What Do Lorentz-Equivariant Jet Taggers Learn?

Jay Agarwal [1] **Siddharth Khare** [1] **Dhruv Kumar** [1]

Abstract

We study what Lorentz-equivariant jet taggers learn internally, using equivariance tests, linear probes and grade ablations across five models including L-GATr, L-GATr-slim and LLoCa-T. Linear probes show that equivariant models suppress frame-dependent pseudorapidity to zero while encoding jet mass and N-subjettiness strongly. Grade ablations on L-GATr reveal that bivector channels are negligible for top-quark tagging while vector-like channels are dominant but seed-variable, consistent with the network exploiting multiple representational pathways. These results characterize which physical features and algebraic grade structures carry discriminative information in equivariant taggers and may inform future development of such models.

1. Introduction

Equivariant neural networks are becoming central tools for scientific machine learning because they build known physical symmetries directly into the model. In high-energy physics (HEP), Lorentz symmetry is especially important: jets are observed in a particular detector frame, but the underlying physics is constrained by relativistic transformations. Recent Lorentz-equivariant architectures achieve strong performance on jet tagging benchmarks, yet high accuracy alone does not tell us what physics these models have learned, whether their symmetry constraints are active in practice, or how their internal representations differ from ordinary transformers.

This question is timely for trustworthy AI in science. Collider analyses increasingly rely on neural networks to extract subtle structure from complex events. If symmetry-aware models are to become reliable scientific instruments, we need tools that inspect not only their outputs, but also whether their internal representations align with the structure of the physical problem.

We study the Lorentz-equivariant Geometric Algebra Transformer (L-GATr (Spinner et al., 2024)), which builds on the Geometric Algebra Transformer framework (Brehmer et al., 2023) by representing jet constituents as multivectors in the geometric algebra of Minkowski spacetime. This representation decomposes into grades: scalars, vectors, bivectors, trivectors and pseudoscalars. The grade structure gives L-GATr an unusual advantage for interpretability: unlike a generic hidden vector, each component has a transformation law and a geometric meaning.

We study L-GATr and four comparison models: L-GATr-slim, LLoCa-T, Vanilla-T and ParT, using equivariance checks, linear probes, grade ablations and invariant multivector probes on TopTagging. Our main contributions are:

(1) Linear probes reveal invariant physical representations. Linear probes show that all three equivariant architectures suppress frame-dependent pseudorapidity to zero while strongly encoding jet mass and N-subjettiness, a representation-level signature of the imposed symmetry that non-equivariant baselines do not exhibit. LLoCa-T achieves the highest probe scores of all five models for most physics targets, suggesting that probe accessibility reflects representational strategy rather than task utility.

(2) Grade ablations reveal multiple usable pathways. Ablating grade groups within L-GATr's multivector hidden state shows that one algebraic grade (bivectors) is consistently negligible for top-quark tagging across all seeds, while vector-like channels are dominant but vary substantially across independently trained seeds, consistent with the network exploiting multiple representational pathways rather than a single fixed strategy.

These findings demonstrate that equivariance tests, linear probes and grade ablations transfer directly to models with structured Lorentz representations, suggesting they can guide validation and design of future equivariant collider models.

Relation to prior work. Recent interpretability studies of jet transformers have focused mainly on attention maps for non-equivariant architectures. Wang et al. (2024) study ParT attention, finding sparsity and subjet structure, while Legge

[1]BITS Pilani, Pilani, India. Correspondence to: Jay Agarwal <f20221305@pilani.bits-pilani.ac.in>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

et al. (2025) caution that attention sparsity does not by itself explain model decisions. Esmail et al. (2026) use attention maps and CKA to analyze learned representations in IAFormer. Linear probes were used in an early jet-tagging interpretability study by Cheng (2019), who found that individual N-subjettiness values are more accessible than their ratios. We extend this probe-based approach to transformer architectures and Lorentz-equivariant models and add grade-aware interventions specific to geometric-algebra networks. LLoCa-T (Spinner et al., 2025) provides a comparison point between full architectural equivariance and non-equivariant baselines.

The Appendix reports supporting diagnostics: attention maps, CKA layer similarity (Kornblith et al., 2019), full scalar- and multivector-probe trajectories, grouped and subgroup-false ablation tables and details of the combined-bootstrap uncertainty estimates used in the figures and tables.

