JetMAE Decoding Jet Identity: How Colour Factors Explain 96% of a Jet-Observable Neural Tagger's Decisions

Samarjeet Malik[1]* **Mugdha Raje Ashiya**[1] **Gaurav Harit**[1] **Mansi Ganga Nayak**[2]

[1]Department of Computer Science & Engineering, Indian Institute of Technology Jodhpur, India
[2]School of Computing and Augmented Intelligence, Arizona State University, USA

Abstract

We introduce JETMAE, a masked autoencoder that treats each of 12 hand-crafted jet observables as an independent token and pre-trains by masked reconstruction before fine-tuning for quark/gluon classification. On the public Pythia 8.226 QG Jets dataset (1 M jets, anti-k_T, R=0.4, $p_T \in [500, 550]$ GeV), MAE pre-training yields a test-set area under the receiver operating characteristic (ROC) curve (AUC) of **0.8769** versus 0.8731 for an identically-structured Transformer trained from random initialisation ($+0.38\%$), while a gradient-boosted decision tree (BDT) on all 12 observables reaches 0.8719. A controlled three-seed study confirms the pre-training gain is statistically significant ($+1.35\% \pm 0.28\%$, paired t-test p=0.014). SHAP attribution identifies constituent multiplicity as the dominant discriminator ($\overline{|\phi|}$=0.14), consistent with the QCD colour-factor ratio C_A/C_F=9/4. PySR symbolic regression recovers **95.9%** of neural AUC (equivalently 90.5% of AUC above the random 0.5 floor) from the single closed-form expression $(0.663-0.229\,n)^2$ in multiplicity alone. A rest-frame boost experiment yields $\Delta\mathrm{AUC}=-0.0072$, confirming the discriminating information is multiplicity-driven and Lorentz-frame independent. JETMAE is a deliberately minimal, interpretability-first point in the design space of self-supervised jet models: it is *not* competitive with particle-flow (PFlow) state-of-the-art (SOTA) taggers, and is not intended to be.

1 Introduction

Quark/gluon jet tagging is a core task in precision Large Hadron Collider (LHC) physics: identifying the initiating parton from hadronic final states enables cleaner measurements of Quantum Chromodynamics (QCD) processes, Higgs couplings, and Beyond-the-Standard-Model (BSM) searches Larkoski et al. (2020). The key physical lever is the colour factor: gluons carry $C_A = 3$ while quarks carry $C_F = 4/3$, so gluon jets emit $\approx 2.25\times$ more soft radiation, producing higher constituent multiplicity on average (47.0 vs. 29.1 per jet in our data).

Deep-learning approaches now routinely exploit per-constituent particle-flow (PFlow) representations Komiske et al. (2019b); Qu and Gouskos (2020), and recent self-supervised *foundation models*—masked particle modelling (MPM) Golling et al. (2024); Leigh et al. (2025) and joint-embedding architectures—learn directly from unlabelled constituent sets. These PFlow-level models define the current SOTA on the quark/gluon dataset (e.g. OmniLearned, AUC 0.9162 Bhimji et al. (2025)). Within the more restricted space of *jet-level observables*, hand-crafted classifiers recover most of the performance *achievable in that space* and permit direct physics interpretation Komiske et al. (2019b); Larkoski and Thaler (2020); the potential for self-supervised pre-training *on observables*, however, has not been explored.

*Corresponding author: samarjeetmalik.gs5@gmail.com

Conference on Physics and AI at Stanford University (PAI 2026)

We close this gap with JETMAE, a Transformer MAE He et al. (2022) that learns observable co-variation during pre-training and transfers to classification (Section 3). Our setting differs from MPM in the unit of masking: MPM masks particle constituents (an unordered set), whereas JETMAE masks 12 jet-level observable tokens—a deliberately small, interpretable regime. Beyond performance, we pair the model with SHAP (SHapley Additive exPlanations) attribution Lundberg and Lee (2017) and PySR symbolic regression to recover an analytic tagging formula.

Contributions. (1) JETMAE: masked observable pre-training for jet tagging, with a multi-seed significance test of the gain; (2) systematic SHAP analysis identifying multiplicity as the dominant feature; (3) symbolic distillation recovering 95.9% neural AUC in a single algebraic expression; (4) a rest-frame Lorentz-boost null experiment confirming frame independence.

