Towards Foundational Models for HEP: Learning Universal Top-Quark Event Representations

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

Perfilov M., Volkov P., Vorotnikov G.

Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University

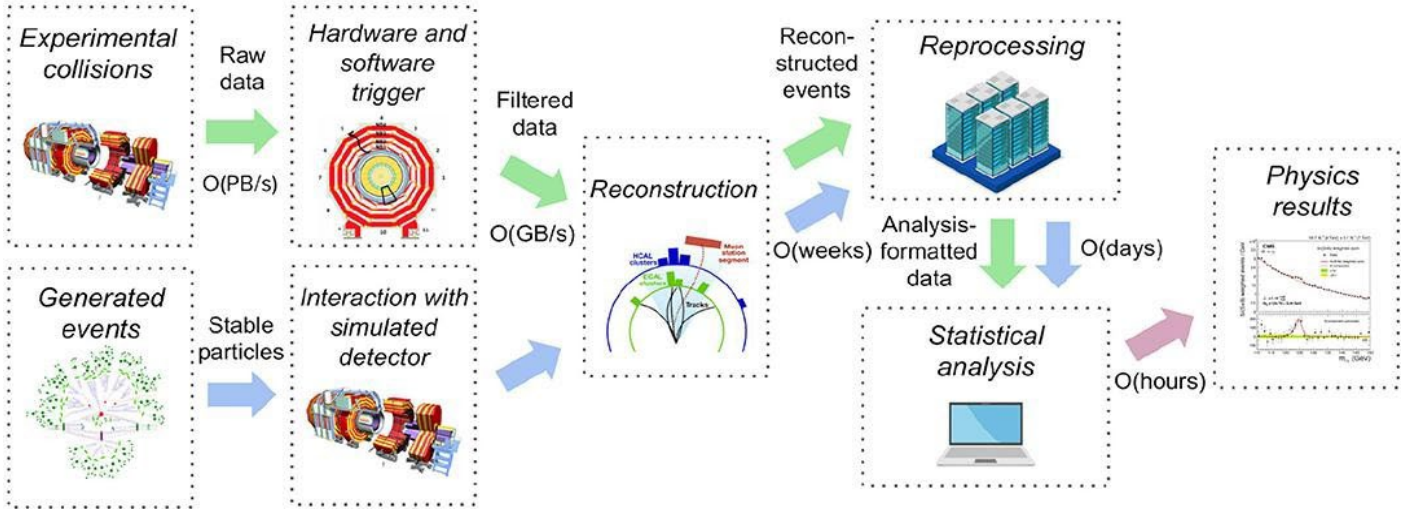
This study was conducted within the scientific program of the National Center for Physics and Mathematics, section #5 «Particle Physics and Cosmology». Stage 2026-2027

Motivation

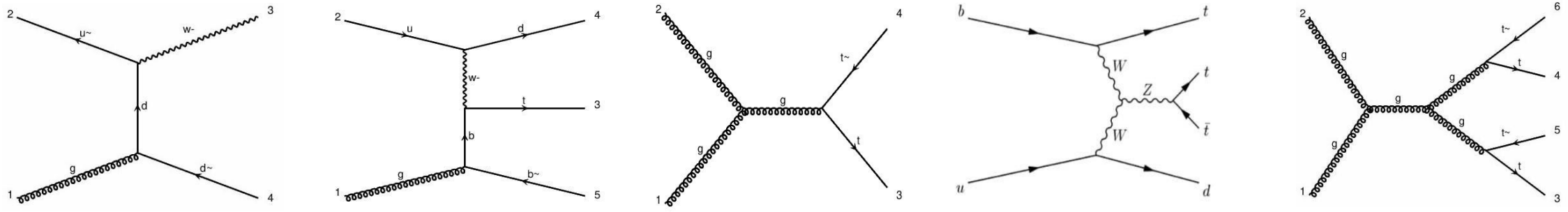
- Modern experiments in High Energy Physics generate enormous amounts of data
- Classical ML models require complex manual customisation for each task

