Developing universal foundational models for top-quark physics

Abasov E., Dudko L., Iudin E., Markina A.,

Perfilov M., Volkov P., Vorotnikov G., Zaborenko A.

Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University

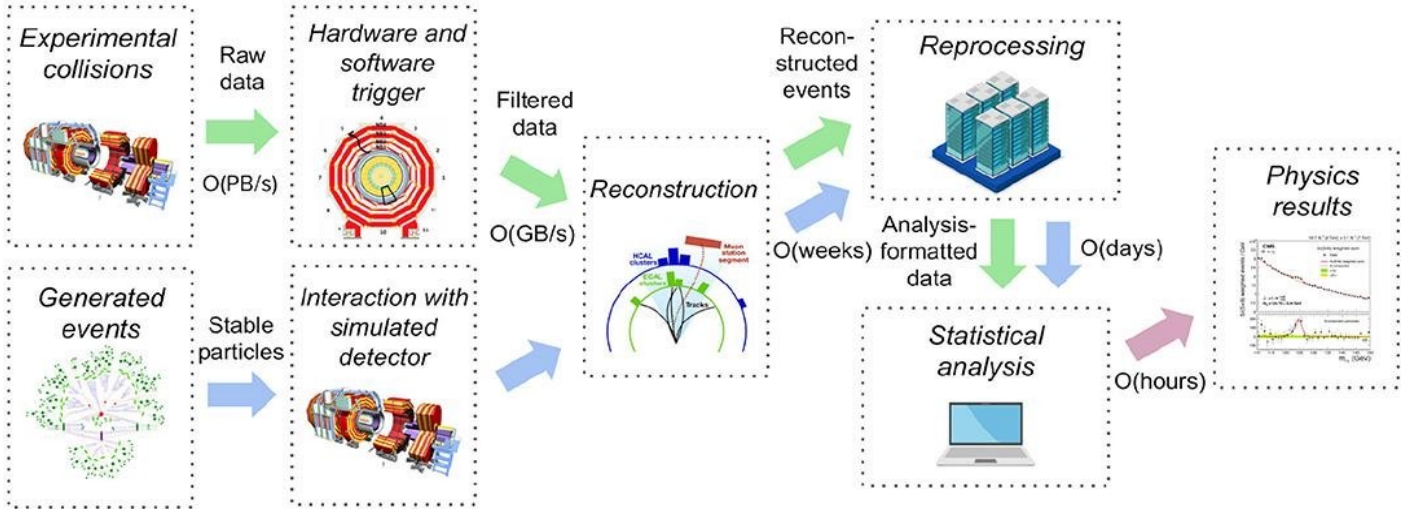
This study was conducted within the scientific program of the National Center for Physics and Mathematics, section #5 «Particle Physics and Cosmology». Stage 2026-2027

Motivation

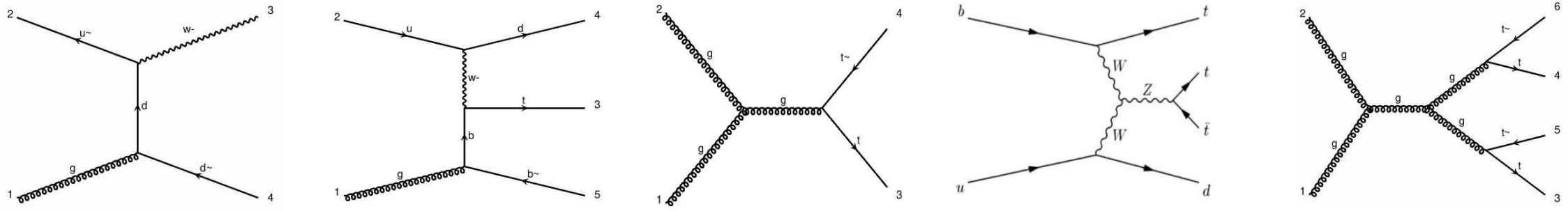
- Modern experiments in High Energy Physics generate enormous amounts of data
- Classical ML models require complex manual customisation for each task

- CV and NLP domains have seen the development of “foundational models”, which are pretrained to infer basic rules of the domain and then finetuned for specific downstream tasks

- Our goal is to apply the “foundational model” approach to HEP

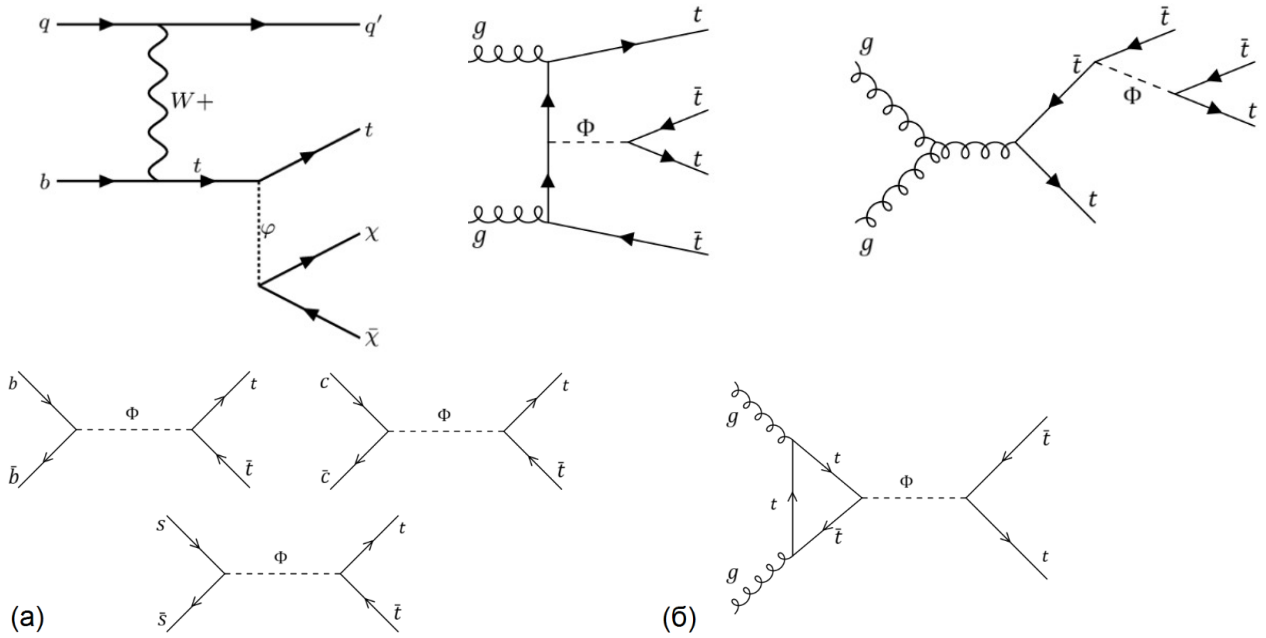


Data



SM processes

- 5 process groups (0, 1, 2, 3, 4 top-quarks) in order to cover wide range of parameters and final states
- Generators: MadGraph5 and CompHEP
- ~8M events in total

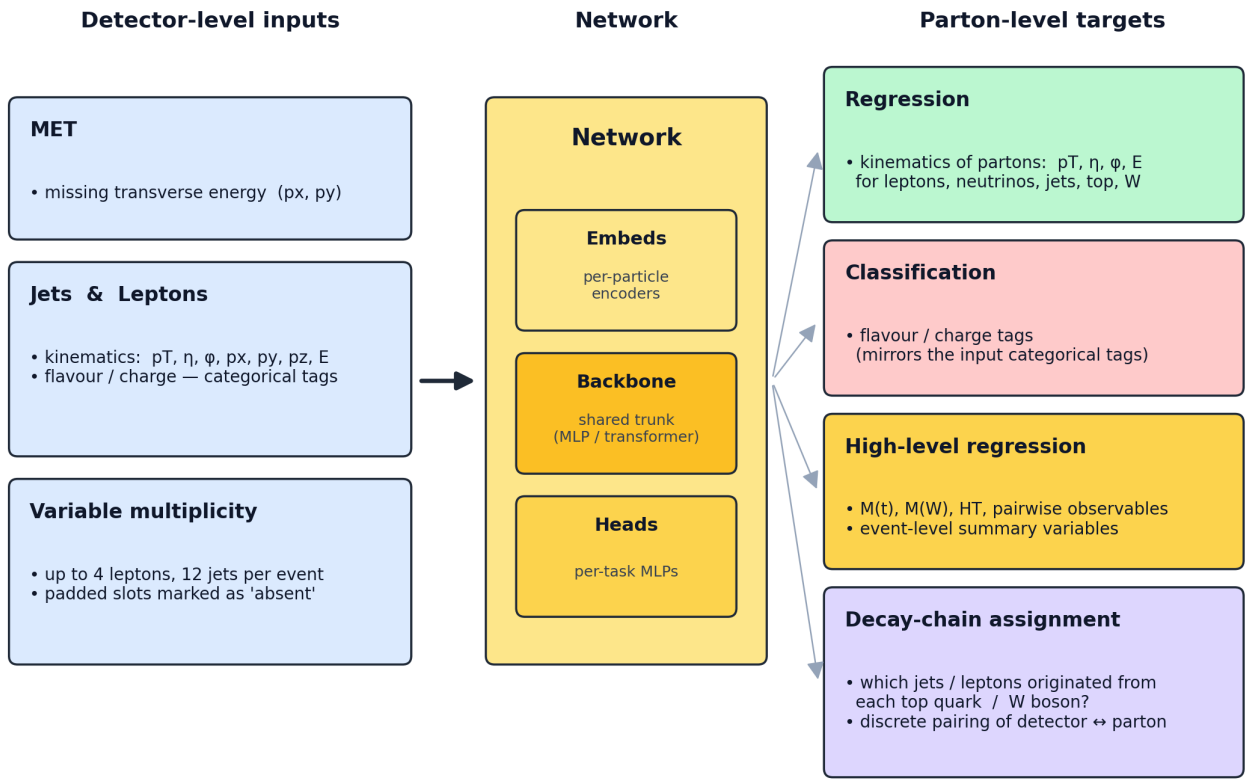


BSM processes

Pre-training task

Pretraining structure:

- Inputs – detector-level variables
  - MET
  - p_T, η, φ, P_x, P_y, P_z, E
  - Flavour/charge categorical tags
- Outputs – properties of parton-level particles
  - p_T, η, φ, E – target variables for the regression task
  - Flavour/charge categorical tags – targets for classification task
  - "High-level" variables (M_tW, H_t, pairwise variables)
  - Decay chain reconstruction (assignment task) – which jets/leptons originated from a given t/W



Pretraining: predict parton-level truth from the detector-level event description (four heads share the same backbone, trained jointly on a multi-task loss).