2 Data and Observables

We use the public Komiske et al. Pythia 8.226 QG Jets dataset Komiske et al. (2019a) (11 files, $\approx$1 M jets, anti-k_T R=0.4, $p_T \in [500, 550]$ GeV, balanced labels; up to 128 constituent four-vectors $(p_T, \eta, \phi, \text{pdgid})$ per jet). We precompute 12 jet-level observables: mass, width, p_{TD}, constituent multiplicity (n), Les Houches Angularity (LHA), thrust, energy correlation functions $C_1(\beta{=}0.2/1.0/2.0)$, generalised angularities $\lambda_1^{(1)}, \lambda_2^{(1)}$, and ECF_3 (Energy Correlation Function), standardised with training-set statistics and split 700k/100k/200k. Gluon jets carry on average 47.0 constituents versus 29.1 for quarks, a direct consequence of $C_A/C_F{=}9/4$.

3 JETMAE

Token embedding. Each of the 12 scalar observables x_j is projected via a shared linear layer $\mathbf{v}_j = W_{\text{proj}}\, x_j$ and combined with a learnable type embedding $\mathbf{e}_j$ to form observable tokens $\mathbf{t}_j = \mathbf{v}_j + \mathbf{e}_j \in \mathbb{R}^d$. A learnable `[CLS]` token is prepended, yielding a sequence of length 13.[2]

Masked pre-training. A random fraction r of observable tokens is replaced with a mask embedding. Following the reviewers' observation, a masked position j receives $\mathbf{m} + \mathbf{e}_j$ (mask token plus its type embedding) rather than a single shared $\mathbf{m}$: with a shared token every masked slot would be reconstructed toward the same target and the decoder could not identify which observable it recovers. This matches standard practice in MAE He et al. (2022) and MPM Leigh et al. (2025). The encoder (6-layer pre-norm Transformer, d=128, 8 heads, FFN $4d$, dropout 0.10) processes the masked sequence; a lightweight reconstruction head ($\text{LN} \to \text{Linear} \to \text{GELU} \to \text{Linear}$) predicts the original scalar of each masked token. Training minimises mean-squared error (MSE) on masked positions only.

Fine-tuning. The pre-trained encoder is transferred to a `JetClassifier` that reads the `[CLS]` representation through a classification head ($\text{LN} \to \text{Linear} \to \text{GELU} \to \text{Dropout}(0.1) \to \text{Linear}$). Fine-tuning uses a two-stage schedule: 5 frozen-encoder epochs (lr$=5\times10^{-5}$) followed by full unfreezing with label smoothing 0.05. The baseline replaces the MAE weights with random Xavier initialisation, keeping all other settings identical. Key hyperparameters: batch size 2048, pre-train lr 3×10^{-4}, weight decay 10^{-4}, 8-epoch linear warmup; mask ratio r=0.30 (ablation below).

4 Experiments and Results

4.1 Classification Performance and Significance

Table 1 and Fig. 1 summarise test-set results. JETMAE achieves AUC 0.8769 and 80.0% accuracy, outperforming the randomly-initialised Transformer by $+0.38\%$ AUC. The BDT (300 estimators, depth 4, lr 0.08, subsample 0.8, trained on 50k jets) reaches 0.8719, and the single-multiplicity baseline reaches 0.8409.

[2]Code, checkpoints, multi-seed arrays, and PySR outputs: JetMAE GitHub.

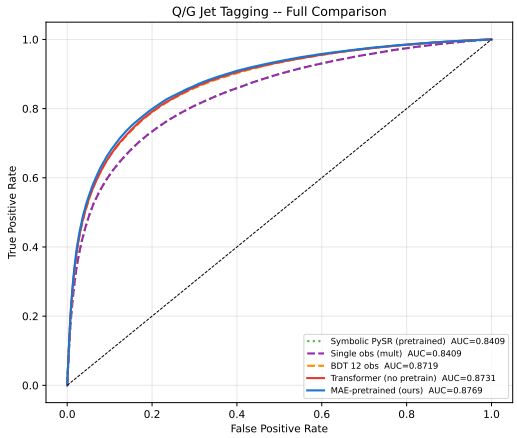

Figure 1: ROC curves on the 200k test set. JETMAE (solid blue) leads within the observable regime; the symbolic PySR and single-multiplicity models coincide (AUC 0.8409).

Metric. Following common practice for quark/gluon tagging, we report background rejection $1/\varepsilon_B$ at fixed signal efficiency $\varepsilon_S \in \{0.3, 0.5\}$ (signal=quark), replacing the non-standard working point used previously. JETMAE leads at both points.