2. Background

2.1. Geometric Algebra and the Lorentz Group

The geometric algebra over Minkowski spacetime, with metric signature $(+,-,-,-)$, is a 16-dimensional associative algebra generated by four basis vectors e_0, e_1, e_2, e_3 satisfying $e_0^2 = +1$, $e_i^2 = -1$ for $i = 1, 2, 3$, and $e_\mu e_\nu = -e_\nu e_\mu$ for $\mu \neq \nu$. Elements are called *multivectors* and decompose into five *grades*:

- **Grade 0** (scalar, 1 component): Lorentz-invariant by construction.
- **Grade 1** (vectors, 4 components): transform as 4-vectors under $\mathrm{SO}^+(1,3)$.
- **Grade 2** (bivectors, 6 components): oriented spacetime planes.
- **Grade 3** (trivectors, 4 components): Hodge-dual to G1 vectors via the pseudoscalar I.
- **Grade 4** (pseudoscalar, 1 component): parity-odd invariant.

We write grade k as Gk throughout the paper; for example, G2 denotes bivectors.

The geometric product of two multivectors mixes grades in a way determined by the algebra; for instance, the product of two G2 elements produces G0 and G4 components, while a G2 element times a G1 element produces G1 and G3 components.

2.2. Lorentz-Equivariant Geometric Algebra Transformer

L-GATr (Spinner et al., 2024) adapts the Geometric Algebra Transformer framework (Brehmer et al., 2023) from Euclidean to Minkowski spacetime. It represents each jet constituent as a multivector token encoding its 4-momentum, supplemented by beam spurion tokens (a lightlike vector along the beam axis and a time reference) that break the remaining boost symmetry as a fixed reference frame. The network consists of N_{blocks} GATrBlocks, each applying equilinear transformations (linear maps that commute with the Lorentz group action) and geometric attention over multivector tokens. Each block produces multivector channels $h_{\text{mv}} \in \mathbb{R}^{N \times C_{\text{mv}} \times 16}$ and scalar channels $h_{\text{s}} \in \mathbb{R}^{N \times C_{\text{s}}}$.

The output projection layer is equilinear: it maps grade k inputs to grade k outputs only, with no grade mixing. Since the classification logit is a scalar (G0), the output layer reads *only* from the G0 component of h_{mv} at the final layer. This architectural constraint implies that all non-scalar components at the final layer are architecturally inaccessible to the output, regardless of their information content. We use this fact explicitly in interpreting the layer-resolved ablations.

The scalar channels h_{s} are Lorentz-invariant by construction.

By default L-GATr enforces equivariance to the connected, proper orthochronous Lorentz subgroup $\mathrm{SO}^+(1,3)$. Under this subgroup, G0 and G4 carry scalar-type representations, while G1 and G3 carry equivalent vector-type representations via Hodge duality; the implementation therefore uses the mixed groups G0+G4 and G1+G3. An alternative setting that keeps all five grades strictly independent is studied as a side check in Appendix E.3.

L-GATr-slim (Petitjean et al., 2025) is a compact variant that retains only scalars (G0) and 4-vectors (G1) internally, explicitly dropping the outer product and G2, G3, and G4 channels. It matches L-GATr's task AUC while being $6\times$ faster (Petitjean et al., 2025).

2.3. LLoCa-Transformer

LLoCa-T (Spinner et al., 2025) is designed to enforce Lorentz equivariance through local canonicalization rather than through multivector-valued layers. It maps each jet to a canonical Lorentz frame and applies a standard transformer. Since the classification output is a Lorentz scalar, no inverse frame map is required. By design LLoCa-T is exactly Lorentz-equivariant: the canonicalization is itself an equivariant operation, so the full pipeline inherits exact equivariance. However, the canonicalization step is numerically more sensitive than the architectural approach of L-GATr, which likely accounts for the larger measured errors in our evaluation. LLoCa-T uses no multivector representation; its hidden state is a flat vector, so grade decomposition does not apply directly.

3. Experimental Setup

3.1. Dataset

TopTagging (Butter et al., 2018; Kasieczka et al., 2019) is a binary jet tagging benchmark with 1.2M training, 400k validation and 400k test jets at $\sqrt{s} = 14\,\text{TeV}$. Signal jets arise from hadronic top quark decays $t \to Wb \to qqb$ (three-pronged substructure); background jets are QCD-initiated single-prong jets. Constituent particles are ordered by transverse momentum p_T (descending).