Models

An event is a set of 4-momenta. We compare three models that respect this to an increasing degree.

1. **MLP** – *no prior*
   - Event → flat 130-d vector [MET | 4 ℓ × 8 | 12 j × 8], zero-padded.
   - Permutations and particle types must be learned from data
   - 4-momenta enter as normalized scalars → Lorentz symmetry broken
2. **DEtection Transformer (DETR)** – *set + decay-tree prior*
   - Event → 17 tokens {MET, ℓ, j}; attention is permutation-equivariant within each type
   - Lorentz symmetry still only learned
3. **Lorentz Geometric Algebra Transformer (LGaTr)** - *Lorentz-equivariant prior*
   - Each token = multivector in Cℓ(1,3)
   - Encoder is Lorentz-equivariant by construction
   - Invariant masses and angles are exact, not learned

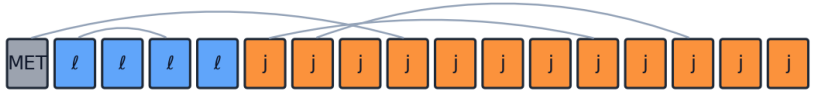
Physics priors built into the architecture

MLP *no symmetry*

flat 130-d vector [MET | 4ℓ×8 | 12j×8]

+ permutation symmetry

DETR *permutation-equivariant*



set of 17 typed tokens {MET} ∪ {ℓ_i} ∪ {j_k}

+ Lorentz symmetry $SO^+(1,3)$

LGaTr *Lorentz-equivariant*

multivectors in Cℓ(1, 3)

$\Phi(\Lambda \cdot x) = \Lambda \cdot \Phi(x)$

grade-4 (1): pseudoscalar, CP-odd
grade-3 (4): 3-body volumes
grade-2 (6): decay planes
grade-1 (4): (E, p_x, p_y, p_z)
grade-0 (1): scalars, masses

exact Lorentz covariance by construction

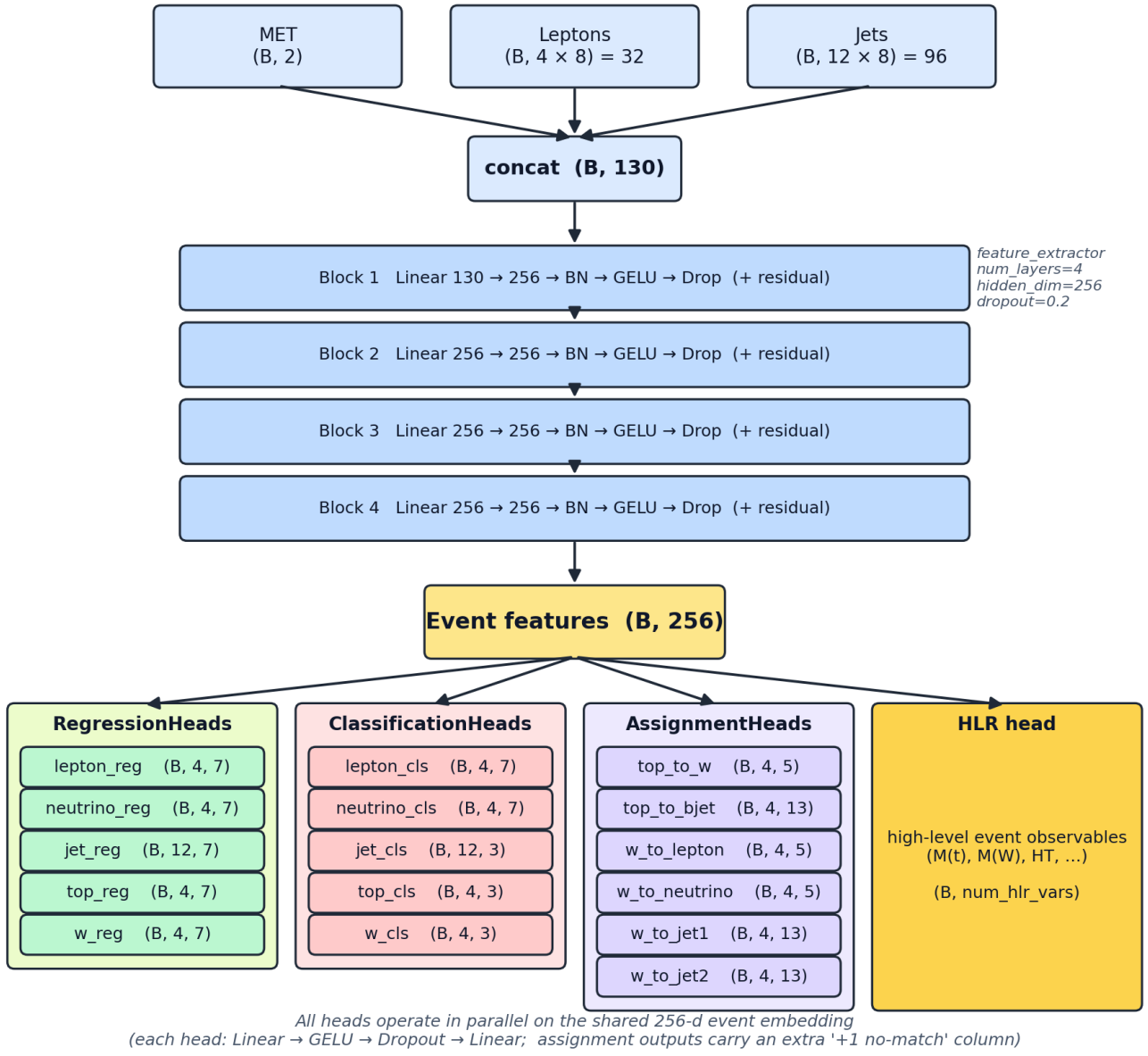
MLP: concepts

Data:

- 130-dimensional flat vector

Architecture:

- Backbone: Simple MLP, 4 blocks with 256 neurons each and skip-connections, GELU activation
- Heads: separate Linear layers for every particle-task type combination
  - Regression – 5 heads (jet, lepton, neutrino, t, W)
  - Classification – 5 heads (same)
  - Assignment – 6 heads ($t \to W, t \to b, W \to l, W \to \nu, W \to j_1, W \to j_2$)
  - HLR – 1 head for all variables
- Losses: MSE for regression and HLR, Cross-entropy (CE) for classification and assignment.



DETR: concepts

Data:

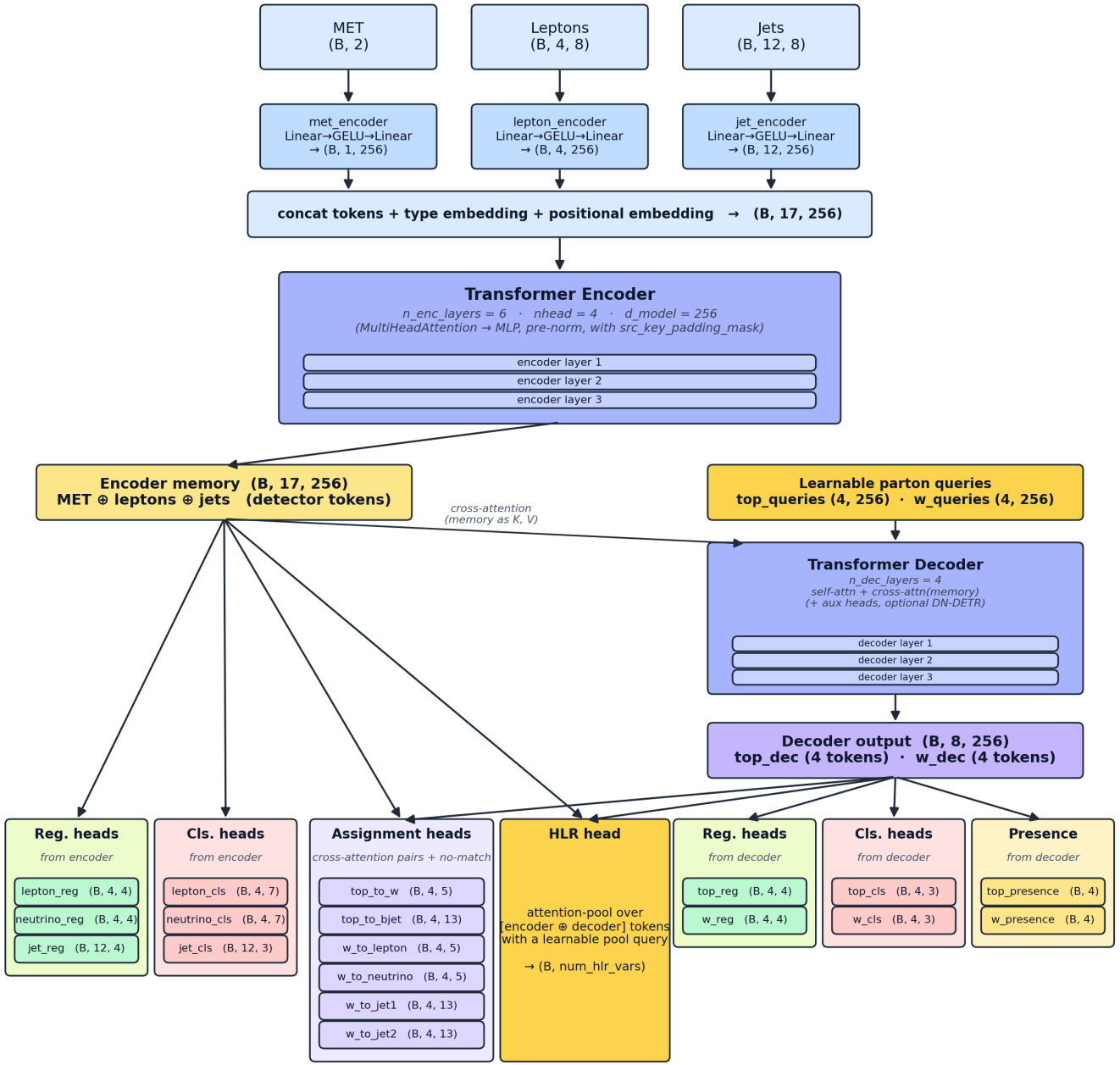
- 5-dimensional token for every particle, 25 tokens in total

Architecture:

- Embedder:
  - MLP $R^5 \to R^D, D = 256$
  - Learnable type and positional embeddings

$$t_k^\ell = \underbrace{h_k^\ell}_{\text{content (MLP)}} + \underbrace{e_{\text{type}}^{[1]}}_{\text{type embedding}} + \underbrace{p_k^\ell}_{\text{positional embedding}}$$

- Backbone: Transformer Encoder-Decoder with Multi-head attention, used configuration is 4 layers with 4 heads for each part
  - Encoder only deals with particles present in the input (jets, leptons, neutrinos (indirectly via MET))
  - Decoder uses learned query tokens and the full decoder output to attend to t/W. These tokens are initialized as $N(0, 0.02)$ and then trained by the model in parallel



DETR-style encoder–decoder. Lepton / neutrino / jet heads read encoder memory; top / W / presence heads read decoder output. Assignment and HLR heads consume both encoder memory and decoder output.

DETR: concepts

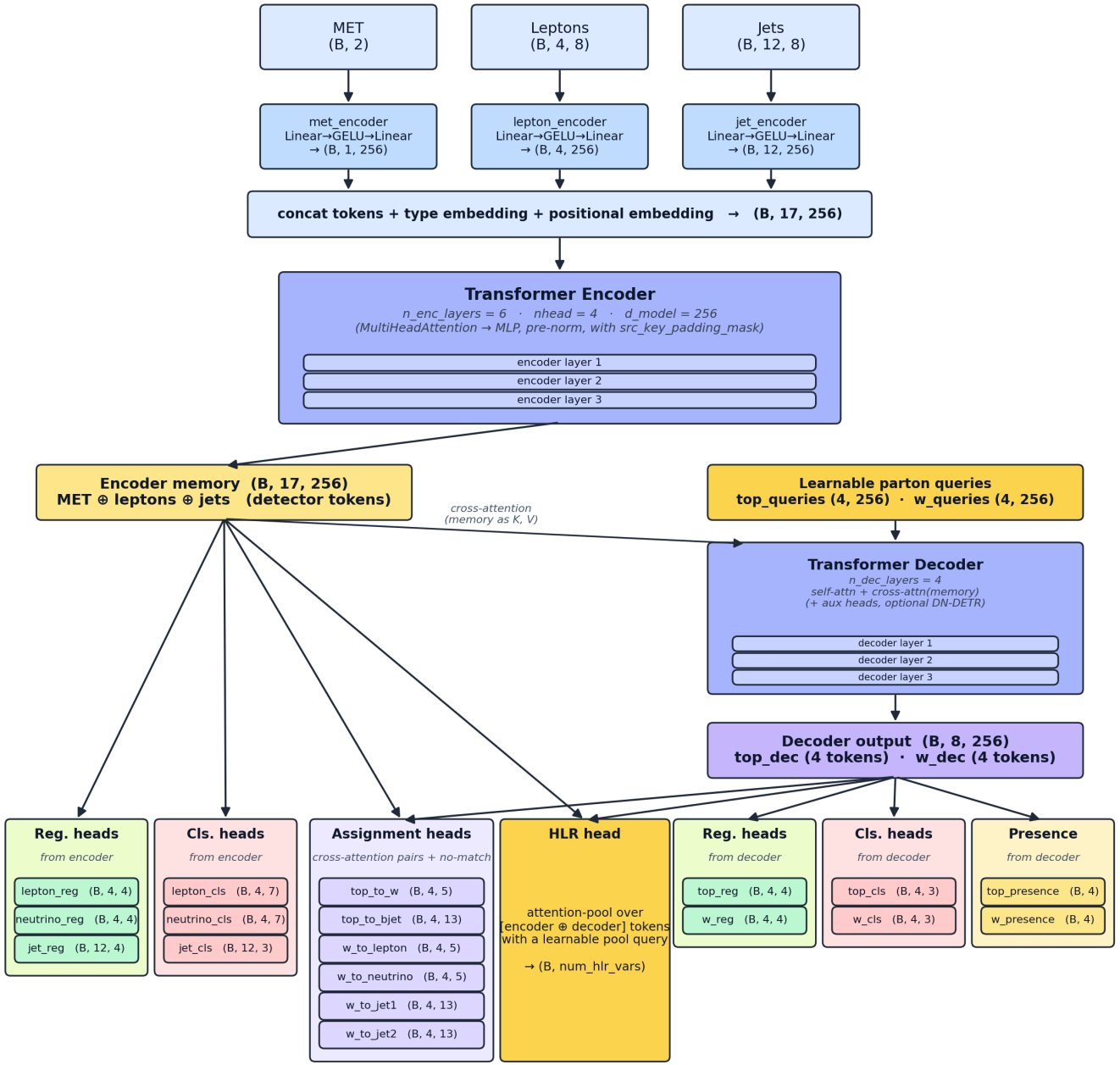
- Heads: Per-Particle heads; each token in the backbone output carries information about the relevant particle, can ignore others
  - Regression and classification – 1 for each token
  - HLR – Attention pooling with separate HLR token

$$\alpha_i = \frac{\exp(\boldsymbol{q}_{\mathrm{HLR}}^{\top} \boldsymbol{z}_i / \sqrt{D})}{\sum_{j \notin \mathrm{pad}} \exp(\boldsymbol{q}_{\mathrm{HLR}}^{\top} \boldsymbol{z}_j / \sqrt{D})}, \qquad \boldsymbol{h}_{\mathrm{HLR}} = \sum_i \alpha_i \, \boldsymbol{z}_i$$

  - Assignment – Bilinear assignment heads (key – parent, query – each child)
  - Presence – new type of heads analogous to classification which only predicts if the current slot has a particle

Outputs of the decoder are *unordered*, in order to match them with targets, set matching is used

$$C_{ij}^{W} = \underbrace{\mathrm{FL}(\hat{y}_{W_i}^{\mathrm{cls}}, y_{W_j}^{\mathrm{cls}})}_{\text{classification}} + \underbrace{R(\hat{y}_{W_i}^{\mathrm{reg}}, y_{W_j}^{\mathrm{reg}}) \cdot m_j}_{\text{regression}} + \underbrace{\sum_{k \in \mathcal{A}_W} \mathrm{FL}(\hat{a}_i^k, a_j^k)}_{\text{assignment}}$$



DETR-style encoder–decoder. Lepton / neutrino / jet heads read encoder memory; top / W / presence heads read decoder output. Assignment and HLR heads consume both encoder memory and decoder output.