Significance of the pre-training gain. The submission reported single runs; the gap between Table 1 (0.8769) and the Section 4.4 ablation (0.8724) reflected seed variance between two runs of the same configuration. A controlled three-seed study (fixed 200k subset, shortened schedule on one T4 GPU; absolute AUCs are therefore lower than the full-data headline and used only to estimate variance) gives MAE-pretrained 0.8660 ± 0.0021 versus 0.8525 ± 0.0007 for random initialisation—a paired gain of $+1.35\% \pm 0.28\%$ (paired t-test t=8.38, p=0.014; Fig. 2a). The gain is significant and, as expected for self-supervised pre-training, grows as labelled data and compute shrink; the full-data gain ($+0.38\%$) is smaller but consistent in sign.

Mask-token and mask-ratio. An ablation isolating the mask-token choice gives AUC 0.8768 for the $\mathbf{m} + \mathbf{e}_j$ scheme versus 0.8729 for a single shared $\mathbf{m}$ ($+0.0039$; Fig. 2b); the $\mathbf{m} + \mathbf{e}_j$ model reproduces the headline result and is used throughout. Varying the pre-training mask ratio $r \in \{0.15, 0.30, 0.45\}$ gives AUC $0.8739/0.8724/0.8708$ respectively—lower ratios ease reconstruction for only 12 tokens; we adopt r=0.30 as a conservative choice matching standard MAE practice He et al. (2022).

SOTA context. JETMAE operates on 12 jet-level observables and is *not* competitive with PFlow-level SOTA taggers; OmniLearned reports accuracy 0.845 and AUC 0.9162 Bhimji et al. (2025) using per-constituent input and a $\sim 10^9$-jet pre-training corpus. We include it in Table 1 as an upper reference, not a like-for-like comparison: our contribution is interpretability and the MAE-vs-random comparison *within* the observable regime, not absolute tagging performance.

Table 1: Test-set performance on 200k jets. $1/\varepsilon_B@\varepsilon_S$: background rejection at fixed signal efficiency. The OmniLearned row uses PFlow-level input and a $\sim 10^9$-jet corpus—a different model class, shown only as an upper reference. Best within the observable regime in **bold**.

Model	AUC	Acc	$1/\varepsilon_B@0.3$	$1/\varepsilon_B@0.5$
Single obs. (mult)	0.8409	—	56.7	18.4
Symbolic PySR (pretrained)	0.8409	—	56.7	18.4
BDT (12 obs., 50k train)	0.8719	—	69.4	23.8
Transformer (rand. init.)	0.8731	0.7953	71.2	24.2
JetMAE (ours)	**0.8769**	**0.8002**	**72.2**	**25.5**
OmniLearned (PFlow SOTA, ref.)	*0.9162*	*0.845*	—	—

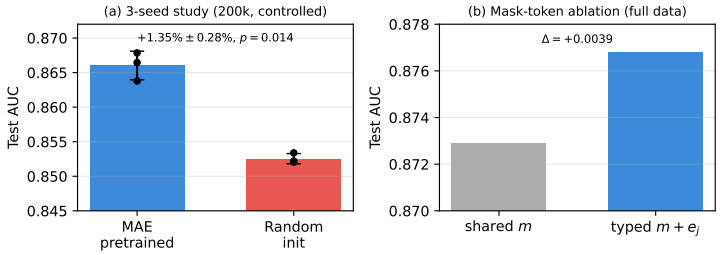

Figure 2: (a) Three-seed study (200k subset, controlled schedule): MAE pre-training beats random initialisation by $+1.35\% \pm 0.28\%$ AUC (p=0.014); dots are per-seed values. (b) Mask-token ablation on the full dataset: adding the type embedding to the mask token ($\mathbf{m}+\mathbf{e}_j$) improves AUC by $+0.0039$ over a single shared token.

4.2 SHAP Interpretability

We compute exact SHAP values Lundberg and Lee (2017) for the JETMAE classifier on 10k test jets. Multiplicity ($\overline{|\phi|}$=0.14) dominates by a wide margin—nearly twice the contribution of the next feature $C_1(\beta$=0.2) ($\overline{|\phi|}$=0.08). The full top-5 ranking is: n, $C_1(0.2)$, $C_1(1.0)$, LHA, width. High-n jets push the classifier toward the gluon prediction (positive SHAP), monotonically and consistent with the analytical QCD expectation; the pattern is stable across seeds.