3.2. Models

We study five TopTagging models with comparable classification performance (Table 1). L-GATr (Spinner et al., 2024) has 12 GATrBlock layers with $C_{\text{mv}} = 16$ and $C_{\text{s}} = 32$; L-GATr-slim (Petitjean et al., 2025) has 12 LGATrSlimBlock layers with $C_v = 32$ and $C_s = 96$; and LLoCa-T (Spinner et al., 2025) has 12 transformer blocks with Lorentz equivariance enforced by local canonicalization. These three models provide the symmetry-aware comparison set. We compare them to two non-equivariant baselines: Vanilla-T, a 10-block standard transformer using the same general architecture family as L-GATr, and ParT (Qu et al., 2022), a 10-layer kinematics-only ParticleTransformer checkpoint. ParT augments self-attention with learned pairwise physics biases, giving it a soft geometric prior without exact equivariance. All AUC values are 3-seed means.

Table 1. Test AUC on TopTagging (mean $\pm$ std across 3 seeds).

Model	AUC
L-GATr	0.9869 ± 0.0001
L-GATr-slim	0.9867 ± 0.0001
LLoCa-T	0.9867 ± 0.0001
ParT	0.9857 ± 0.0001
Vanilla-T	0.9856 ± 0.0001

3.3. Probe Methods

Equivariance test. For each model, we apply 5 random Lorentz transforms to 200 test jets and measure the mean relative change in the output logit. For L-GATr and L-GATr-slim, the transform is applied to the full embedded input, including beam-spurion/reference tokens, so the tested representation transforms consistently. We additionally perform a boost sweep with $\gamma \in \{1.0, 1.25, 1.5, 1.75, 2.0, 2.5, 3.0, 4.0, 5.0\}$ to probe behavior from the identity transform through large boosts; the range is capped at $\gamma \approx 5$, motivated by the dataset's hard rapidity cut of $|\eta_j| < 2$, which makes boosts well beyond $\gamma \sim 5$ poorly represented in the data. The output metric is $|\ell(\Lambda x) - \ell(x)|/(|\ell(x)| + \varepsilon)$ with $\varepsilon = 10^{-8}$. This relative-logit metric is a sensitive invariance diagnostic, not a calibrated performance metric.

Linear probes on h_{s}. We extract scalar channel representations at each layer for 10,000 TopTagging jets, then fit linear probes for 14 physics targets: jet mass, jet multiplicity, N-subjettiness values $\tau_1, \tau_2, \tau_3, \tau_{21}, \tau_{32}$ (Thaler & Van Tilburg, 2011), top/QCD classification and particle-level p_T, η, ϕ, E, p_T-quartile and ΔR. The main probe table and probe figures report combined 95% bootstrap intervals obtained by pooling draws across the 3 training seeds ($3 \times 200 = 600$ draws per target). Regression targets use Ridge regression; binary and multi-class targets use logistic regression. For particle-level probes, train/test separation is performed at the jet level.

Grade decomposition. We zero one grade group of h_{mv} at all layers simultaneously and measure the AUC drop (*zero-grade ablation*), and separately activate only one group while zeroing all others (*keep-only ablation*). Layer-resolved ablation zeros one grade group at one layer at a time. Under the default connected-subgroup setting (see Section 2), L-GATr enforces equivariance to the proper orthochronous Lorentz subgroup, so scalar and pseudoscalar components (G0+G4) may mix, as may vector and trivector components (G1+G3); we therefore report three groups: *scalar-like* (G0+G4), *vector-like* (G1+G3) and *bivector* (G2). All ablations are applied to hidden multivector outputs after the corresponding layer module, not to the raw input embedding.

Geometric-algebra-invariant probes on h_{mv}. We construct 832 Lorentz-invariant scalar features per token from h_{mv} via the geometric-algebra inner product $\langle \tilde{X} Y \rangle_0$ across grade pairs, where $\tilde{\cdot}$ denotes reversion. Features include: G0 and G4 direct components (32 features), G1 pairwise inner products (136), G3 pairwise inner products (136), G1 $\times$ G3 cross terms (256) and G2 scalar and pseudoscalar invariants (136+136). Implementation details are in Appendix D.

