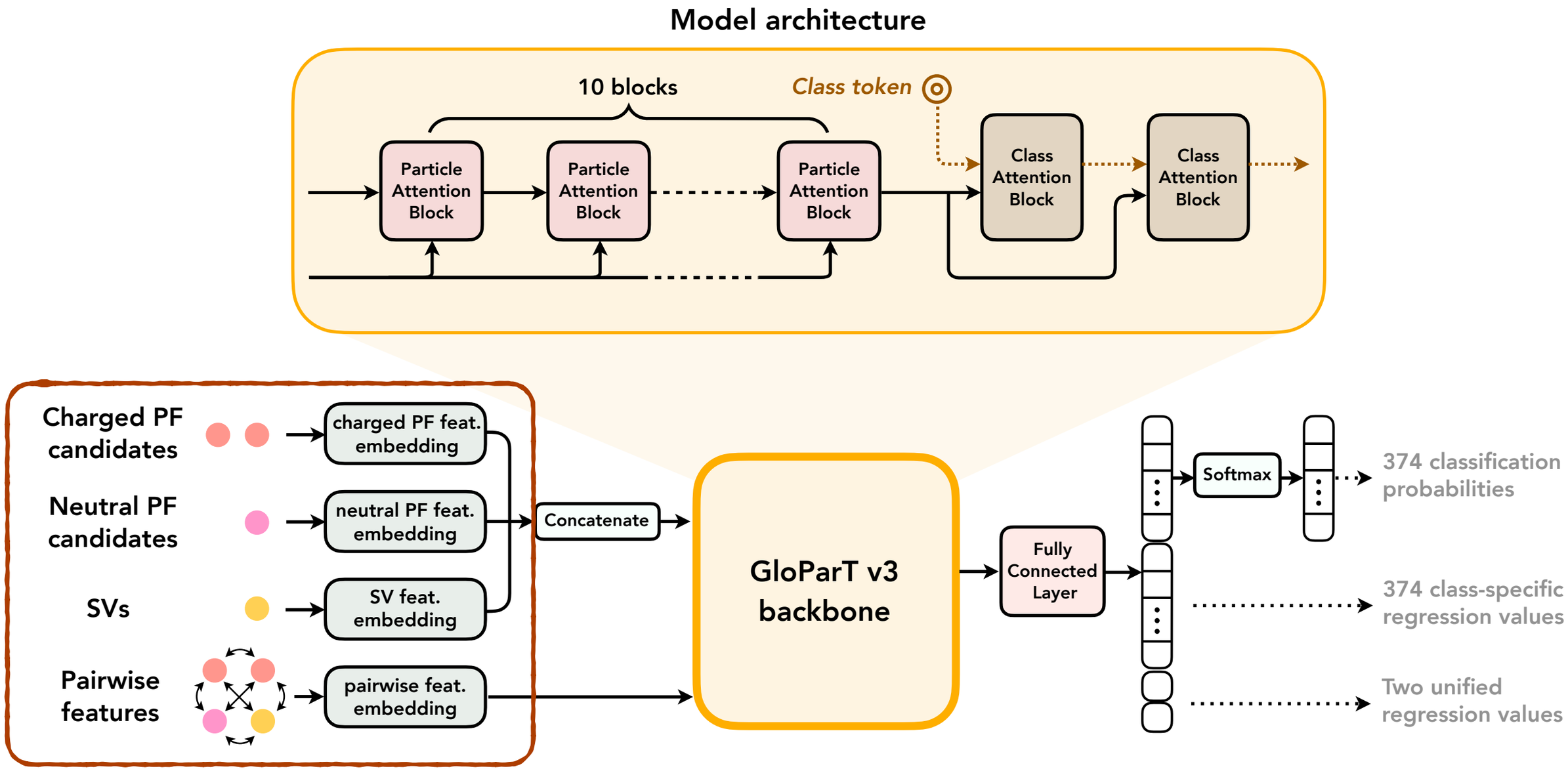
sizeable role for identification of 2 abd 3-prong jets.

The azimuthal angle has no role to play, a validation test.

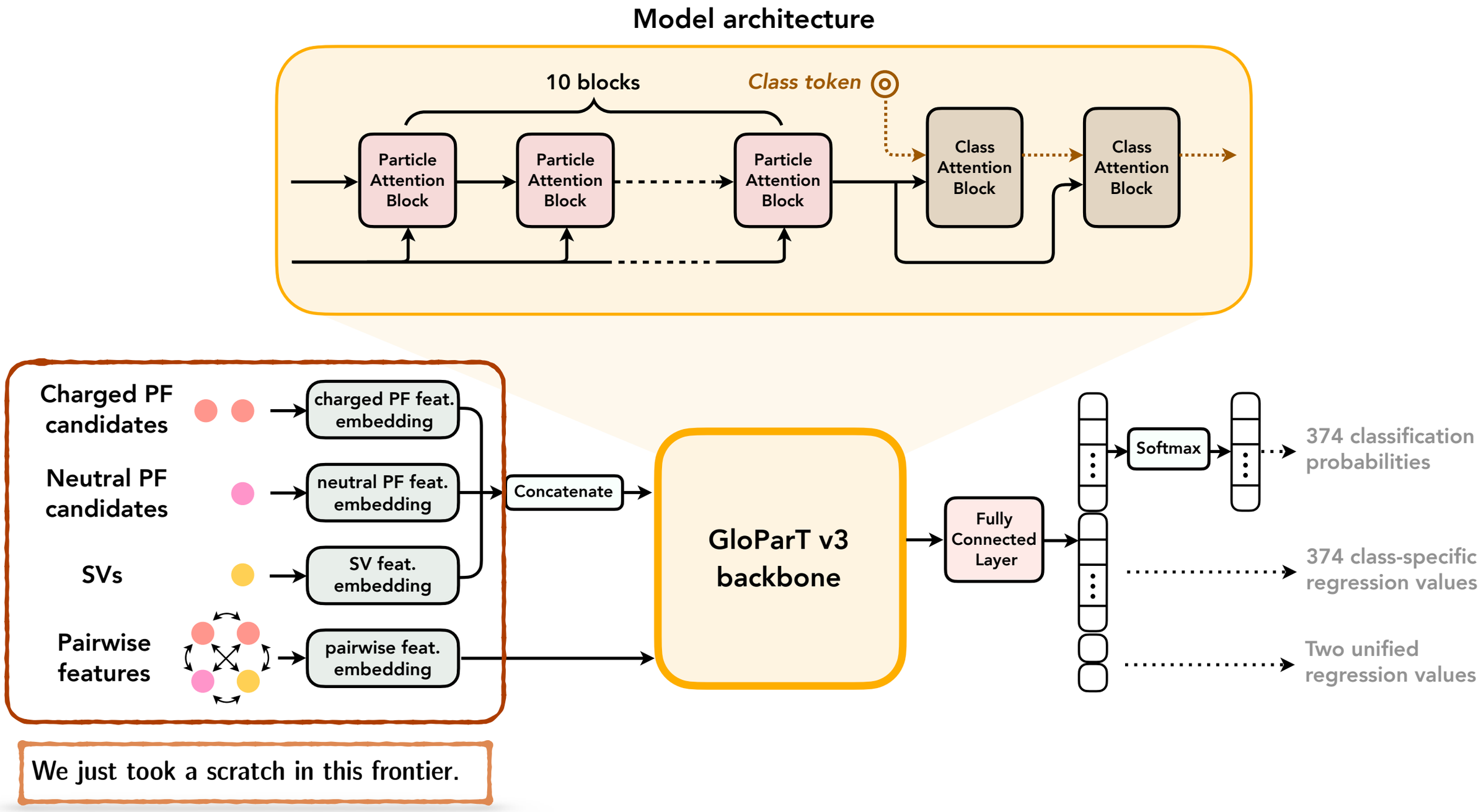
Transiting Towards Architecture Dynamics