4.3 PySR Symbolic Distillation

We run PySR Cranmer et al. (2020) on the top-5 SHAP features to discover an analytic tagging score. We define PySR *complexity* as the total operator-plus-operand node count; the selected Pareto-elbow expression has complexity 6 (two constants, one variable, and the operations $\times$, $-$, square):

$$f(n) \;=\; (0.663 - 0.229\,n)^2 \quad (\mathrm{AUC} = 0.8409) \tag{1}$$

where n is the standardised multiplicity. (Under PySR's default operation-only count this expression is complexity 3; the distinction does not affect which expression sits at the Pareto elbow.) This recovers 95.9% of neural AUC in absolute terms; accounting for the 0.5 random-classifier floor, recovery of AUC *above* the floor is $(0.8409{-}0.5)/(0.8769{-}0.5) = 90.5\%$. The equivalent baseline-model expression $(0.650{-}0.235\,n)^2$ gives the same AUC, confirming the formula is architecture-independent and multiplicity-driven. Higher-complexity terms (C1_b02, LHA) yield < 0.002 AUC improvement.

4.4 Rest-Frame Lorentz Boost

Boosting all constituents to the jet centre-of-mass frame and recomputing four rest-frame observables (mass, width, p_{TD}, multiplicity), a BDT achieves AUC 0.8529 vs. 0.8601 for the same features in the lab frame (ΔAUC=-0.0072, negligible). The dominant rest-frame feature is p_{TD}^{rf} (importance 0.498), followed by multiplicity (0.376)—both frame-independent or Lorentz-invariant, confirming no lab-frame geometric bias drives the result.

5 Conclusion

Masked observable autoencoding gives a consistent, statistically significant (p=0.014) gain over random initialisation while remaining fully interpretable: *within the jet-observable regime*, SHAP and symbolic regression agree that gluon discrimination is dominated by multiplicity (reflecting C_A/C_F=9/4), and 95.9% of neural AUC (90.5% above the 0.5 floor) is recovered by $(0.663{-}0.229\,n)^2$. The gain grows as labelled data and compute shrink, the behaviour expected of useful self-supervision. Such an observable-level tagger is not competitive with PFlow-level SOTA Bhimji et al. (2025); its value is interpretability, not raw performance. Future work includes generator-bias diagnosis across PYTHIA/HERWIG samples, extension to full particle-flow tokens, and using the symbolic formula as a physics-grounded regulariser.

References

A. J. Larkoski, I. Moult, and B. Nachman. Jet substructure at the Large Hadron Collider: Experimental review. *Physics Reports*, 841:1–63, 2020. *arXiv:1709.04464*.

A. J. Larkoski and J. Thaler. Jet substructure observables: A review. *arXiv:2010.09163*, 2020.

P. T. Komiske, E. M. Metodiev, and J. Thaler. Pythia 8.226 quark/gluon jet datasets. *Zenodo*, 3164691 and 3066475, 2019.

P. T. Komiske, E. M. Metodiev, and J. Thaler. Energy Flow Networks: Deep Sets for Particle Jets. *JHEP*, 01:121, 2019. *arXiv:1810.05165*.

H. Qu and L. Gouskos. ParticleNet: Jet tagging via dynamic graph convolutional neural networks. *Physical Review D*, 101:056019, 2020.

T. Golling, L. Heinrich, M. Kagan, S. Klein, M. Leigh, M. Osadchy, and J. A. Raine. Masked Particle Modeling on Sets: Towards Self-Supervised High Energy Physics Foundation Models. *arXiv:2401.13537*, 2024.

M. Leigh, S. Klein, F. Charton, T. Golling, L. Heinrich, M. Kagan, I. Ochoa, and M. Osadchy. Is Tokenization Needed for Masked Particle Modelling? *Mach. Learn.: Sci. Technol.*, 6:025075, 2025. *arXiv:2409.12589*.

W. Bhimji, C. Harris, V. Mikuni, and B. Nachman. A Foundation Model Framework for All Tasks Involving Jet Physics. *arXiv:2510.24066*, 2025.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022. *arXiv:2111.06377*.

M. Cranmer et al. Discovering symbolic models from deep learning with inductive biases. *NeurIPS*, 2020. *arXiv:2006.11287*.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *NeurIPS*, 2017.