- 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

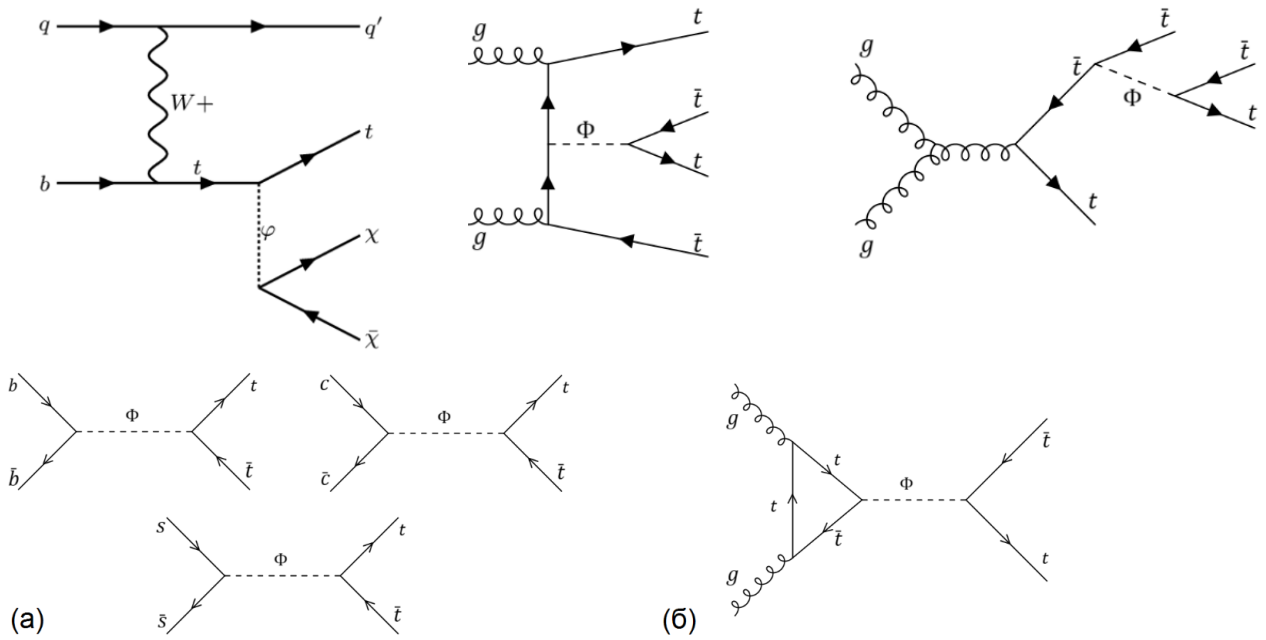


Data



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

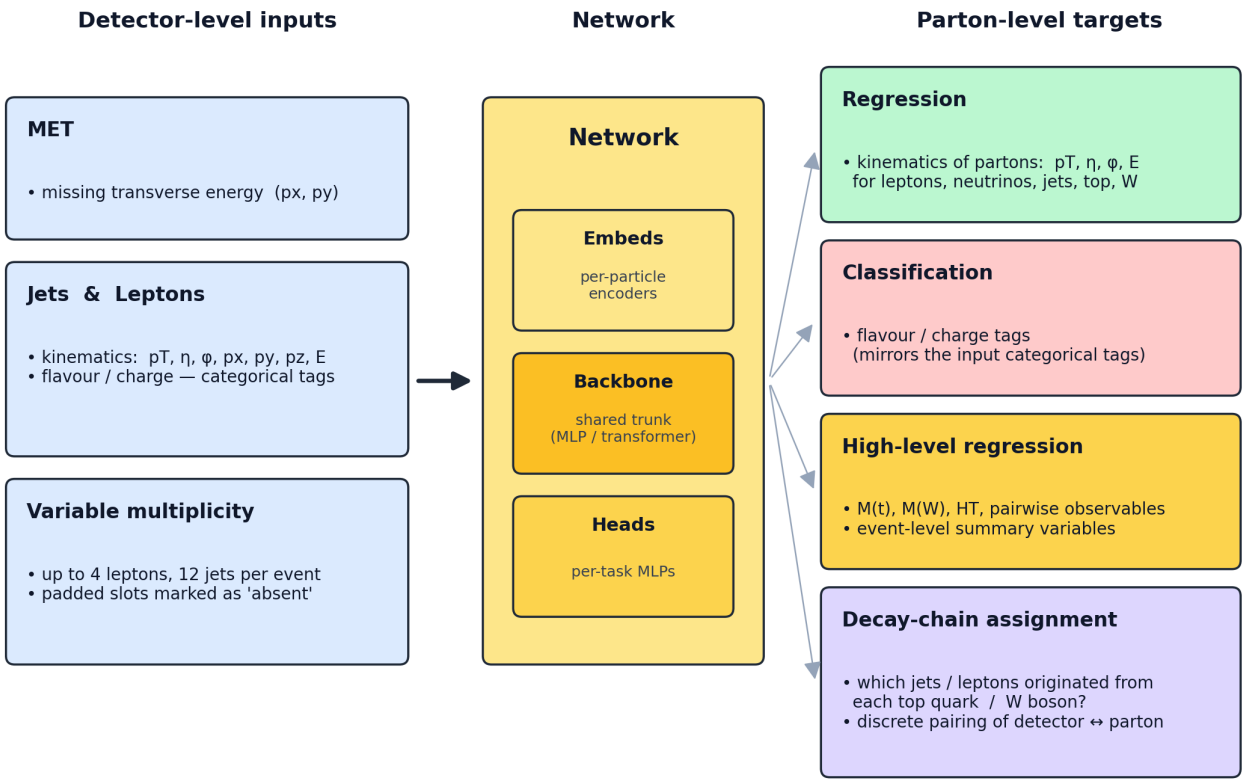


BSM processes

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

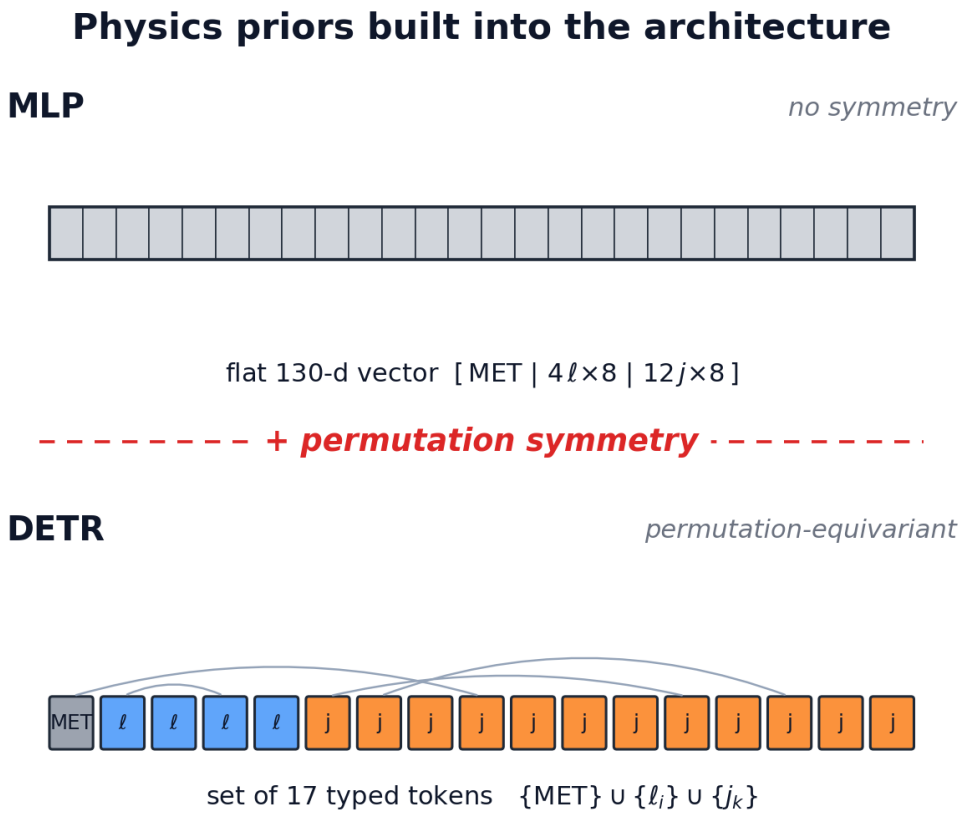