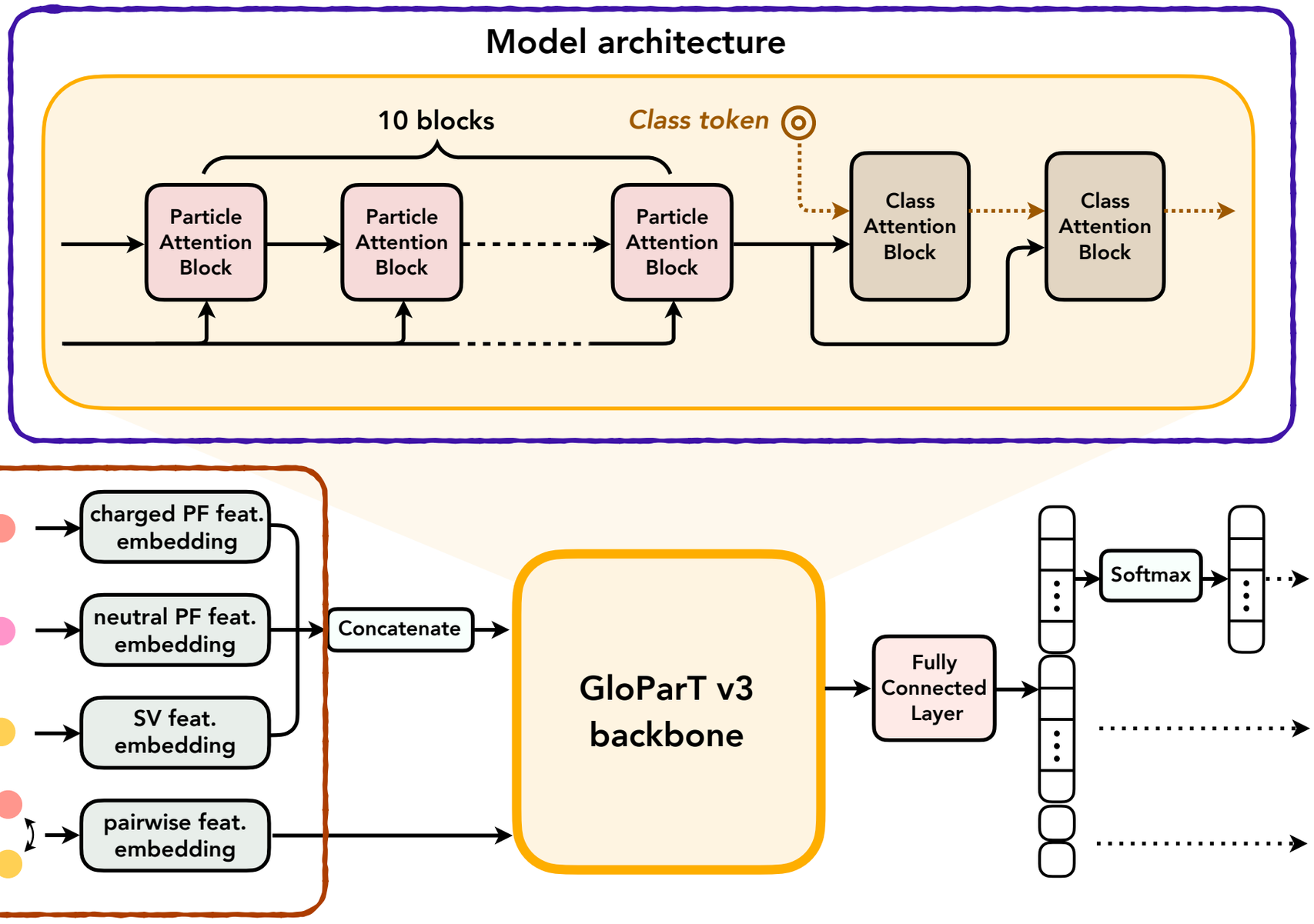
Transiting Towards Architecture Dynamics



Transiting Towards Architecture Dynamics



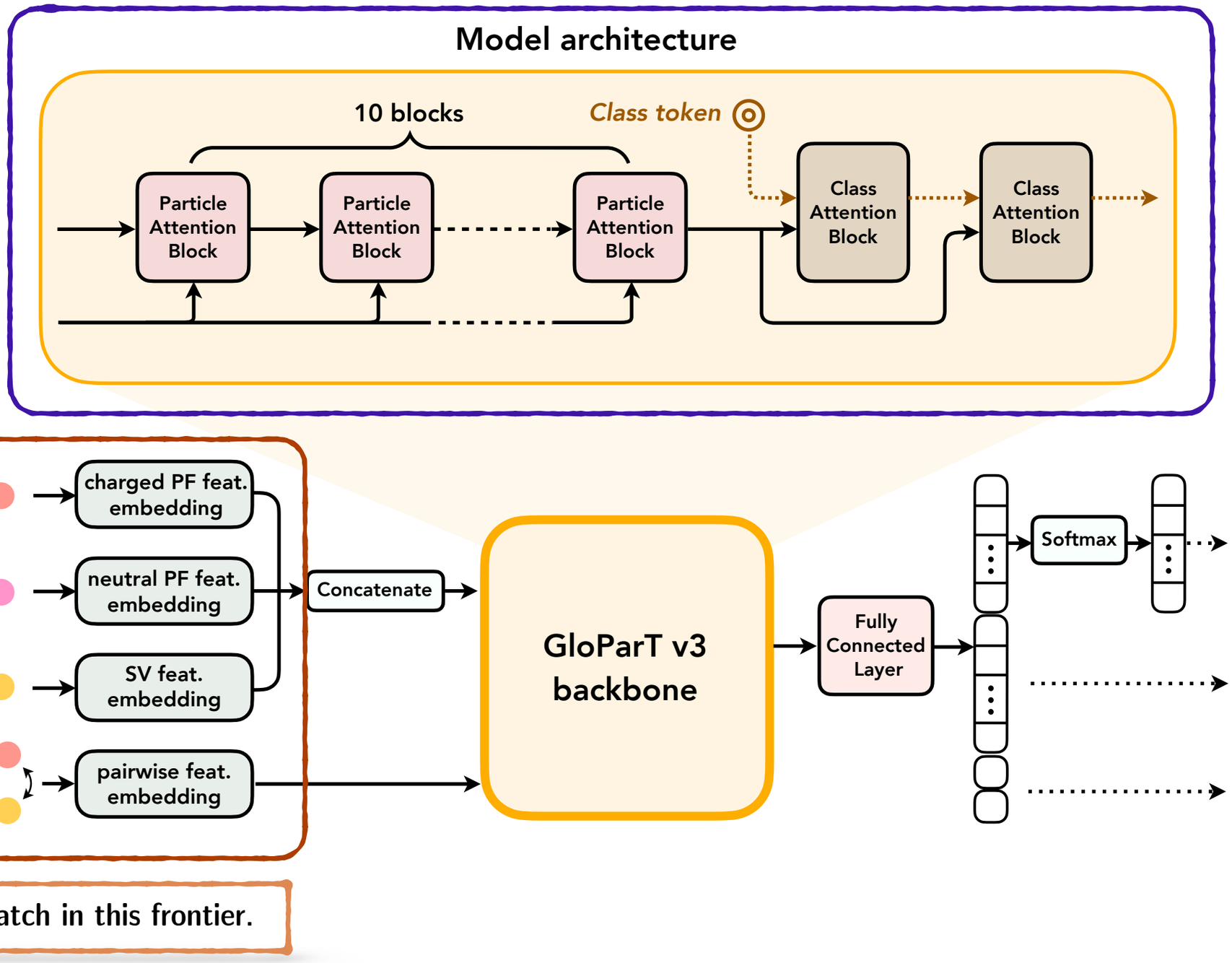
Transiting Towards Architecture Dynamics



We just took a scratch in this frontier.