DETR: concepts

- Losses:
  - Regression – MSE for p_t, η, E, Circular loss for ϕ (naturally periodic)

$$\mathcal{L}_{\text{reg}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \underbrace{\frac{1}{|D \setminus \{2\}|} \sum_{d \neq 2} (\hat{y}_d - y_d)^2}_{\text{MSE on } (p_T, \eta, E)} + \underbrace{1 - \cos\big((\hat{y}_2 - y_2) \cdot \sigma_\phi\big)}_{\text{circular } \phi \text{ loss}}$$

  - Classification – Focal loss to deal with heavy class imbalance

$$\text{FL}(p_t) = -(1 - p_t)^\gamma \log(p_t), \qquad \gamma = 2$$

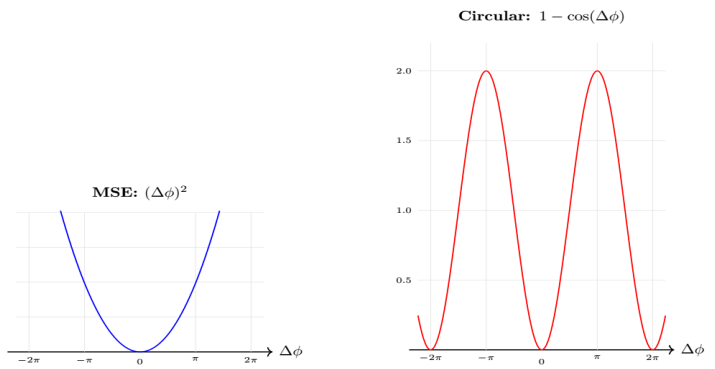
  - HLR – MSE, same as for baseline
  - Assignment – Focal loss
  - Denoising – since the changes introduced complexity, a simple additional task of removing Gaussian noise from inputs is added to help during first training steps
  - "Auxiliary" or deep supervision – all losses computed on every intermediate decoder layer

$$\mathcal{L}_{\text{aux}} = \sum_{l=1}^{L-1} \underbrace{0.5^{(L-1-l)}}_{\text{exponential decay}} \cdot \mathcal{L}_{\text{total}}^{(l)}$$

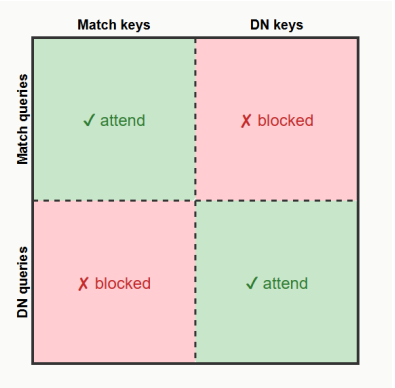
$$\mathcal{L}_{\text{total}} = \underbrace{1.0}_{\text{reg}} \cdot \mathcal{L}_{\text{reg}} + \underbrace{1.0}_{\text{cls}} \cdot \mathcal{L}_{\text{cls}} + \underbrace{2.0}_{\text{assign}} \cdot \mathcal{L}_{\text{assign}} + \underbrace{5.0}_{\text{hlr}} \cdot \mathcal{L}_{\text{hlr}}$$

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{total}}}_{\text{main loss}} + \underbrace{\mathcal{L}_{\text{aux}}}_{\text{deep supervision}} + \underbrace{\mathcal{L}_{\text{dn}}}_{\text{denoising}}$$