Pretraining: predict parton-level truth from the detector-level event description (four heads share the same backbone, trained jointly on a multi-task loss).

Models

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



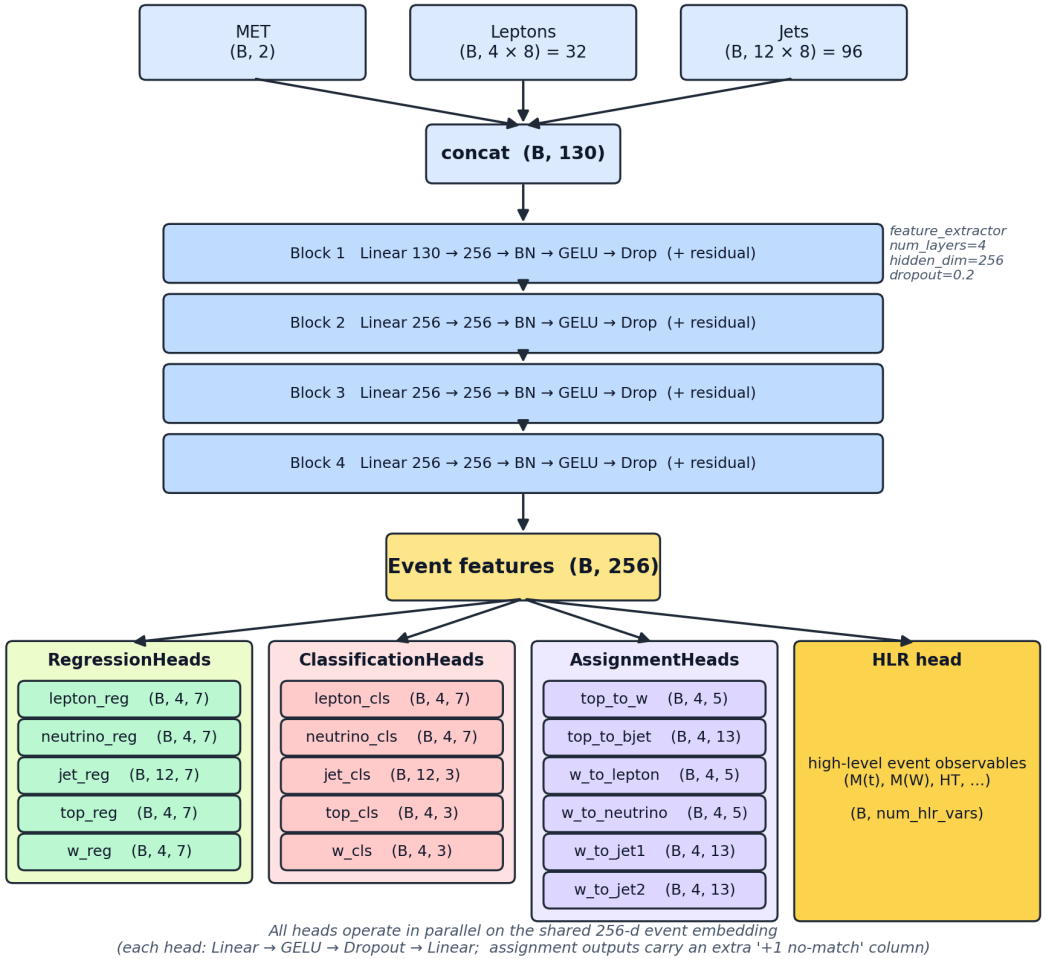
MLP: concepts

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.