Transiting Towards Architecture Dynamics

How do we understand the architecture dynamics?



We just took a scratch in this frontier.

Dissecting A Tagger Through Mech-Int

Boosted top jet

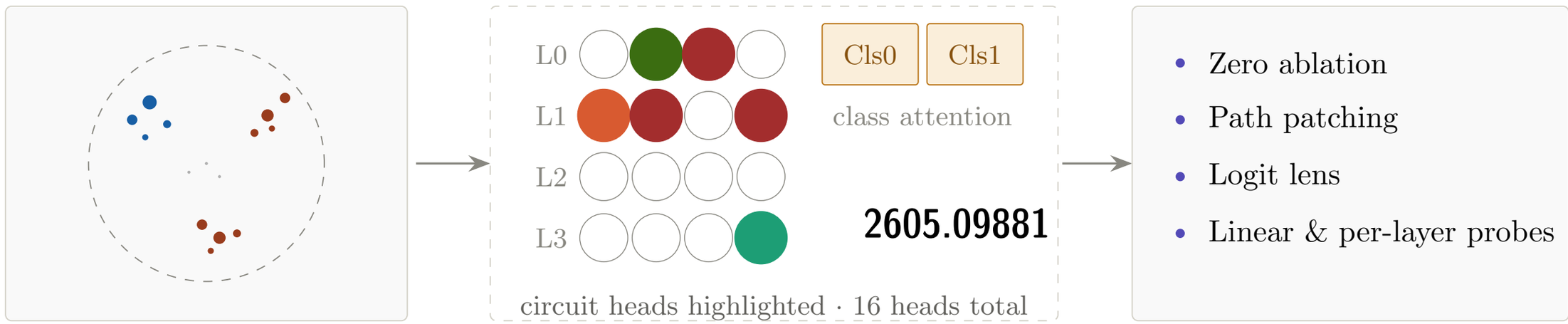
$t \to W(q\bar{q}) + b$

Particle Transformer

4 particle attention + 2 class attention

Mech-interp toolkit

applied to internal activations



Mechanistic interpretability tries to identify a minimal set of components (circuits) that are causally responsible for a learned behavior. The intervention methods are as following :

Zero ablation : switch off one head, measure the AUC drop (or whatever the performance metric is).

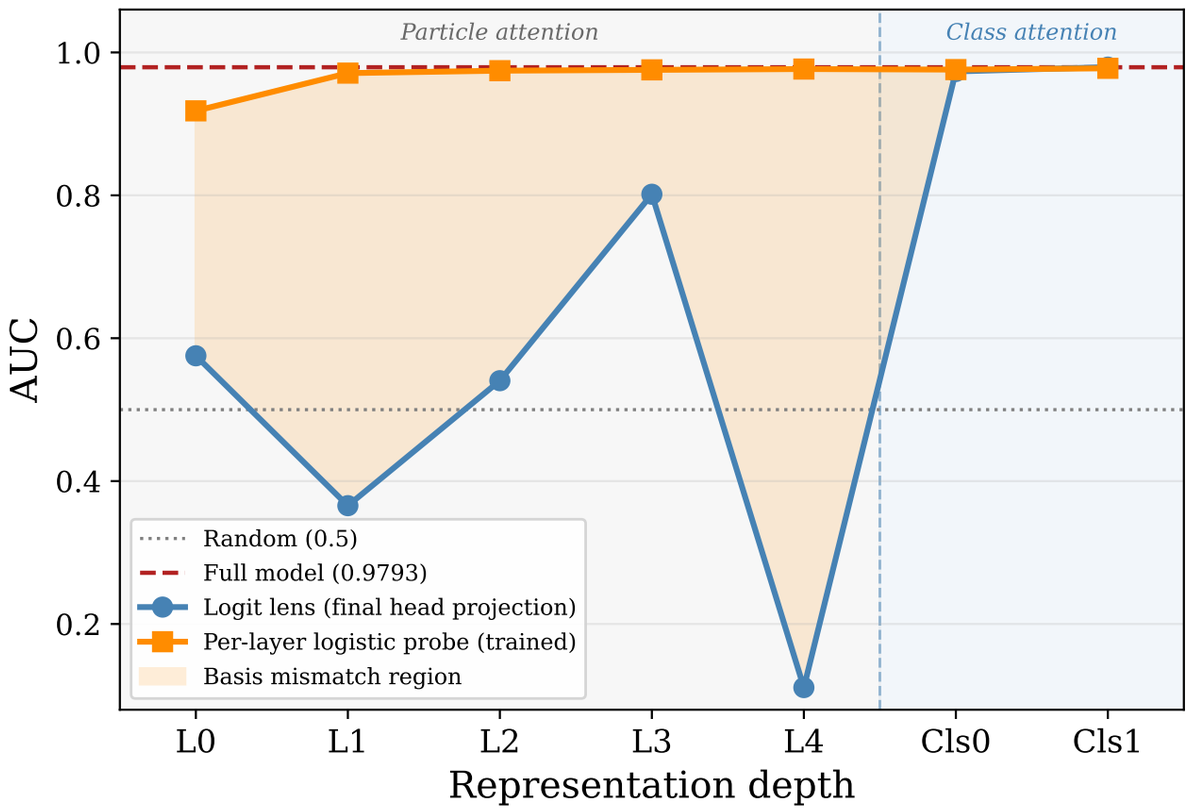
Path patching : swap one head's activation clean ==> corrupted run, see how much recovers (conceptual swap test).

Logit Lens : read the classification score out at every layer (a partial reconstruction).