Full DETR pipeline

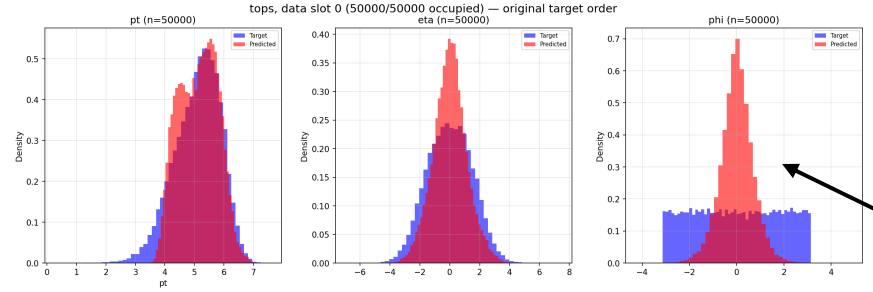


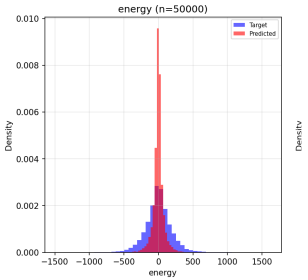
MSE and circular losses



Denoising attention masking

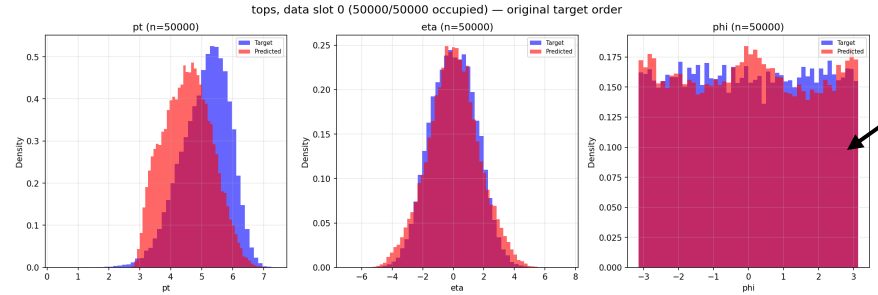
MLP/DETR: Regression

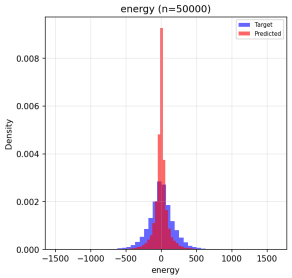




MLP

Circular ϕ loss impact

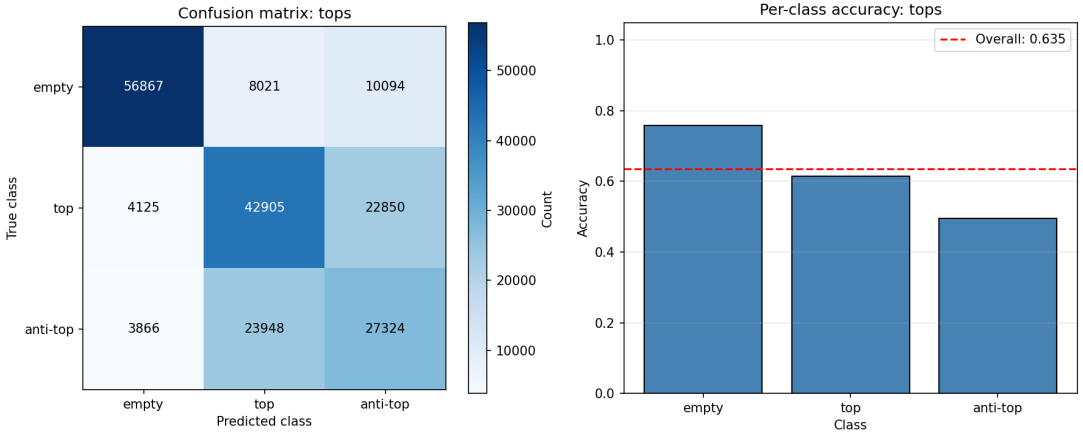




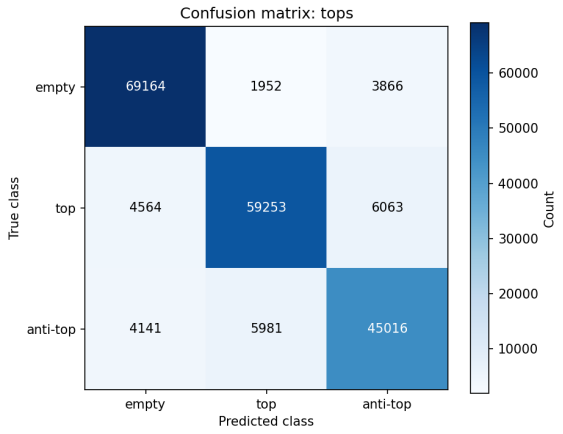
DETR

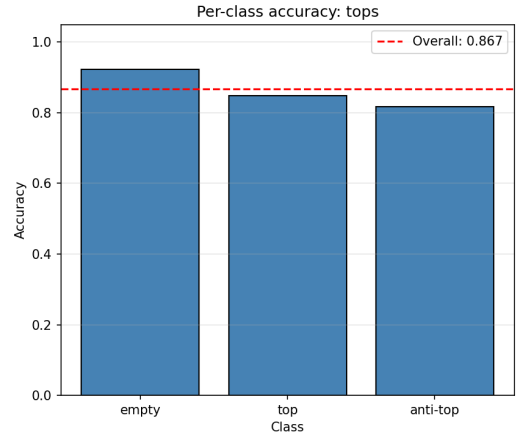
MLP/DETR: Classification

MLP

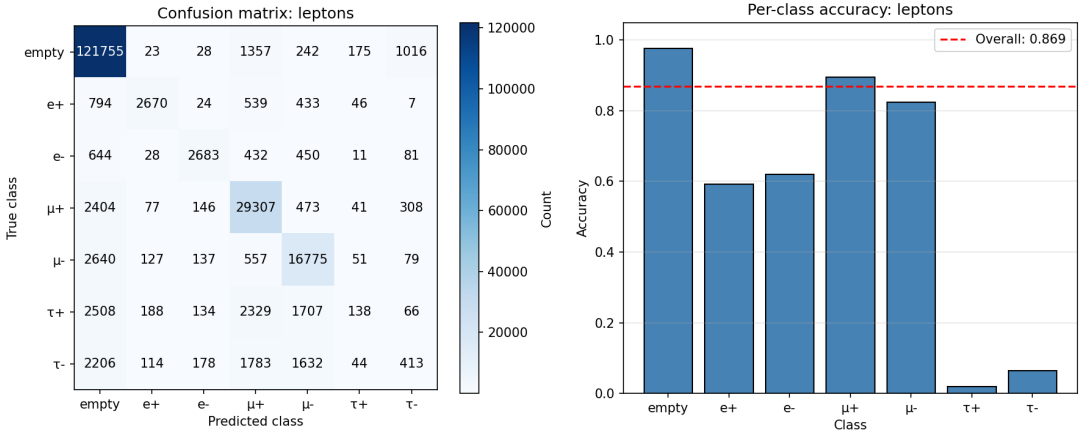


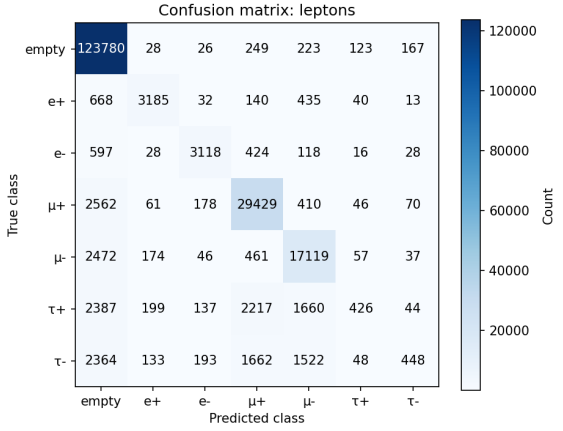
DETR

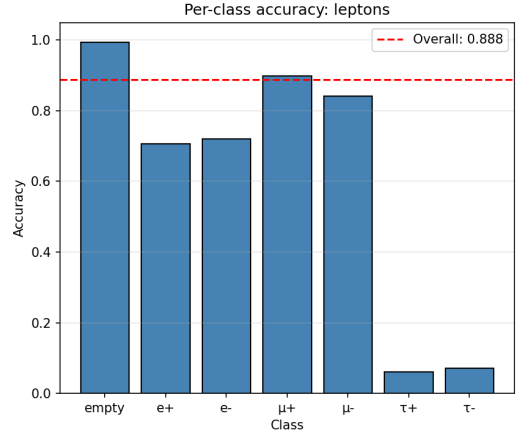




DETR performs much better on complex tasks

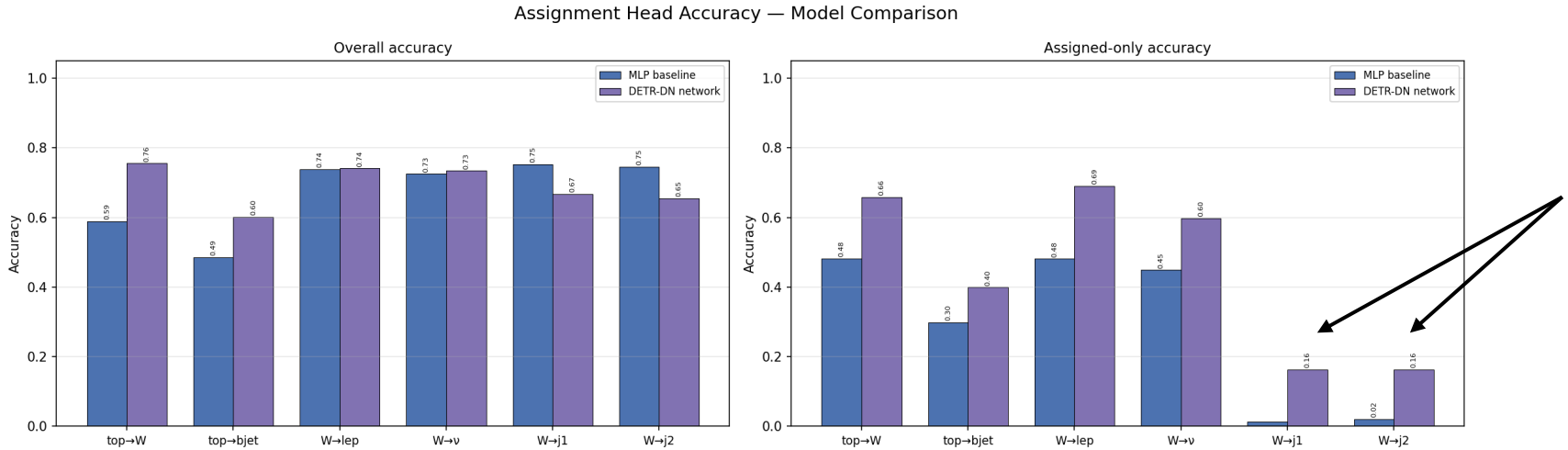




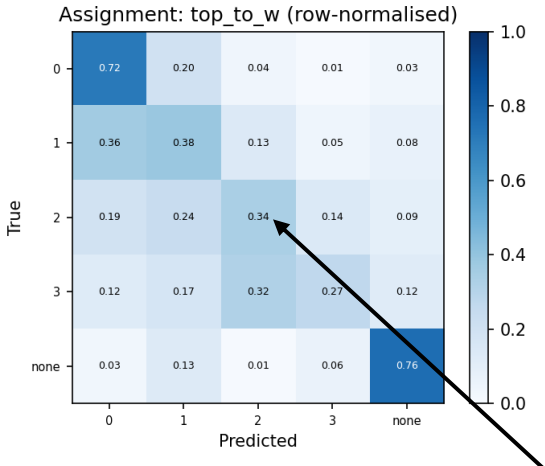


Similar performance on simple tasks

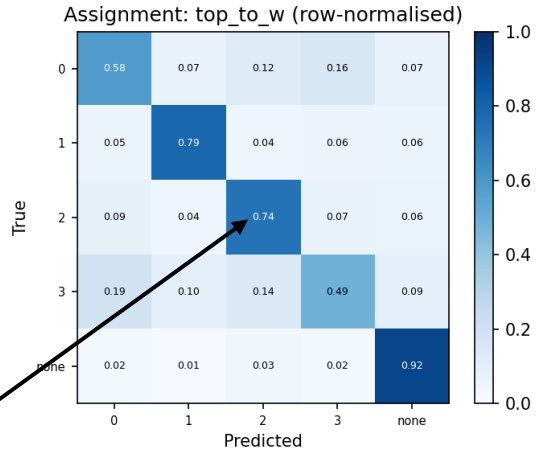
MLP/DETR: Assignment



Very few $W \to \bar{j}j$ events in data → poor performance



MLP



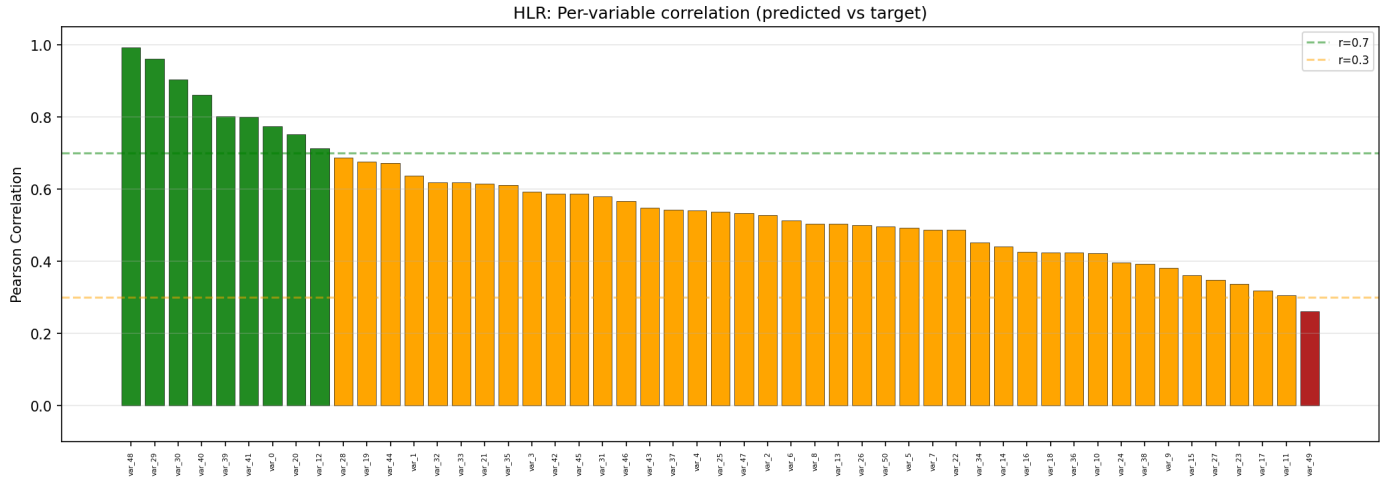
DETR

Much better separation of rarer classes

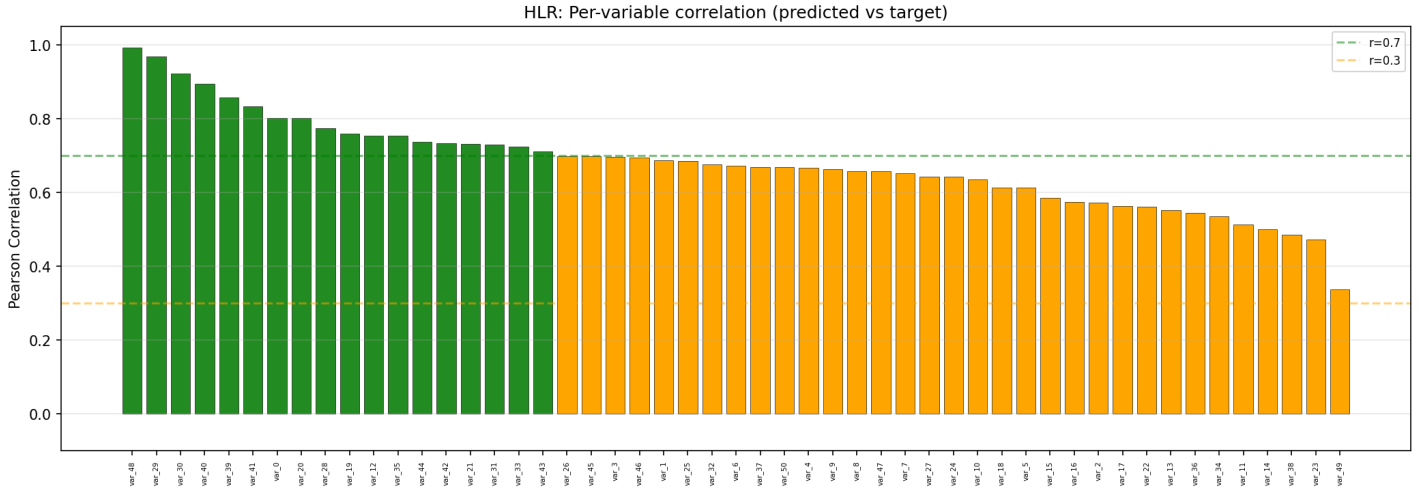
MLP/DETR: HLR

- Comparison metric – per-variable correlation between target and predicted distributions
- DETR shows better reconstruction of all variables

MLP

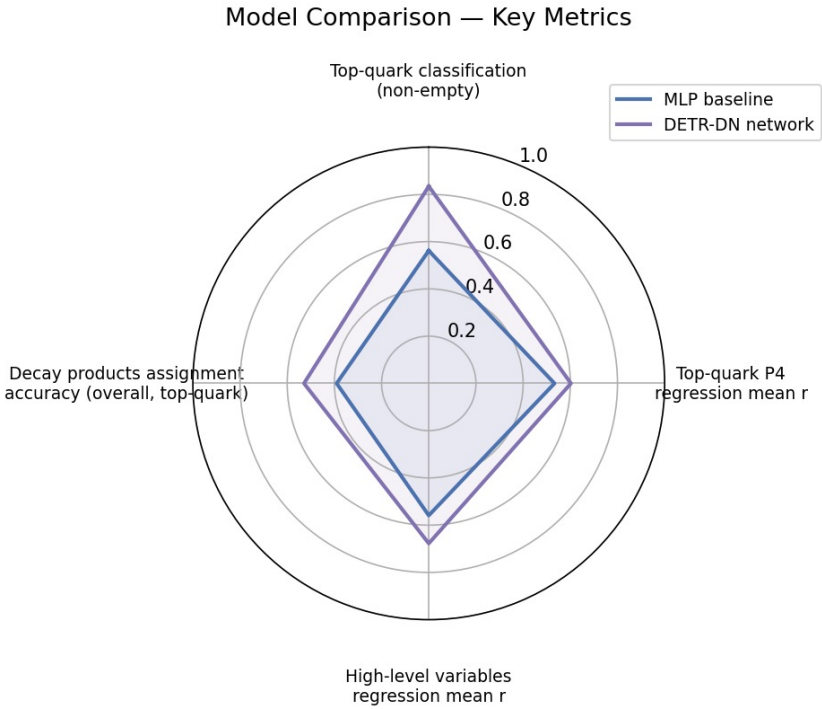


DETR



MLP/DETR: verdict

- DETR outperforms MLP in all 4 tasks, with the most significant improvements in assignment and classification
- Circular loss on ϕ yields a more bounded and physical predictions
- Poor $W \to \bar{j}j$ assignment is caused by lepton mode dominating the training sample



Geometric algebra

Standard HEP kinematics uses three separate operations: scalar product pq, antisymmetric tensor $p^\mu q^\nu - p^\nu q^\mu$, and Levi-Civita contraction $\epsilon_{\mu\nu\rho\theta}$.Geometric algebra (GA) unifies all three into one associative, invertible product.

- GA operates on multivectors
- Multivectors can be decomposed into grades (analogous to tensor rank)
- Main operation between multivectors – geometric product, which can be decomposed into grade-decreasing (dot) and grade-increasing (wedge) products

$$pq = p \cdot q + p \wedge q$$

$$p \cdot q = \frac{1}{2}(pq + qp), \qquad p \wedge q = \frac{1}{2}(pq - qp)$$

Dot product – lower grade

Wedge product – higher grade

Example in HEP: $W \to l\nu$ decay:

$$p_\ell \cdot p_\nu = \frac{m_W^2}{2} \quad \text{(invariant mass)}, \qquad p_\ell \wedge p_\nu = \text{6-component bivector (decay plane)}$$

GA can be defined over any real vector space with a bilinear metric. In HEP Cl(1,3) is used (Minkowski space)

- Basis vectors γ_i