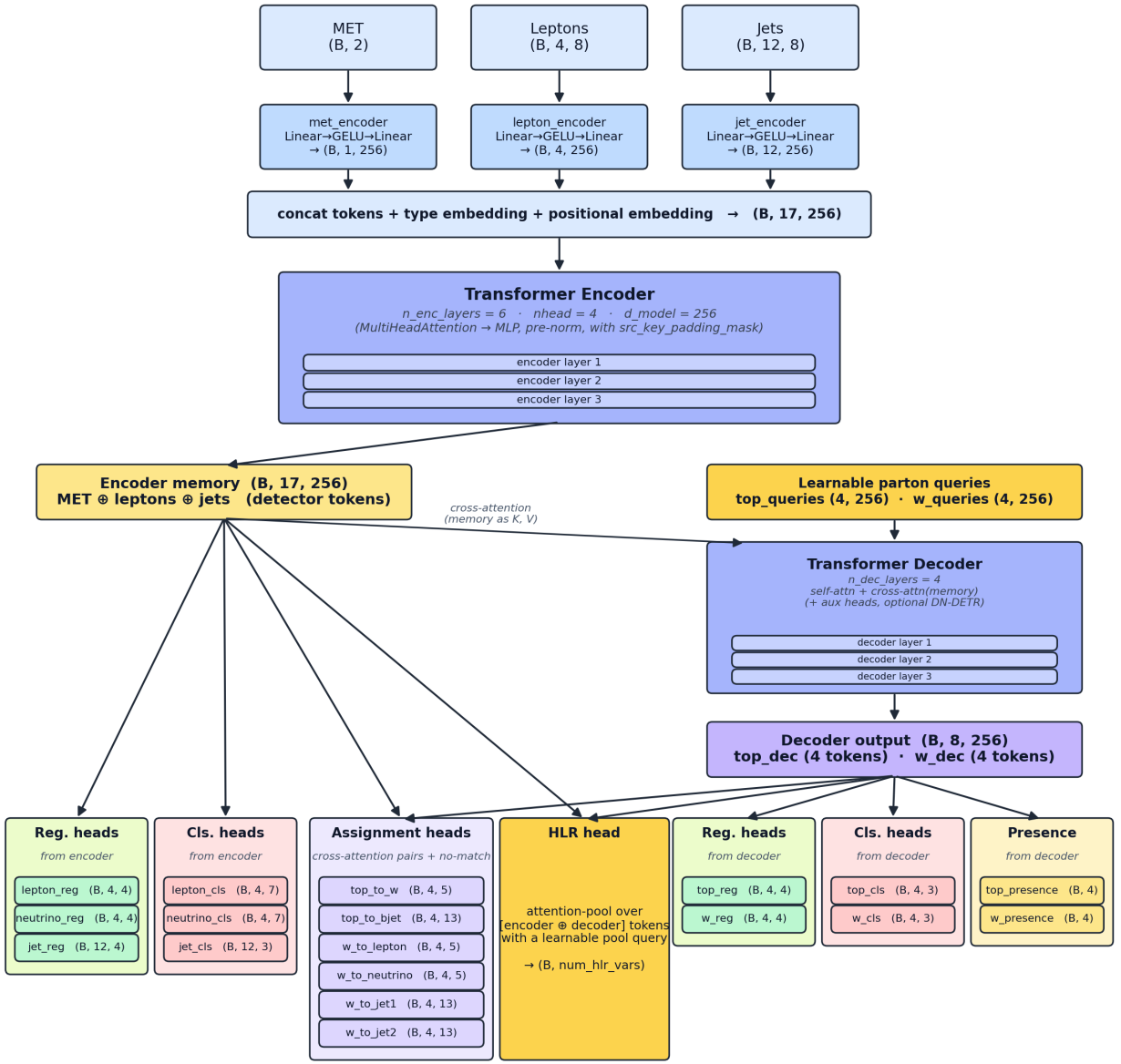


DETR: concepts

- 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

$$\boldsymbol{t}_k^\ell = \underbrace{\boldsymbol{h}_k^\ell}_{\text{content}} + \underbrace{\boldsymbol{e}_{\text{type}}^{[1]}}_{\text{type}} + \underbrace{\boldsymbol{p}_k^\ell}_{\text{position}}$$

  - 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

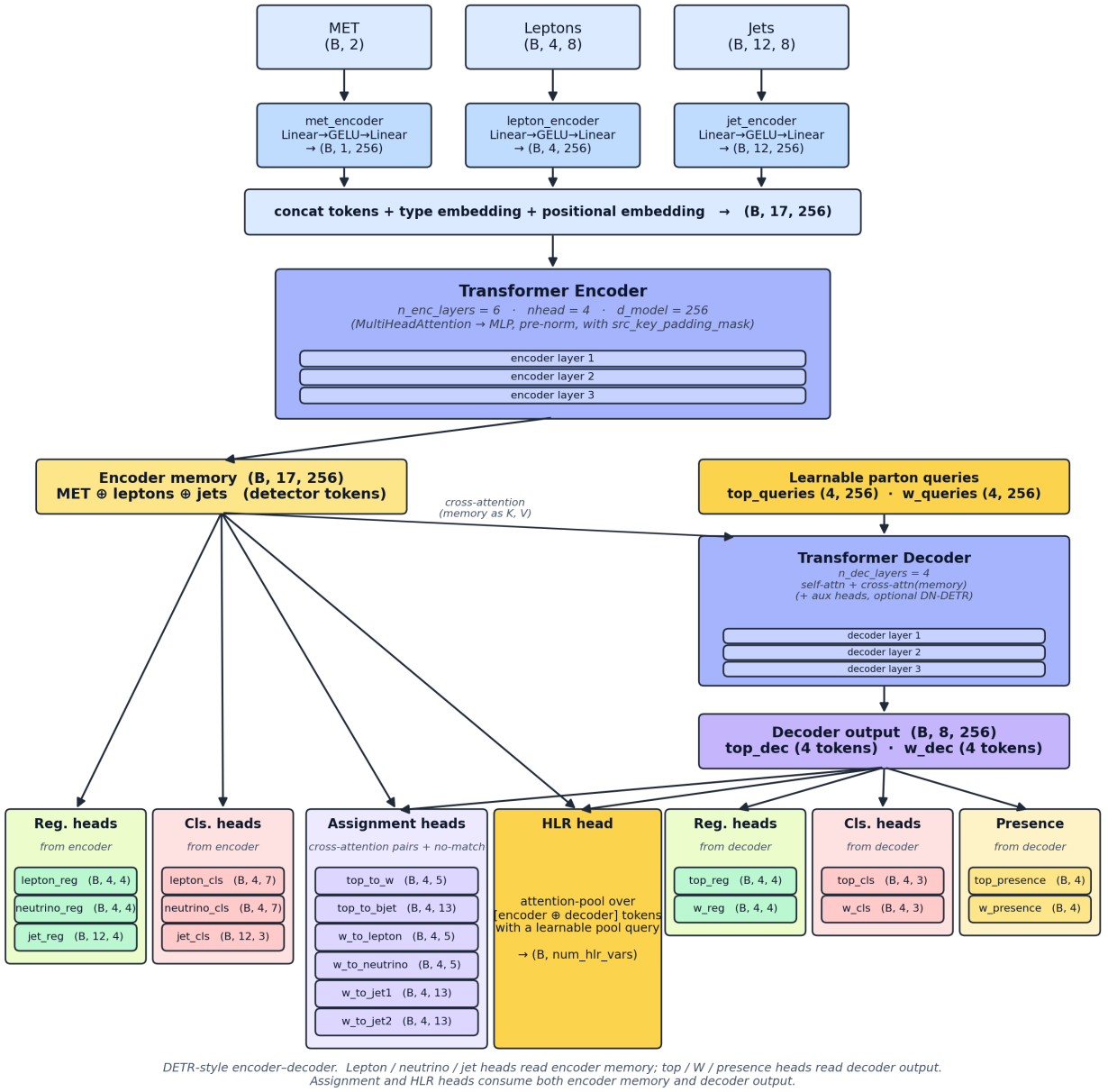
Heads: Per-Particle heads; each token in the backbone output carries information about the relevant particle, can ignore others