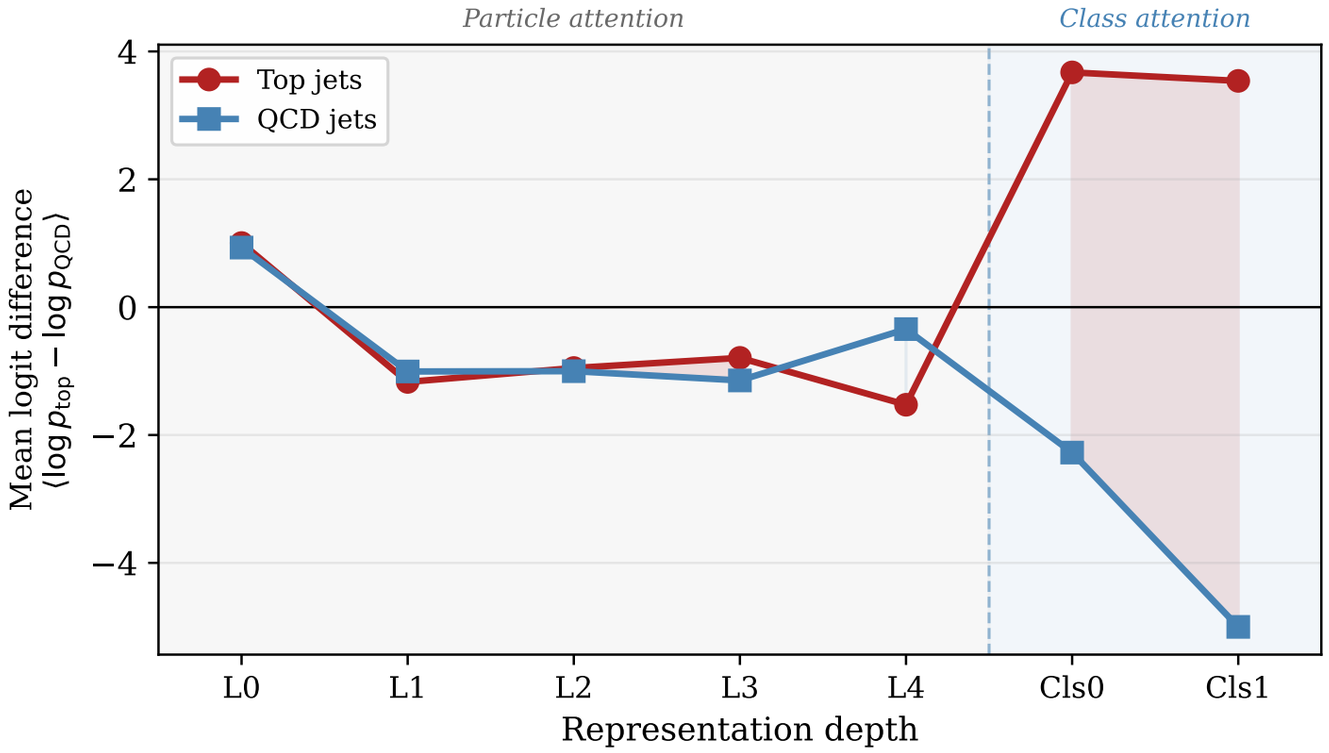
Linear Per-Layer Probing : runs a linear regression from intermediate layers to target (R^2 coefficient).

The Logit-Lens / Per Layer Probe

2605.09881



Depth	Logit lens	Per-layer probe	Difference
L0	0.575	0.918	+0.343
L1	0.366	0.971	+0.605
L2	0.541	0.974	+0.434
L3	0.801	0.976	+0.174
L4	0.111	0.977	+0.866
Cls0	0.973	0.976	+0.003
Cls1	0.979	0.978	−0.002



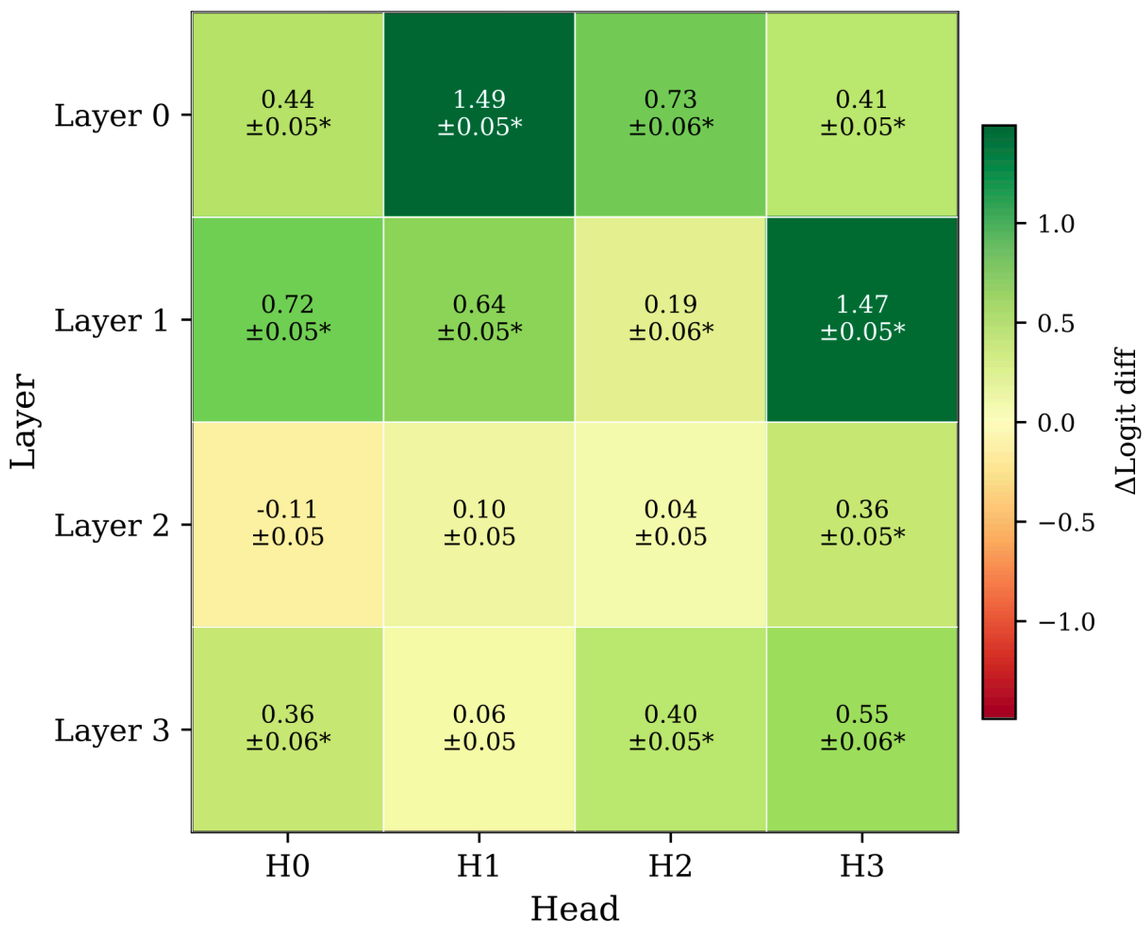
Naive logit-lens AUC :

$$\mathbf{0.575 \to 0.366 \to 0.541 \to 0.801 \to 0.111}$$

Per layer trained probe $\sim \mathbf{0.97}$ after L1.

This indicates a basis rotation, not a new computation.