4. Equivariance Validation

Figure 1 shows the mean relative logit change under pure boosts across all five models, with combined 95% bootstrap CIs pooled across 3 seeds; full combined-bootstrap intervals for random transforms are reported in Appendix F. Under random Lorentz transforms, both L-GATr and L-GATr-slim achieve logit errors below 1% (bootstrap medians 0.52% and 0.39% respectively), with overlapping seed ranges, so the gap between the two equivariant models is within training variability and should not be over-interpreted. LLoCa-T shows a larger measured error ($\approx 13\%$) than the architectural models, consistent with its canonicalization being more numerically sensitive than the architectural approach. ParT and Vanilla-T exhibit logit changes of $\approx 334\%$ and $\approx 120\%$ respectively, roughly two to three orders of magnitude larger

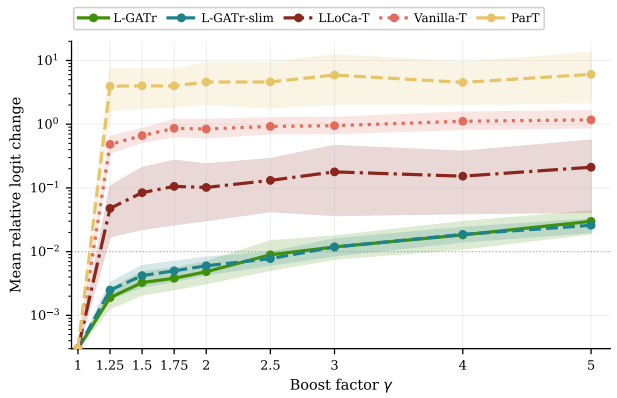

Figure 1. Mean relative logit change under pure boosts versus γ across five models, using a sweep from $\gamma = 1$ to $\gamma = 5$. Shaded bands: combined 95% bootstrap CIs pooled across 3 seeds. Lower is more equivariant.

than the equivariant models. We treat this experiment as validation rather than a main result: it confirms that the trained models preserve the intended symmetries closely enough for the representation-level analyses below.

Both L-GATr and L-GATr-slim remain below 1% logit error at moderate boosts ($\gamma \lesssim 2$). Their errors increase as the boost grows, consistent with accumulated floating-point error at more extreme boosts. LLoCa-T remains much more stable than the non-equivariant baselines, but has wider and larger measured errors than L-GATr/L-GATr-slim, again reflecting greater numerical sensitivity of the canonicalization. ParT and Vanilla-T degrade sharply even at moderate boosts.

5. Physical Representations

5.1. Scalar Channel Probes

Table 3 (Appendix C) and Figure 2 report final-layer linear probe scores for all five models as combined-bootstrap medians. Several consistent patterns emerge.

Pseudorapidity η as an invariance probe. All three symmetry-aware models suppress particle-level η to approximately zero at the final layer (L-GATr: -0.002; L-GATr-slim: 0.000; LLoCa-T: -0.006), while Vanilla-T retains 0.151 and ParT retains 0.287. The particle-level ϕ probe is near zero for all models, largely because inputs use $\Delta\phi$ relative to the jet axis rather than absolute azimuth. The η result is a representation-level signature of the imposed symmetry: in the architecturally equivariant models this absence is largely enforced by the scalar channel construction, while LLoCa-T shows a similar empirical pattern through its canonicalization-based representation. The comparison to Vanilla-T and ParT is the non-trivial part: ordinary transformers retain linearly accessible rapidity information.

LLoCa-T makes physics observables most linearly accessible. Despite its larger equivariance error, LLoCa-T achieves the highest final-layer probe scores for jet mass (0.993), jet multiplicity (0.971), τ_1, τ_2, τ_3 and τ_{21} (0.609) of all five models. This suggests that probe accessibility reflects how linearly the model encodes physics: canonicalization may organize representations in a way that is more linearly decodable in this probe basis, despite its larger measured equivariance error.

N-subjettiness and ratio probes. All models encode individual N-subjettiness values τ_1, τ_2, τ_3 strongly ($R^2 \approx 0.84$–0.97), consistent with these being well-defined jet substructure observables (Thaler & Van Tilburg, 2011). However, the ratio $\tau_{21} = \tau_2/\tau_1$ and the top-tagging-motivated ratio $\tau_{32} = \tau_3/\tau_2$ are substantially weaker for all models. This is consistent with Cheng (2019): networks encode the building blocks τ_N more accessibly than their ratios, a pattern that persists across all five architectures studied here.