- 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

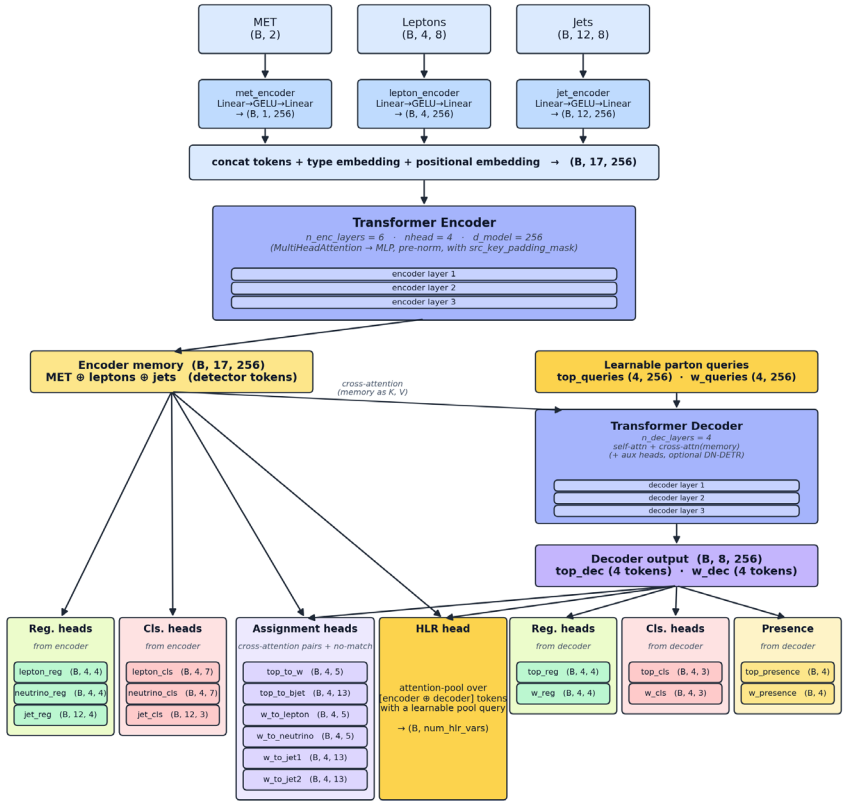
$$\text{FL}(p_t) = -(1-p_t)^\gamma \log(p_t), \qquad \gamma = 2$$

- 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

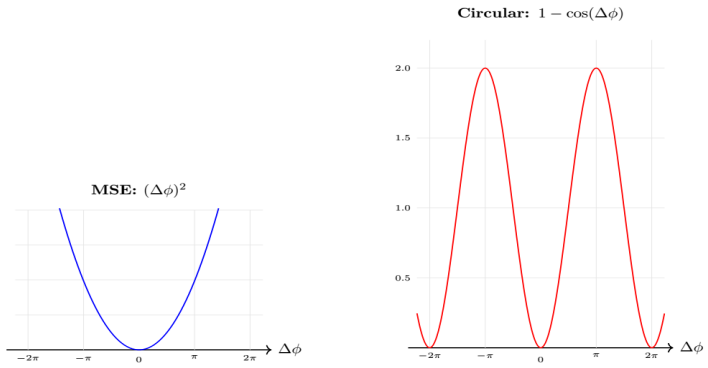
$$\mathcal{L}_{\text{aux}} = \sum_{l=1}^{L-1} 0.5^{(L-1-l)}\, \mathcal{L}_{\text{total}}^{(l)}$$

$$\mathcal{L}_{\text{total}} = w_{\text{reg}}\,\mathcal{L}_{\text{reg}} + w_{\text{cls}}\,\mathcal{L}_{\text{cls}} + w_{\text{assign}}\,\mathcal{L}_{\text{assign}} + w_{\text{hlr}}\,\mathcal{L}_{\text{hlr}}$$

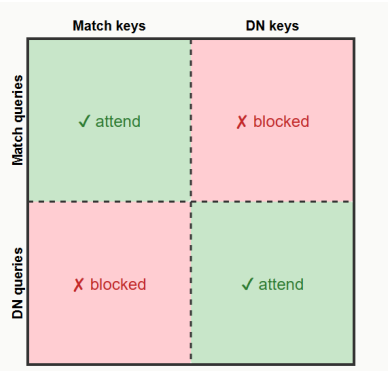
$$\mathcal{L} = \quad \mathcal{L}_{\text{total}} \quad + \quad \mathcal{L}_{\text{aux}} \quad + \quad \mathcal{L}_{\text{dn}}$$