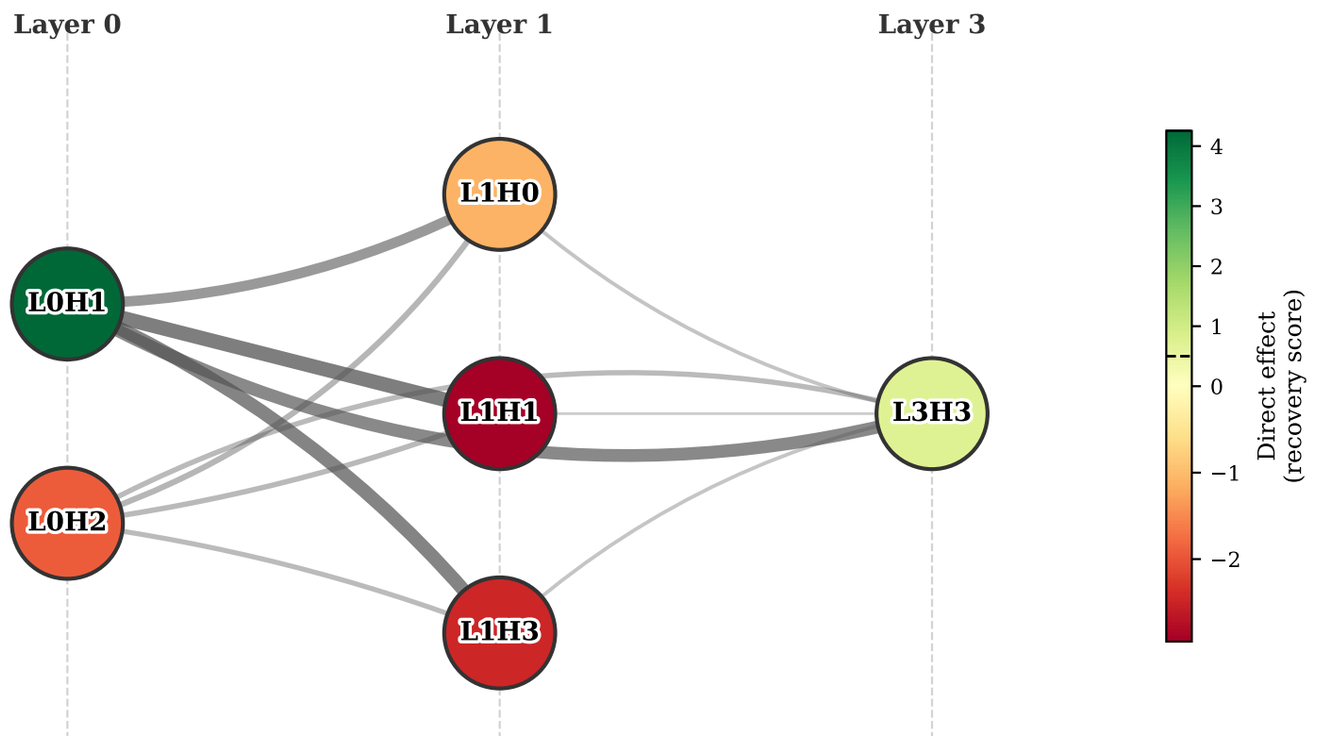
Head Importance Zero Ablation Method



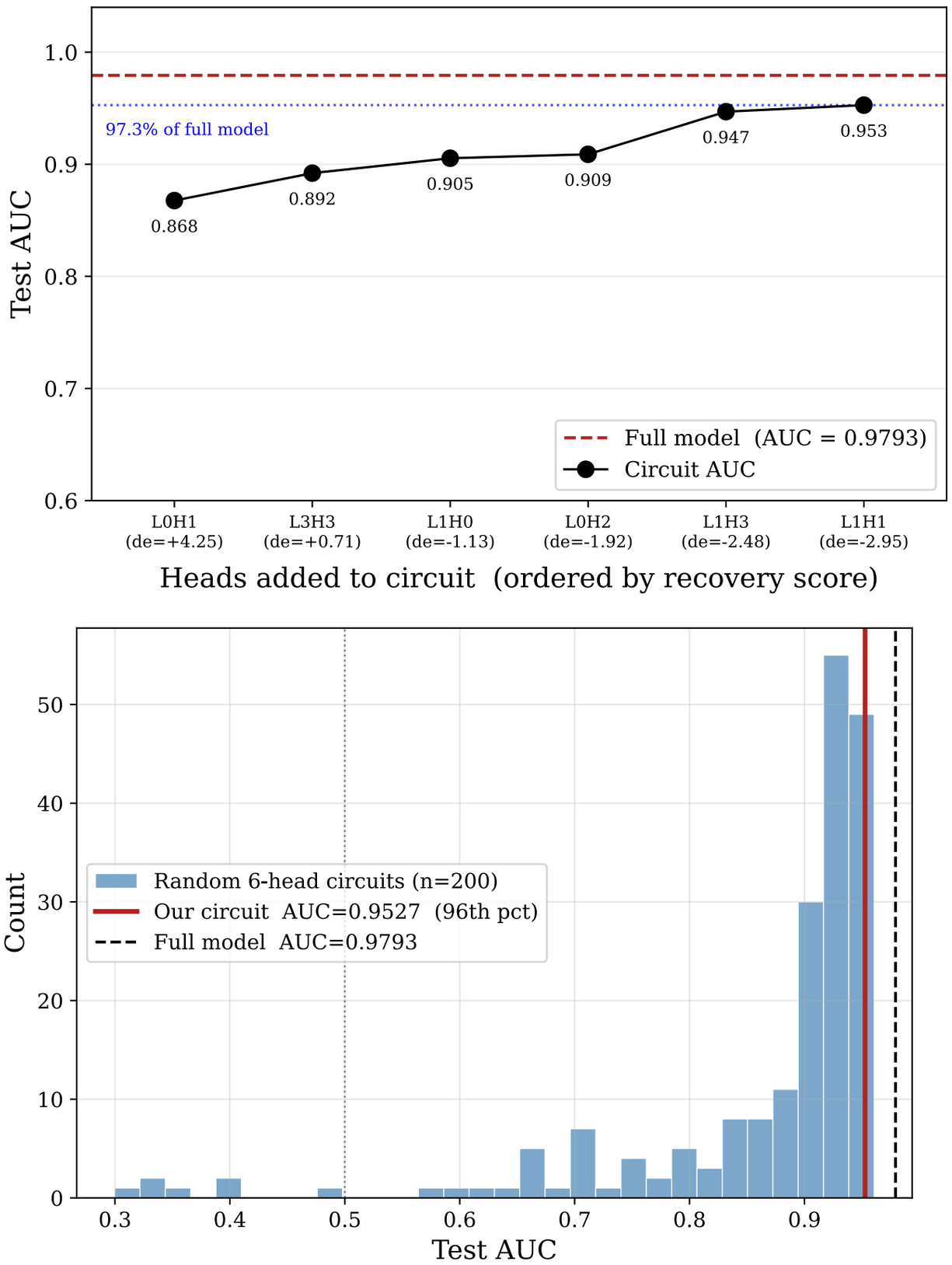
- ☑ 6 out of 16 heads recover $\sim 97\,\%$ of accuracy.
- ☑ L0H1 primary source (soft/collinear context).
- ☑ L1 (H0, H1, H3) ==> the relay heads recovers the W substructure.
- ☑ L3H3 performs the readout/aggregation.
- ☑ Tested over 5 training seeds & 2 independent corruption strategies.

Edge thickness encodes the path effect magnitude.

Head	Direct effect	Role
L0H1	+4.25	Primary source
L0H2	−1.92	Secondary source
L1H0	−1.13	Relay
L1H1	−2.95	Relay
L1H3	−2.48	Relay
L3H3	+0.71	Readout