Full L-GATr makes jet-level observables more linearly accessible. Despite comparable task AUC, L-GATr achieves higher final-layer probe scores than L-GATr-slim for jet mass (0.983 vs. 0.953), jet multiplicity (0.923 vs. 0.827) and τ_3 (0.809 vs. 0.766). This advantage is clearest for jet-level substructure targets; several particle-level kinematic probes are higher for L-GATr-slim.

L-GATr-slim self-corrects input leakage. L-GATr-slim takes $\Delta\eta$ as a scalar input feature and its particle-η probe starts above zero at early layers before decaying to ≈ 0 by the final layer, consistent with progressively suppressing linearly accessible non-invariant information in the probed scalar channels.

Figure 3 shows layerwise probe trajectories for η, τ_{32}, and jet mass. The η suppression is visually immediate: equivariant models remain near zero throughout depth, while ParT and Vanilla-T retain substantial non-zero signal.

5.2. Multivector Invariant Probes

We fit linear probes on the 832 Lorentz-invariant features from h_{mv} at each layer of L-GATr (Section 3). Two findings inform the grade discussion below.

Multiple usable pathways in the final layer. At the final layer, probes on scalar-like invariants (G0+G4 components, 32 features) and vector-like invariants (G1+G3 inner products and cross terms, 528 features) both approach full-model classification AUC: scalar-like achieves ≈ 0.986 and vector-like achieves ≈ 0.982–0.986 under the combined-bootstrap summary, matching the task AUC. This is consistent with L-GATr encoding discriminative information through both the vector-like (G1+G3) pathway and the scalar-like (G0+G4)

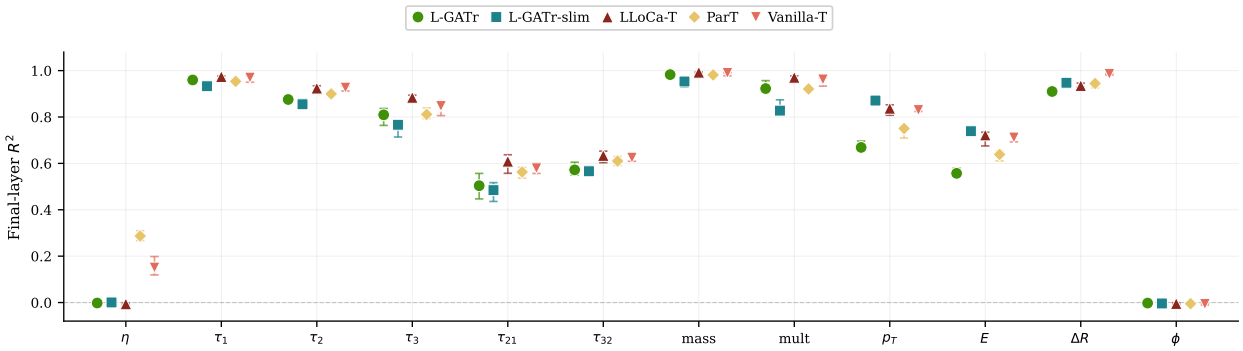

Figure 2. Final-layer linear probe scores for selected targets across five models. Points: combined-bootstrap medians; error bars: 95% CIs pooled across 3 seeds. Targets shown: particle η, the N-subjettiness family, jet-level mass and multiplicity, then particle kinematics; particle ϕ is placed last as it is near zero for all models. Top/QCD classification and particle p_T-rank quartile are reported in Table 3; full per-layer trajectories are in Appendix C.