Full DETR pipeline



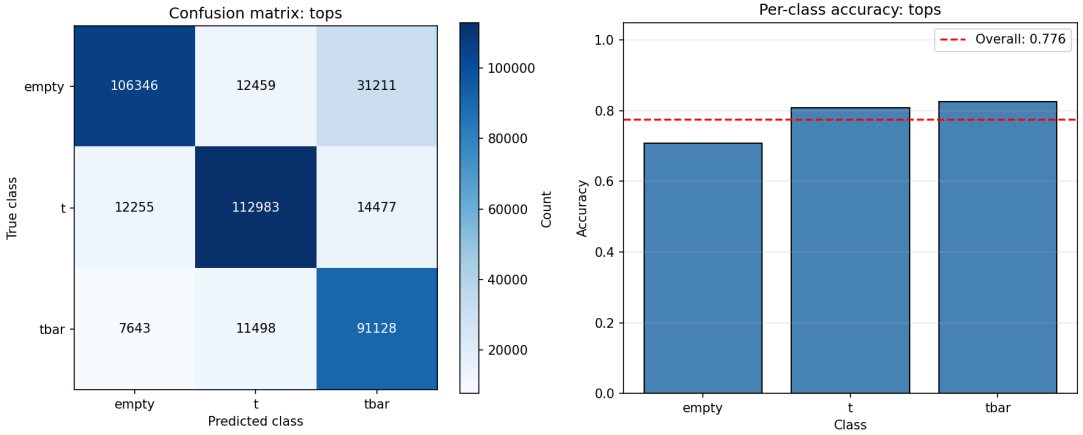
MSE and circular losses

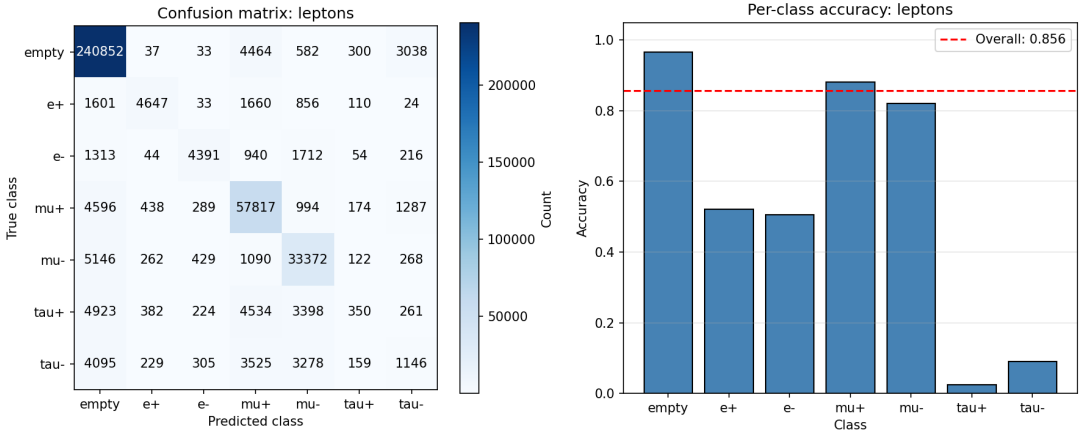


Denoising attention masking

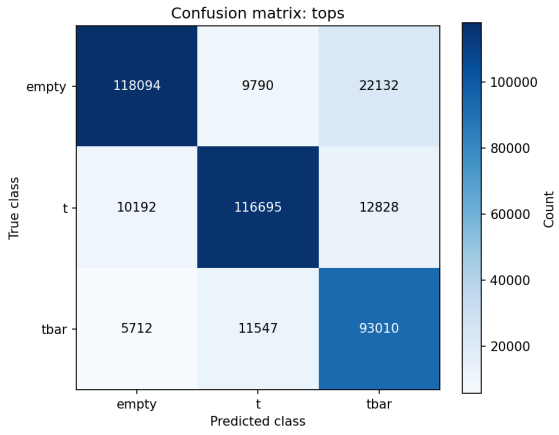
MLP/DETR: Classification

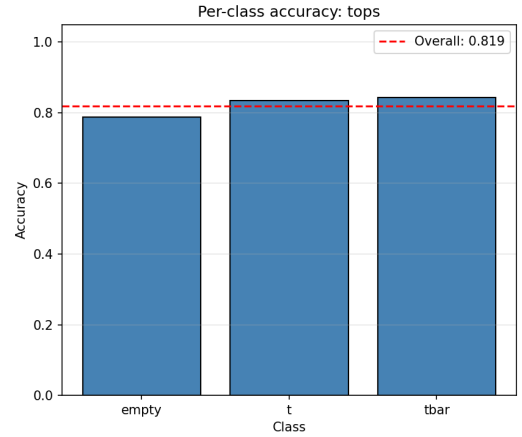
MLP

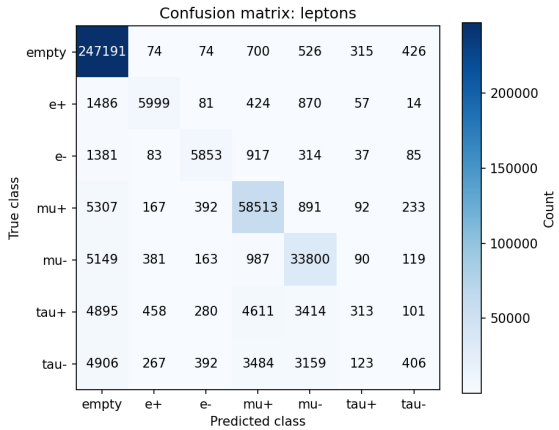


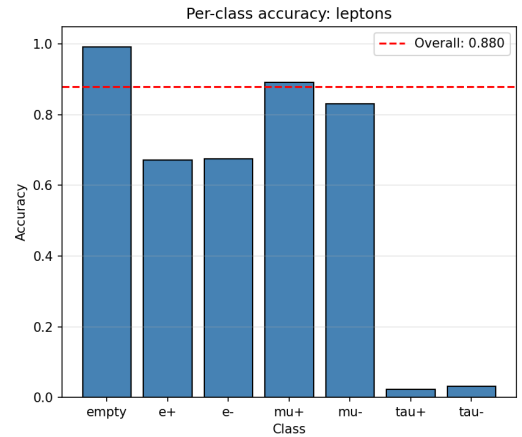


DETR