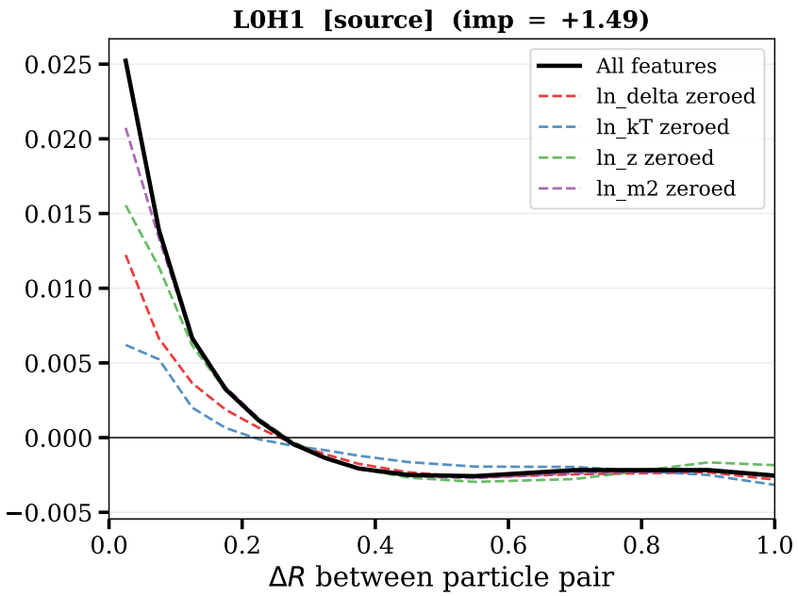
Head Importance Zero Ablation Method

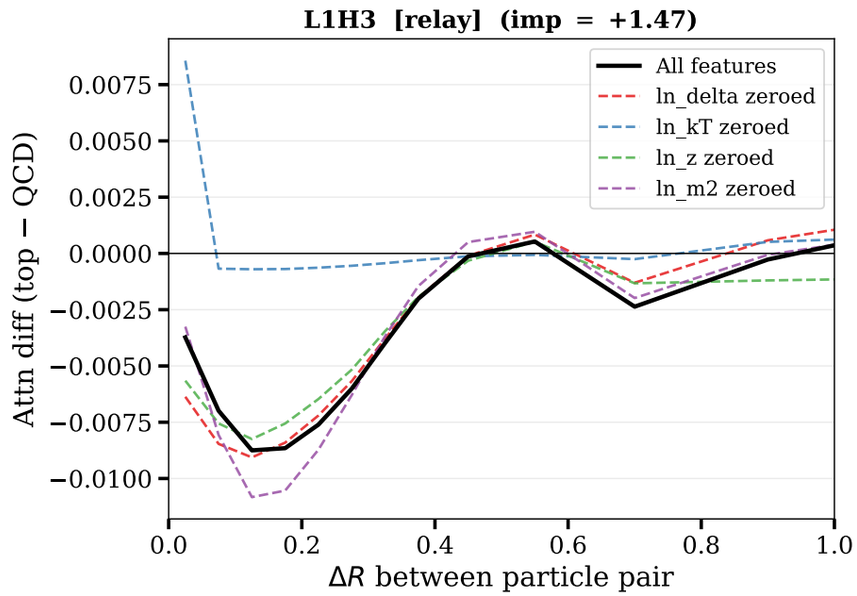


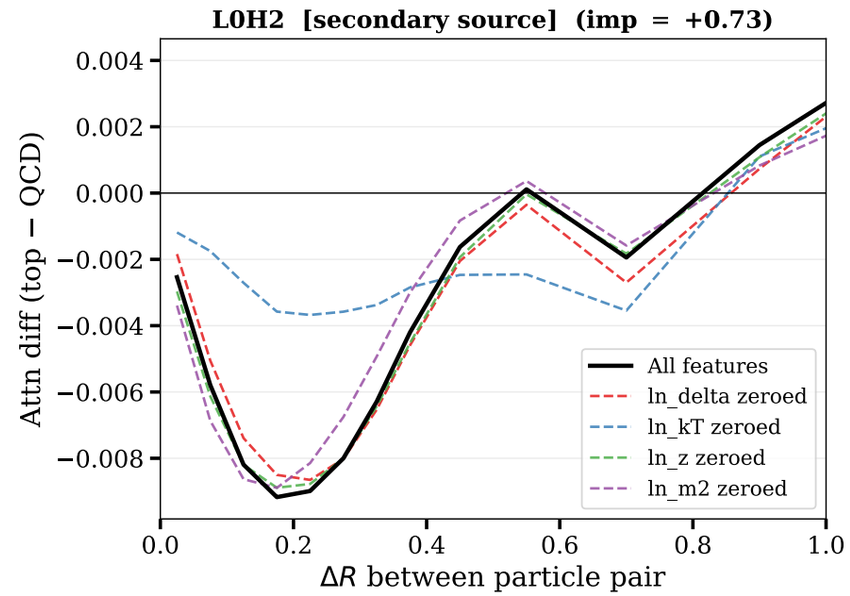
The two largest single-step gains occur when L0H1 is added (Step-1, +0.868 AUC) and the L1H5 head is added (Step 5, + 3.8% points)

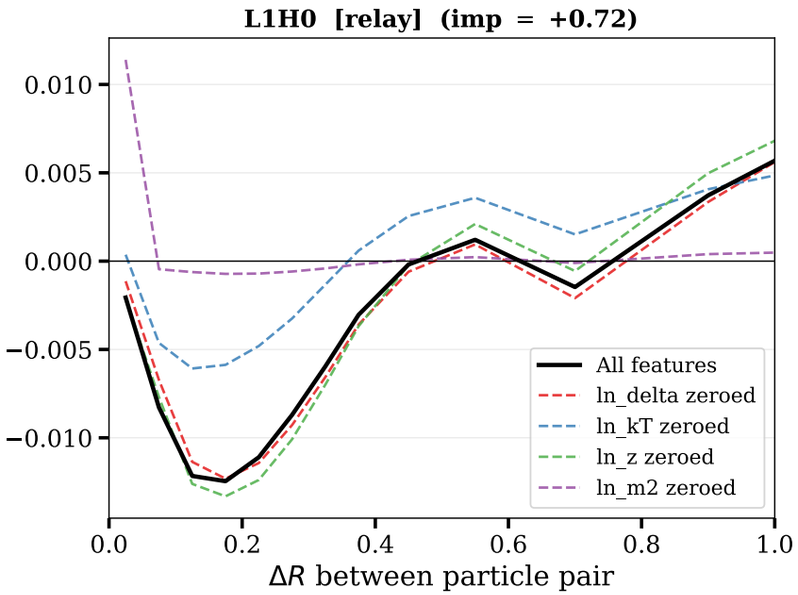
The random sparse subsets of six headed circuits have varied distribution. This shows part of the subset circuit is actually redundant.