$$\gamma_\mu\gamma_\nu + \gamma_\nu\gamma_\mu = 2\,\eta_{\mu\nu}, \qquad \eta = \text{diag}(+1,-1,-1,-1) \qquad \gamma_0^2 = +1, \qquad \gamma_i^2 = -1 \quad (i = 1,2,3) \qquad I = \gamma_0\gamma_1\gamma_2\gamma_3, \qquad I^2 = -1$$

Geometric algebra

- Every grade can be universally transformed with a single multivector – rotor – using a sandwich product

$$X \longmapsto R\,X\,\widetilde{R}, \qquad R\widetilde{R} = 1$$

$$R = \exp\left(\frac{B}{2}\right), \qquad B \in \langle \mathrm{Cl}(1,3)\rangle_2$$

$$B^2 = -1 \;\Rightarrow\; \text{spatial rotation}, \qquad B^2 = +1 \;\Rightarrow\; \text{Lorentz boost}$$

Examples:

2D rotation

$$R = e^{\theta \mathbf{e}_1\mathbf{e}_2/2} = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}\,\mathbf{e}_1\mathbf{e}_2$$

$$\mathbf{v}' = R\,\mathbf{v}\,\widetilde{R}$$

Lorentz boost

$$R_{\text{boost}} = e^{\alpha\,\gamma_3\gamma_0/2} = \cosh\frac{\alpha}{2} + \sinh\frac{\alpha}{2}\,\gamma_3\gamma_0$$

Since every grade transforms with the same rotor – neural network using GA can be equivariant

$$\Phi\big(\Lambda \cdot x\big) = \Lambda \cdot \Phi(x) \longleftrightarrow \Phi\big(R\,X\,\widetilde{R}\big) = R\,\Phi(X)\,\widetilde{R}$$

Geometric algebra in HEP

- Event grade decomposition

Grade k	$\binom{4}{k}$	Name	HEP meaning
0	1	scalar	masses m_i^2, Mandelstam $p_i \cdot p_j$
1	4	vector	4-momentum p^μ
2	6	bivector	decay planes $p_i \wedge p_j$, Lorentz generators
3	4	trivector	3-body oriented volumes $p_i \wedge p_j \wedge p_k$
4	1	pseudoscalar	CP-odd sign, chirality

- **Cayley–Menger collapse -** Any Lorentz-invariant scalar built from a higher-grade blade collapses to a polynomial in $\{p_i \cdot p_j, m_i^2\}$ – GA does not generate new Lorentz-invariant scalars beyond what you already have
- Main benefit of GA is not in new inputs (the only exception is CP-odd pseudoscalar), but in forcing Lorentz-equivariance by construction
- Additional example – top-quark spin correlations
  - Standard approach: (1) boost 4-momenta to $\bar{t}t$ rest frame, (2) boost to respective top/antitop rest frame, (3) project lepton directions onto spin-quantization axis – complex operation needs to be learned
  - GA approach – one operation

$$B_t = p_b \wedge p_\ell \wedge p_\nu, \qquad B_{\bar{t}} = p_{\bar{b}} \wedge p_{\bar{\ell}} \wedge p_{\bar{\nu}}$$

$$\cos \Delta\phi_{\text{spin}} \sim \frac{\langle B_t \widetilde{B}_{\bar{t}} \rangle_0}{|B_t|\,|B_{\bar{t}}|}$$

LGaTr: concepts

Data: each detector token → multivector in Cl(1,3)

- Grade-1 channel: raw (E, p_x, p_y, p_z) / E_scale, E_scale = 100 GeV
- Grade-0 channels: b-tag, lepton flavour, particle type (5 scalar channels)
- Potentially grade-2 pairwise features

Architecture: GATr encoder blocks:

- Equivariant linear maps: mix multivector channels, preserve grade structure

$$Y^{(c')} = \sum_{c=1}^{C} W_{c'c}\, X^{(c)}, \qquad W_{c'c} \in \mathbb{R}$$

- Geometric product layer: channel-wise pairwise products

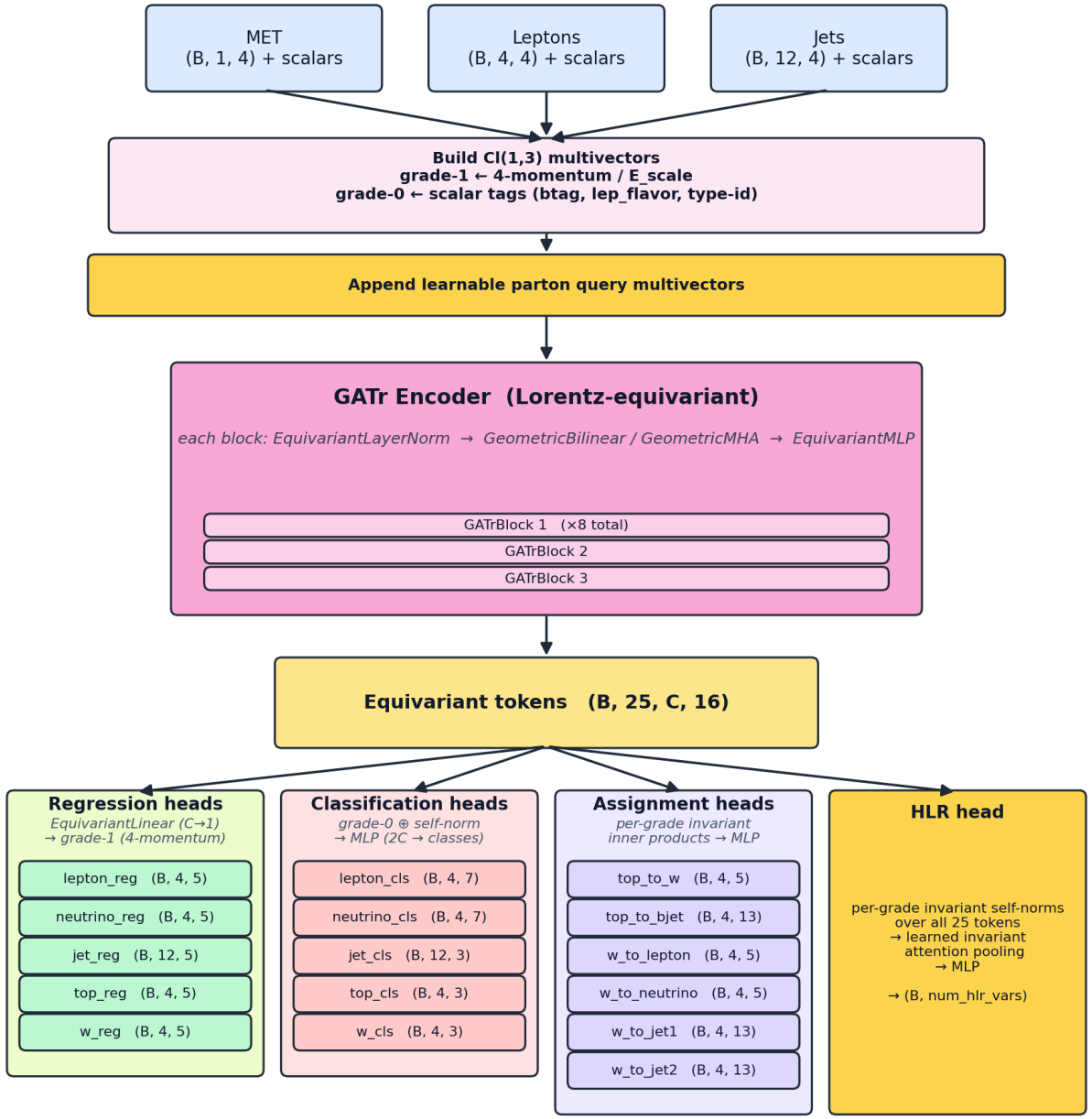
$$Z^{(c_1,c_2)} = X^{(c_1)}\, X^{(c_2)}, \qquad \langle Z \rangle_k = \sum_{\substack{r,s \geq 0 \\ |r-s| \leq k \leq r+s}} \langle X^{(c_1)} \rangle_r \langle X^{(c_2)} \rangle_s$$

- Gated nonlinearity on scalar norm

$$X \longmapsto \sigma\big(\langle X \widetilde{X} \rangle_0\big) \cdot X$$

- Attention: Q, K, V are multivectors

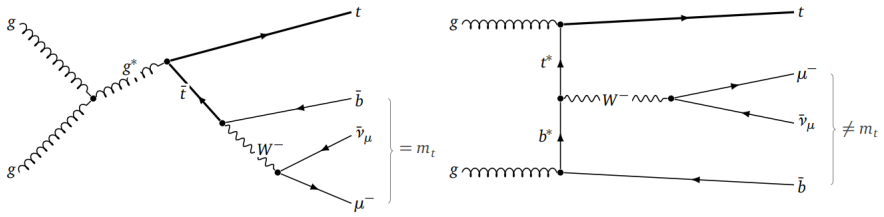
$$\mathrm{score}_{ij} = \langle Q_i\, \widetilde{K}_j \rangle_0, \qquad \mathrm{Attn}(Q,K,V)_i = \sum_j \mathrm{softmax}\left(\frac{\mathrm{score}_{ij}}{\sqrt{d}}\right) V_j$$



Single Lorentz-equivariant stack over detector + parton-query tokens
all heads read Lorentz invariants of the multivector tokens, so outputs transform covariantly with boosts.

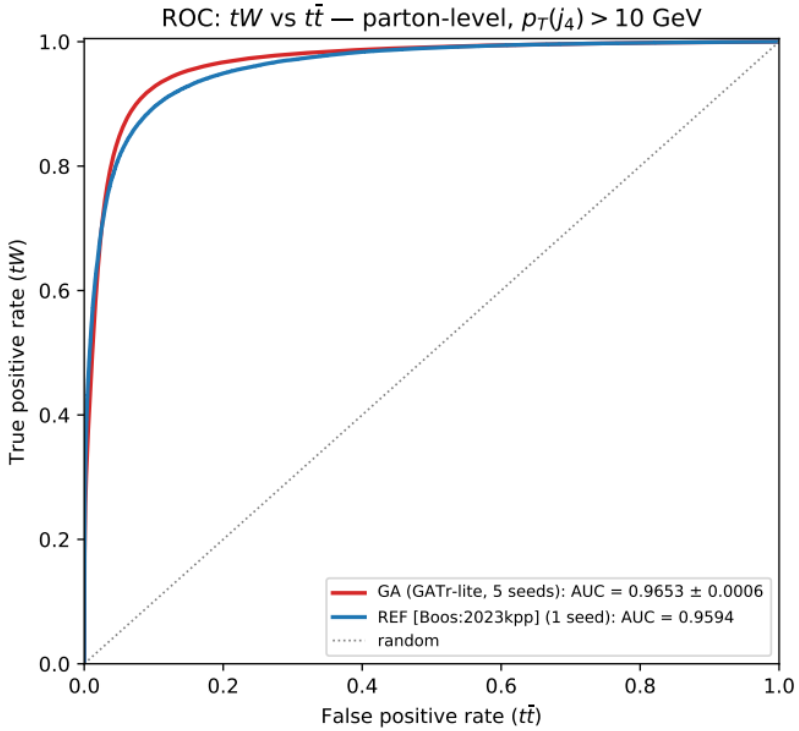
LGaTr proof-of-concept

- LGaTr was tested on a simpler task - $\bar{t}t$ / tWb separation



(a) double-resonant: $t\bar{t}$ (b) single-resonant: tW

- Event multivector
  - Grade-0 – one-hot encoding of the object type $(\mu^-, \overline{\nu_\mu}, u, \bar{d}, b\ \bar{b})$
  - Grade-1 – 4-momentum
  - Grade-2/3 (new) – wedge products of 4-momenta,
- Baseline – MLP on hand-crafted scalar features
- Results show LGaTr beating the baseline with a significant margin



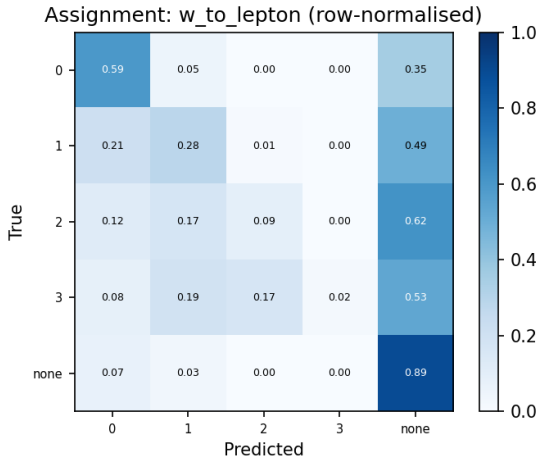
Conclusion

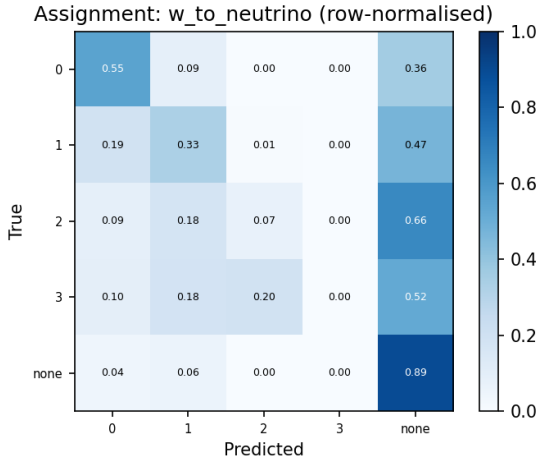
- Foundational models provide a unified framework for various top-physics analyses
- Neural-networks-based solutions are capable of complex event reconstruction
- DETR-like architecture significantly outperforms MLP across all 4 tasks; largest gains on complex tasks (assignment, top-quark classification)
- Geometric algebra provides further physics-based improvements by introducing Lorentz-covariance and CP-odd observables. More detailed study is published in the preprint 2605.15910
- LGaTr shows potential on a proof-of-concept $\bar{t}t$ / tWb separation task