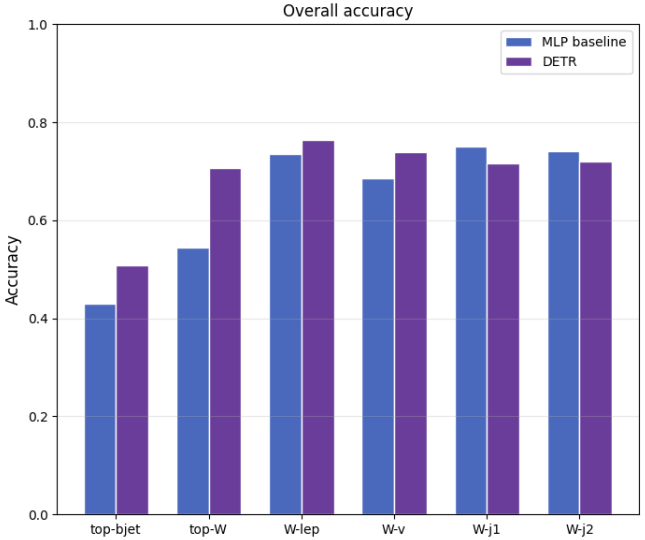


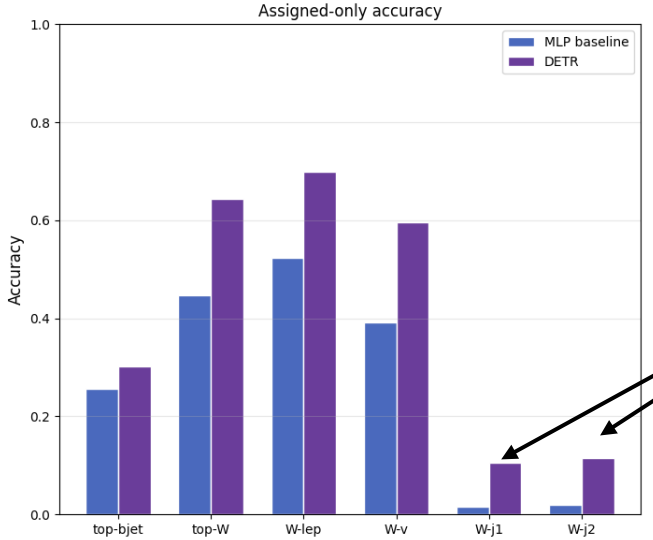




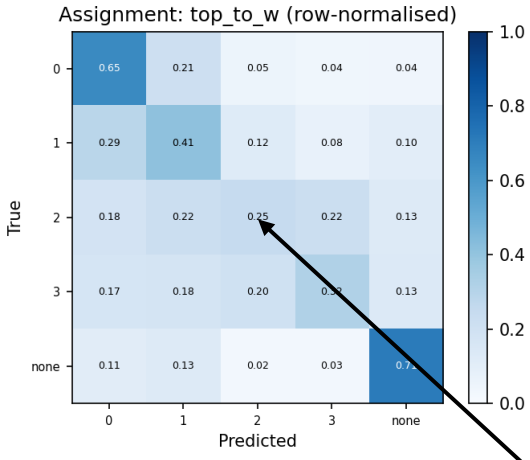


MLP/DETR: Assignment

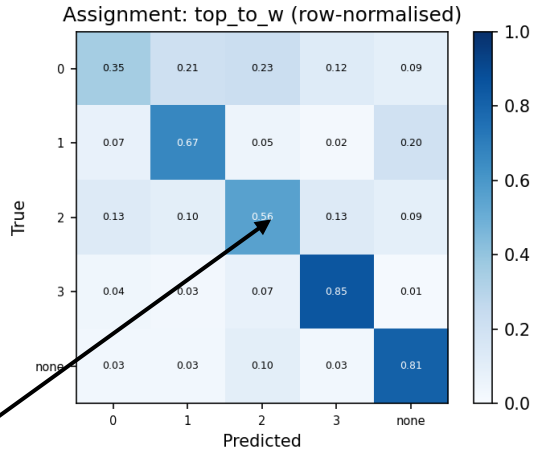




Very few $W \to \bar{j}j$ events in data → poor performance



MLP



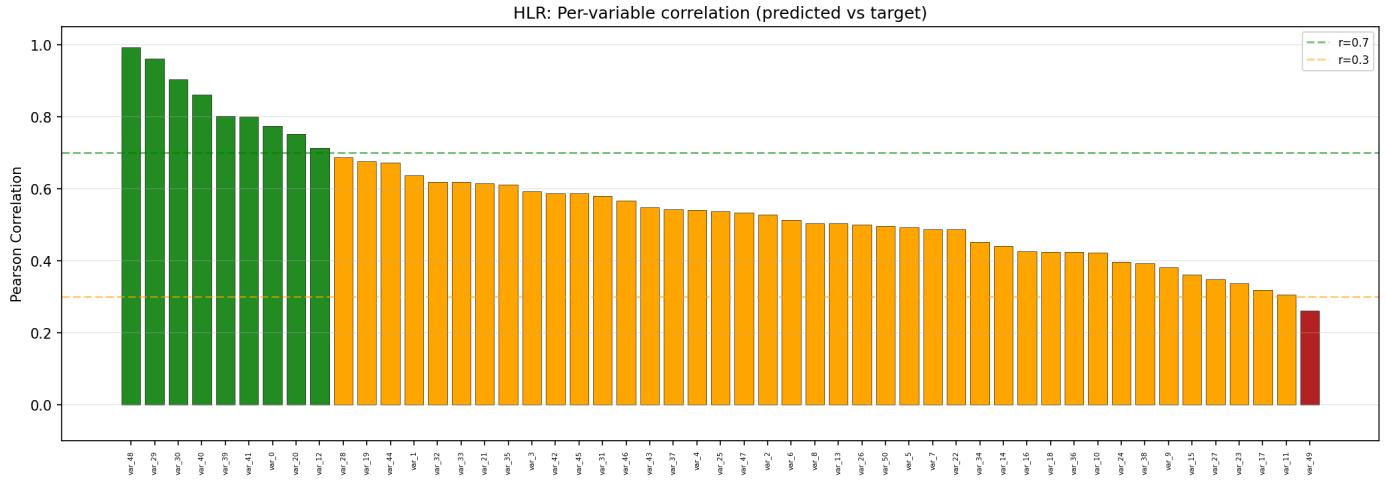
DETR

Much better separation of rarer classes