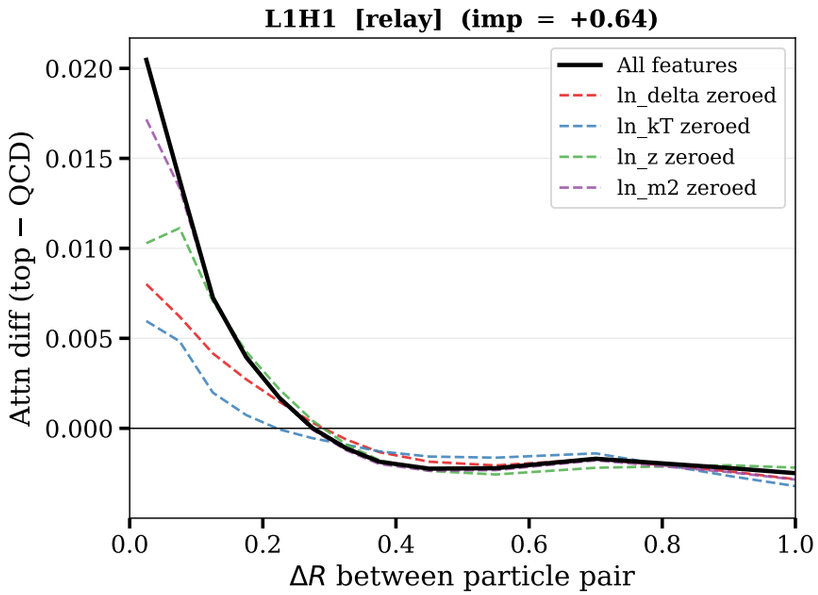
Ablation Applied To Physics Variables

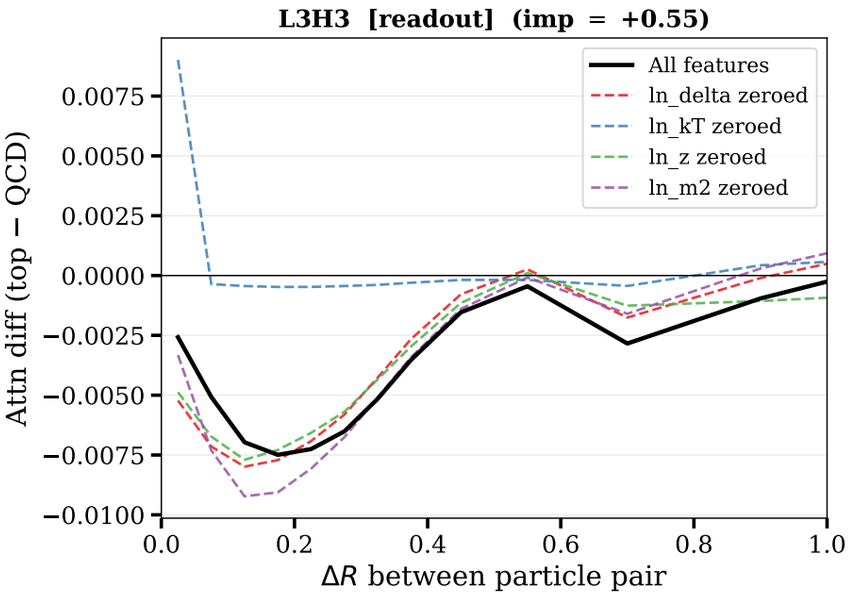










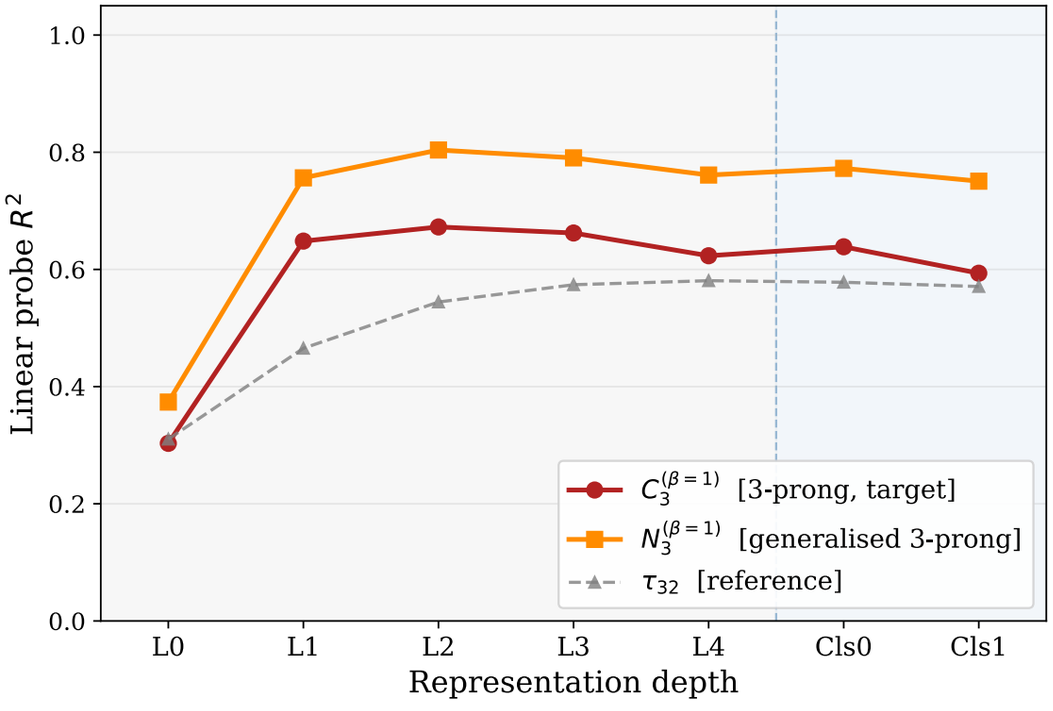


A large deviation from baseline indicates that for a given head, the corresponding feature has a large causal impact.

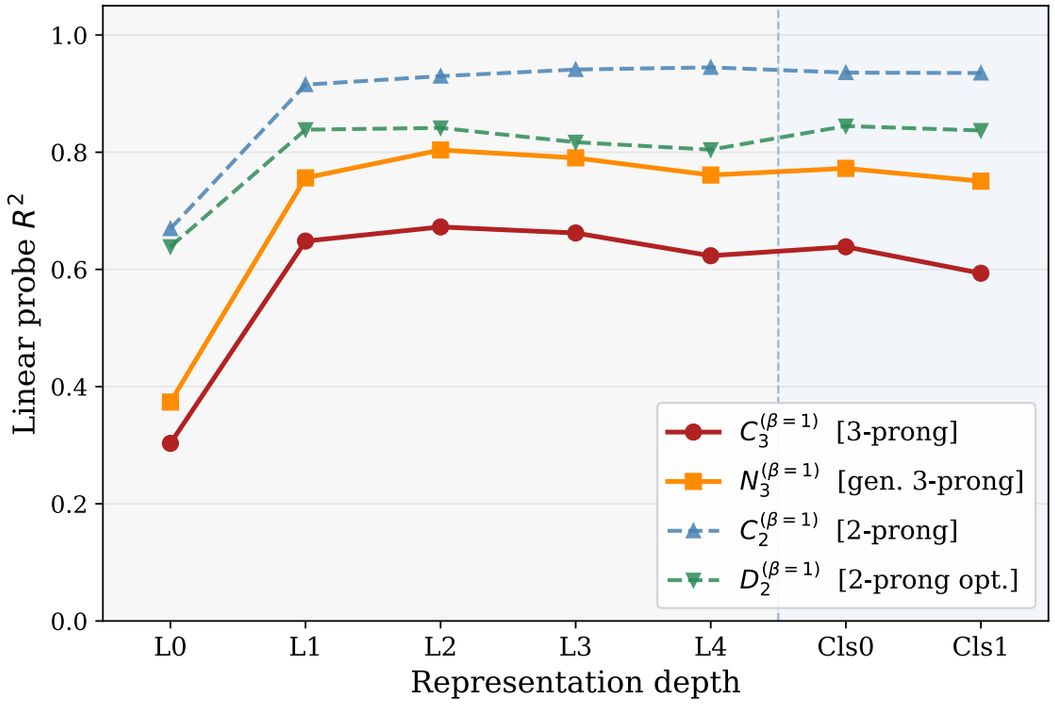
Which Layer Learning What Physics?

Observable	L0	L1	L2	L3	L4	Cls0	Cls1	Peak R^2	Peak layer
2-prong energy correlators:									
$C_1^{(\beta=1)}$	0.856	0.986	0.989	0.988	0.989	0.987	0.986	0.989	L2
$C_2^{(\beta=1)}$	0.670	0.915	0.930	0.941	0.945	0.936	0.935	0.945	L4
$D_2^{(\beta=1)}$	0.639	0.839	0.842	0.817	0.804	0.845	0.837	0.845	Cls0
3-prong energy correlators:									
$C_3^{(\beta=1)}$	0.303	0.648	0.672	0.662	0.623	0.639	0.594	0.672	L2
$N_3^{(\beta=1)}$	0.374	0.756	0.804	0.790	0.761	0.772	0.751	0.804	L2
N-subjettiness reference:									
τ_{21}	0.189	0.597	0.663	0.655	0.672	0.641	0.646	0.672	L4
τ_{32}	0.310	0.465	0.544	0.574	0.581	0.578	0.571	0.581	L4
Jet mass	0.862	0.969	0.972	0.971	0.971	0.971	0.969	0.972	L2

3-prong energy correlators



3-prong vs 2-prong energy correlators



2-prong features are given more weightage over 3-prong structure.