Backup

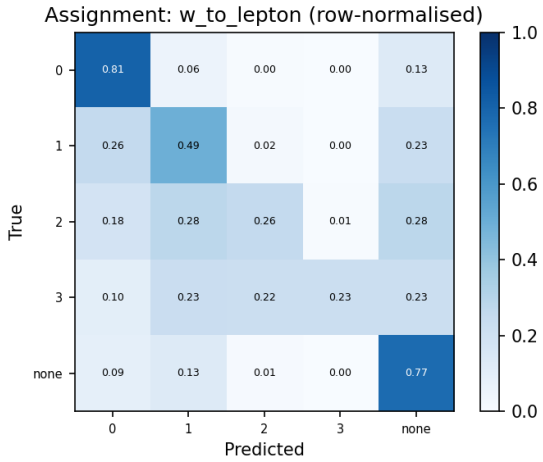
Assignment quality plots

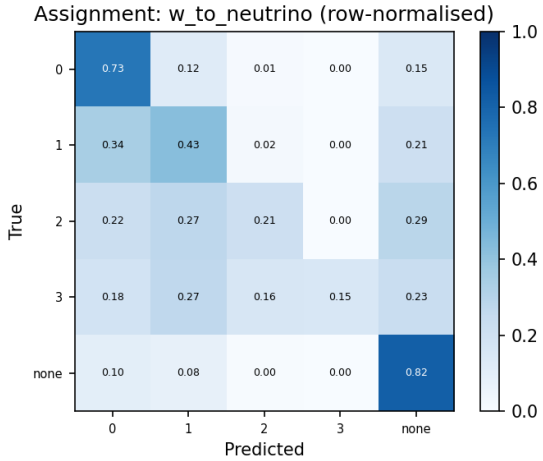
MLP





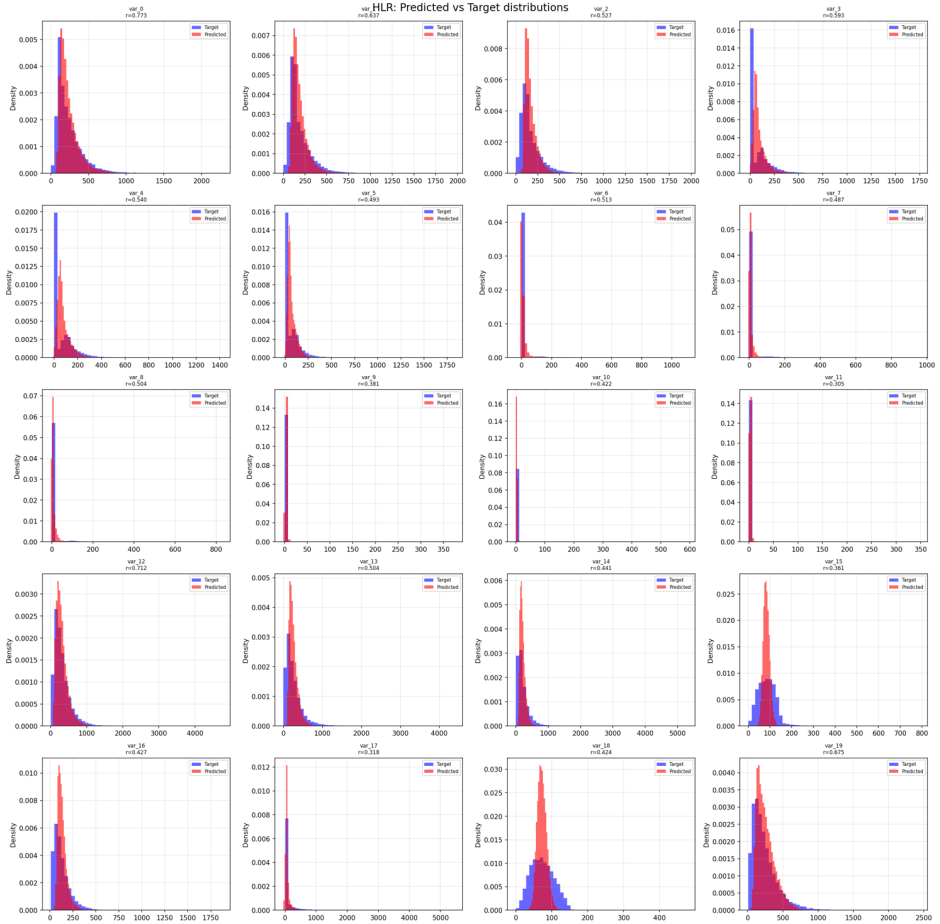
DETR



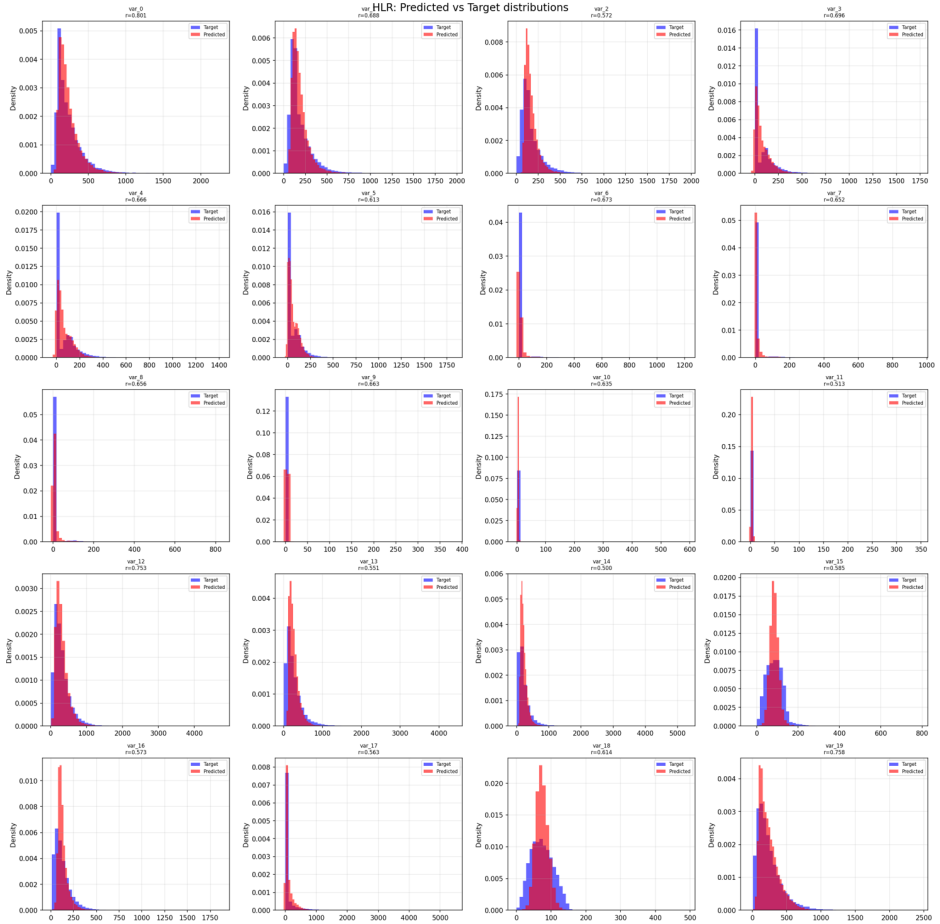


HLR quality plots

MLP



DETR



GA variables

#	Variable	Source	Grade	Frame	Inv. subgroup
1	m_i^2	[4]	0	inv.	Y
2	$\hat{s}$	[4,5]	0	inv.	Y
3	$\hat{t}$	[4]	0	inv.	Y
4	m_{ij}	[4,5]	0	inv.	Y
5	p_T^i	[5]	1	beam	H_{LHC}
6	η_i, y_i	[5]	1	beam	$\mathrm{SO}(1,1)_z$
7	$\phi_i{}^{\mathrm{a}}$	[5]	1	beam	$\mathrm{SO}(2)_\phi$
8	H_T, S_T	[4,5]	1	beam	H_{LHC}
9	$\cos\theta_\ell^*$	[4,20]	0	hel.	Y
10	$\Delta\phi_{ij}$, ΔR_{ij}	[5,11]	1→0	beam	H_{LHC}
11	$C_{ij}^{\mathrm{Bernreuther}}$	[21,22]	0/2	inv.	raw: N; inv. proj.: Y
12	E_T^{miss}	[5]	$1_{(\text{masked } p_z)}$	beam	H_{LHC}-only
13	b-tag, τ-tag	[5]	V_{flav}	n.a.	trivial
14	charge Q_ℓ	[5]	V_{flav}	n.a.	trivial
15	$T^\star_{ijk}$ trivector dual	this work; cf. [18]	3→1	inv.	raw: H_{LHC} for T^{123}; covariant: Y
16	$\mathrm{sgn}\langle p_1p_2p_3p_4\rangle_4$	this work; cf. [18,22]	4	inv.	**N** (CP-odd 1-bit)
17	sphericity S, aplanarity A	event-shape (standard)	n.a.	lab	n.a. (orth. complement)
18	thrust T, Fox–Wolfram H_ℓ	event-shape (standard) [23,24]	n.a.	lab	n.a. (orth. complement)