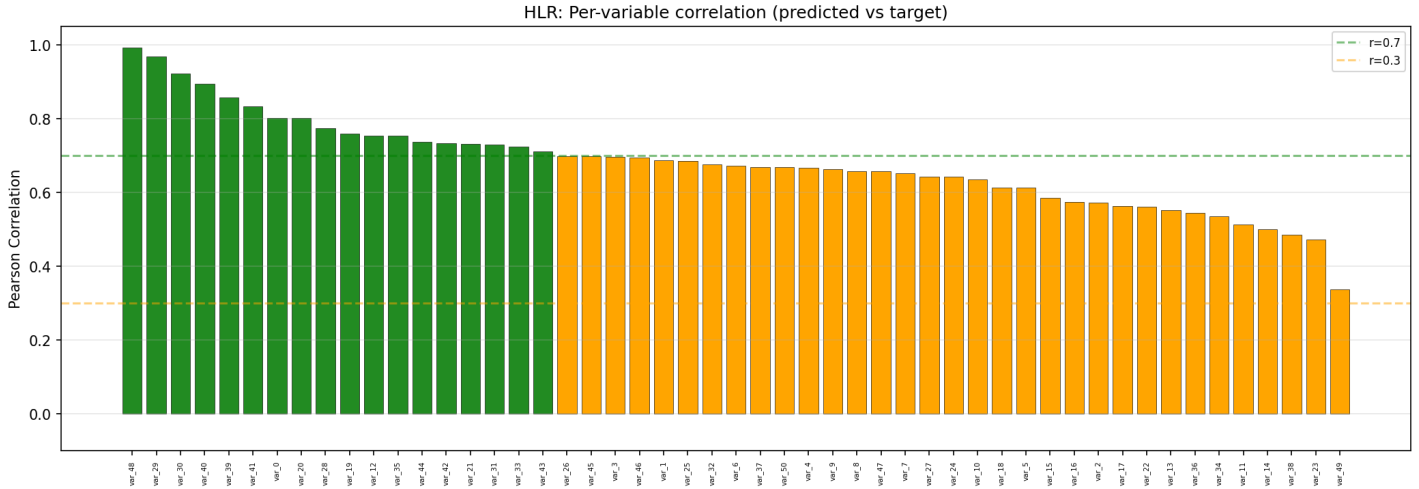
MLP/DETR: HLR

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

MLP

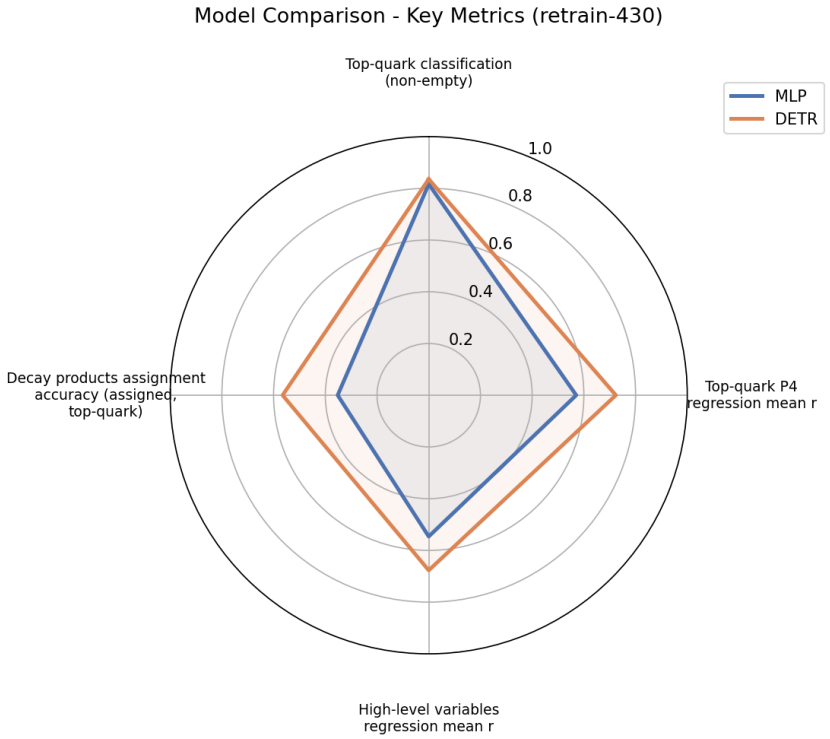


DETR



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



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 outperforms MLP across all 4 tasks; largest gains on complex tasks (assignment, top-quark classification)