Energy correlator variables are preferred over N-subjettiness.

Summary

- ☑ **We approached the question of interpretability by trying to understand what is being learned and how the hidden layers are performing the downstream task.**

- ☑ **Correlation studies show that different layers of a jet tagger learns different Lund-Jet plane features.**

- ☑ **There aren't any best explainability method as of now. Depends on downstream task.**

- ☑ **The takeaway message : if we have a point cloud data representation in HEP, we can do explainable AI on them and extend the methods discussed here.**

- ☑ **Mechanistic interpretability methods allows us to understand what's going on inside a deep network :**
  If some parts are found to be irrelevant, can be trimmed off for faster forward computation.
  It will be great to develop these methods for FastML.

Summary

- ☑ **We approached the question of interpretability by trying to understand what is being learned and how the hidden layers are performing the downstream task.**

- ☑ **Correlation studies show that different layers of a jet tagger learns different Lund-Jet plane features.**

- ☑ **There aren't any best explainability method as of now. Depends on downstream task.**

- ☑ **The takeaway message : if we have a point cloud data representation in HEP, we can do explainable AI on them and extend the methods discussed here.**

- ☑ **Mechanistic interpretability methods allows us to understand what's going on inside a deep network :**
  If some parts are found to be irrelevant, can be trimmed off for faster forward computation.
  It will be great to develop these methods for FastML.

THANK YOU!!

Backup

Path Head Tests

Table 3: Forward path effects (representational deltas) within the candidate circuit, with 95% bootstrap confidence intervals over 500 bootstrap resamples on 2,000 test jets. The three natural groupings (primary source → Layer 1, secondary source → Layer 1, Layer 1 → readout) are pairwise non-overlapping, establishing a quantitative ordering of the information flow.

Path	Mean Δ	95% CI lower	95% CI upper
L0H1 → L1H1	1.454	1.417	1.494
L0H1 → L1H3	1.404	1.360	1.443
L0H1 → L3H3	1.351	1.318	1.383
L0H1 → L1H0	1.189	1.160	1.217
L0H2 → L1H0	0.900	0.881	0.919
L0H2 → L1H1	0.858	0.839	0.876
L0H2 → L3H3	0.855	0.838	0.872
L0H2 → L1H3	0.850	0.833	0.867
L1H3 → L3H3	0.754	0.735	0.771
L1H0 → L3H3	0.752	0.736	0.769
L1H1 → L3H3	0.683	0.668	0.698

Table 4: Partial minimality and sufficiency test. Heads are added to the circuit in order of decreasing recovery score. At each step, all heads outside the current circuit are masked and the test AUC is evaluated on the full 50,000-jet test set. Bootstrap 95% confidence intervals are computed with 1,000 resamples. The six-head circuit recovers 97.3% of the full model AUC of 0.9793.

Step	Head	DE	AUC	95% CI	% of full model
1	L0H1	+4.25	0.8676	[0.8641, 0.8710]	88.6%
2	L3H3	+0.71	0.8921	[0.8888, 0.8953]	91.1%
3	L1H0	−1.13	0.9054	[0.9026, 0.9079]	92.5%
4	L0H2	−1.92	0.9089	[0.9062, 0.9114]	92.8%
5	L1H3	−2.48	0.9469	[0.9450, 0.9488]	96.7%
6	L1H1	−2.95	0.9527	[0.9510, 0.9544]	97.3%

Path Head Tests

Table 5: Direct effects of the six circuit heads under two on-manifold corruption strategies: within batch particle replacement (the primary strategy used throughout this work) and whole jet permutation. The qualitative sign pattern, viz., positive for the primary source L0H1 and the readout L3H3, negative for the secondary source L0H2 and all three relay heads, is fully consistent across the two strategies. Magnitudes are smaller under jet permutation than under replacement, which we attribute to the smaller absolute denominator $\mathrm{LD}_{\mathrm{clean}} - \mathrm{LD}_{\mathrm{corrupt}}$ achieved by the milder corruption.

Head	Role	Replacement	Jet permutation
L0H1	Primary source	+4.25	+3.78
L0H2	Secondary source	−1.92	−1.72
L1H0	Relay	−1.13	−0.67
L1H1	Relay	−2.95	−2.51
L1H3	Relay	−2.48	−2.17
L3H3	Readout	+0.71	+0.66

Table 6: Cosine similarity between the mean jet-level output representations of each circuit head pair, computed over 3,000 test jets. Two clusters emerge: a Layer 0 cluster {L0H1, L0H2} with cosine similarity 0.92, and a Layer 1 cluster {L1H0, L1H1, L1H3} with pairwise similarities in $[0.81, 0.95]$. The L3H3 readout has low similarity to all other heads.

	L0H1	L0H2	L1H0	L1H1	L1H3	L3H3
L0H1	1.000	0.922	0.027	0.144	0.069	0.052
L0H2	0.922	1.000	0.093	0.134	0.112	0.051
L1H0	0.027	0.093	1.000	0.812	0.952	0.451
L1H1	0.144	0.134	0.812	1.000	0.901	0.476
L1H3	0.069	0.112	0.952	0.901	1.000	0.468
L3H3	0.052	0.051	0.451	0.476	0.468	1.000

Pairwise Zero Ablation Test

Pair	Joint	Sum	Super-additivity S
Within Layer 0 *(source cluster):*			
{L0H1, L0H2}	0.732	2.244	−1.512
Within Layer 1 *(relay cluster):*			
{L1H0, L1H3}	0.738	2.236	−1.498
{L1H1, L1H3}	0.638	2.136	−1.498
{L1H0, L1H1}	0.638	1.375	−0.738
L0H1 paired with each remaining circuit head:			
{L0H1, L1H3}	2.570	3.011	−0.441
{L0H1, L3H3}	1.725	2.057	−0.332
{L0H1, L1H1}	1.841	2.150	−0.309
{L0H1, L1H0}	1.943	2.250	−0.307
Cross-cluster (Layer 0 × *Layer* 1 *off-cluster):*			
{L0H2, L1H0}	1.293	1.469	−0.177
{L0H2, L1H1}	1.328	1.369	−0.041
{L1H3, L0H2}	2.250	2.230	+0.020
With readout head L3H3:			
{L1H0, L3H3}	1.151	1.282	−0.131
{L0H2, L3H3}	1.227	1.276	−0.049
{L1H3, L3H3}	2.019	2.043	−0.024
{L1H1, L3H3}	1.184	1.182	+0.002

Mechanistic Interpretability Dictionary

- **Circuit**: sub-graph of features/heads + weights connecting them that causally reproduces model behavior — the network's analogue of a decay chain
- **Direct effect**: path-patching recovery score for one head, holding others fixed
- **Privileged basis**: architecturally distinguished basis directions — why per-head analysis is meaningful; cf. a mass vs. flavor basis, where the physical basis is picked out by the dynamics
- **Logit lens**: projecting an intermediate layer's residual stream through the final classification head
- **Universality (weak)**: models converge on analogous solutions via shared principles; implementing circuits can vary