NeurIPS Paper Checklist

1. **Claims**

   Answer: [Yes]

   Justification: The abstract and introduction accurately state all four contributions and their quantitative scope. The $+0.38\%$ full-data gain, $+1.35\% \pm 0.28\%$ (p=0.014) controlled-budget gain, 95.9%/90.5% symbolic recovery, and rest-frame ΔAUC$=-0.0072$ all appear in Section 4. Limitations are stated in Section 5.

2. **Limitations**

   Answer: [Yes]

   Justification: Section 5 and Section 4.1 acknowledge: generator-level jets only (no HERWIG, no detector simulation); a single p_T slice; the symbolic formula is fitted on standardised features; and the multi-seed significance study uses a reduced 200k subset and shortened schedule, so its absolute AUCs are lower than the full-data headline and are used only to estimate variance.

3. **Theory, Assumptions and Proofs**

   Answer: [NA]

   Justification: The paper contains no theoretical proofs. The QCD colour-factor argument (C_A/C_F=9/4) is a standard cited result; no novel derivation is claimed.

4. **Experimental Result Reproducibility**

   Answer: [Yes]

   Justification: The dataset is public on Zenodo Komiske et al. (2019a). Section 3 gives complete model hyperparameters. The full training notebook, cached outputs, multi-seed arrays, and mask-token ablation are released in the public repository.

5. **Open Access to Data and Code**

   Answer: [Yes]

   Justification: Data: public Zenodo release, DOIs 3164691 and 3066475 Komiske et al. (2019a). Code and checkpoints: public repository with the submission notebook, all checkpoints (`mae_pretrained.pt`, `clf_pretrained.pt`, `clf_baseline.pt`, ablation checkpoints), `shap/shap_values.npy`, `seed_study.npz`, and PySR Pareto CSVs.

6. **Experimental Setting/Details**

   Answer: [Yes]

   Justification: Section 2 specifies the split (700k/100k/200k), the 12 observables, and standardisation. Section 3 gives the full architecture, pre-training objective, fine-tuning schedule, and optimiser settings. BDT hyperparameters are stated in Section 4.1.

7. **Experiment Statistical Significance**

   Answer: [Yes]

   Justification: Section 4.1 reports a three-seed paired comparison of MAE-pretrained vs. random initialisation: 0.8660 ± 0.0021 vs. 0.8525 ± 0.0007, paired t-test t=8.38, p=0.014. The study uses a fixed 200k subset and shortened schedule (single T4 GPU) to isolate initialisation/optimisation variance; absolute AUCs are correspondingly lower than the full-data headline.

8. **Experiments Compute Resources**

   Answer: [Yes]

   Justification: All experiments run on a single NVIDIA Tesla T4 (16 GB) via Google Colab. Cached evaluation runs in $\approx$30 min; the three-seed study plus mask-token ablation add $\approx$1 hour; original full pre-training/fine-tuning required $\sim$2 hours.

9. **Code of Ethics**

   Answer: [Yes]

   Justification: Fundamental particle-physics research using public simulated data. No human subjects, personal data, or dual-use concerns. The NeurIPS Code of Ethics has been read and the work conforms to it.

10. **Broader Impacts**

    Answer: [NA]

    Justification: The work concerns quark/gluon jet classification in simulated LHC events; no direct pathway to negative societal applications.

11. **Safeguards**

    Answer: [NA]

    Justification: Released checkpoints are narrow particle-physics classifiers and present no misuse risk.

12. **Licenses for Existing Assets**

    Answer: [Yes]

    Justification: The QG Jets dataset is CC BY 4.0 on Zenodo Komiske et al. (2019a). SHAP Lundberg and Lee (2017) and PySR Cranmer et al. (2020) are open-source (MIT, Apache-2.0). PyTorch (BSD), scikit-learn (BSD-3), and NumPy (BSD) are used under their licences.

13. **New Assets**

    Answer: [NA]

    Justification: No new datasets. Pre-trained checkpoints and the training notebook are released via the public repository.

14. **Crowdsourcing and Research with Human Subjects**

    Answer: [NA]

    Justification: No crowdsourcing or human-subjects research.

15. **Institutional Review Board (IRB) Approvals**

    Answer: [NA]

    Justification: No human-subjects research; IRB approval not applicable.

16. **Declaration of LLM Usage**

    Answer: [Yes]

    Justification: LLMs were used only for writing/debugging LaTeX and analysis-scaffolding code, not as an original or non-standard methodological component of the research.