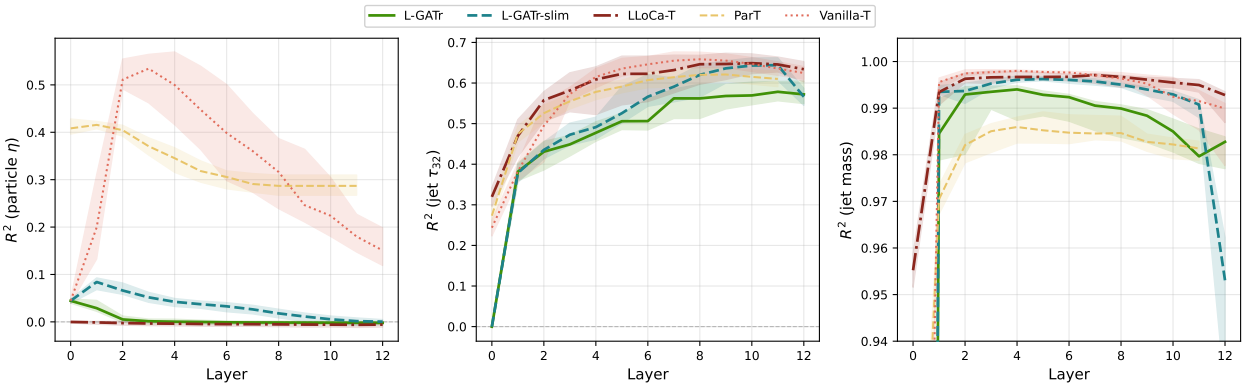

Figure 3. Scalar-channel probe trajectories across all five models. Shaded bands: combined 95% bootstrap CIs pooled across 3 seeds. **Left**: particle η — all three symmetry-aware models remain near zero throughout depth; ParT and Vanilla-T retain substantial signal. **Centre**: τ_{32} is consistently harder to decode than individual τ_N values across all architectures. **Right**: jet mass — LLoCa-T achieves the highest final-layer score. Full probe trajectories for all targets are in Appendix C.

pathway, as reflected also in the high cross-seed variance of vector-like ablations in Section 6.

η is absent from invariant features. The particle η probe on all 832 invariant features gives $R^2 \approx -0.025$ (seed mean, negative for all seeds), confirming that the Lorentz-invariant features carry no frame-dependent rapidity information. The particle ϕ probe is near zero across all layers ($R^2 \approx 0$).

τ_{21} more accessible from h_{mv} than h_{s}. The full 832-feature probe achieves $R^2 \approx 0.648$ for τ_{21} at the final layer, compared to 0.504 from h_{s} alone (Table 3). The ratio appears to be encoded in the inter-grade structure of h_{mv} but less linearly accessible from the scalar channels. Full per-layer trajectories are in Appendix D.

6. Grade Structure

6.1. Grade Ablations: Grouped Pathways and Bivector Redundancy

Table 2 reports grouped grade-ablation results for L-GATr on TopTagging, summarized across three seeds. Figure 4 plots the same zero-grade results and shows how the zero-grade intervention changes with depth.

Bivectors are not load-bearing in this setting. Zeroing all bivector (G2) channels at every layer reduces AUC by only 0.001, the smallest effect of any grade group and consistent with zero. This finding is seed-robust: no seed shows G2 ablation impact above 0.005. This does not imply that bivectors are never useful in L-GATr; rather, for these TopTagging checkpoints the task information is recoverable through the scalar-like and vector-like pathways under the default connected-subgroup parameterization.

Table 2. Grade decomposition ablations for L-GATr on TopTagging (default connected-subgroup setting). Entries give combined-bootstrap medians with asymmetric 95% intervals; the left panel of Figure 4 plots the zero-grade column.

GROUP	ZERO ΔAUC	KEEP-ONLY ΔAUC
SCALAR-LIKE (G0+G4)	$0.118\ ^{+0.056}_{-0.082}$	$0.239\ ^{+0.577}_{-0.027}$
VECTOR-LIKE (G1+G3)	$0.239\ ^{+0.577}_{-0.027}$	$0.149\ ^{+0.014}_{-0.099}$
BIVECTOR (G2)	$0.001\ ^{+0.004}_{-0.001}$	$0.337\ ^{+0.060}_{-0.242}$

Vector-like channels are dominant but seed-variable. Zeroing vector-like (G1+G3) channels gives $\Delta\mathrm{AUC} = 0.239$, the largest zero-ablation effect, but with very high cross-seed variance (seed values: 0.809, 0.239, 0.219). This variability suggests that different training runs may allocate task information differently between the vector-like (G1+G3) pathway and the scalar-like (G0+G4) pathway, as discussed in Section 5.2. More seeds would be needed to characterize this pathway preference definitively.

Scalar-like channels remain competitive. Zeroing scalar-like (G0+G4) channels gives $\Delta\mathrm{AUC} = 0.118$. The keep-only view is complementary: keeping only scalar-like channels gives $\Delta\mathrm{AUC} = 0.239$, matching the vector-like zero ablation. Taken together with the invariant-probe results, this supports redundancy between scalar-like and vector-like pathways.

Note on keep-only bivector. Keeping only G2 channels gives $\Delta\mathrm{AUC} = 0.337$ (Table 2), which may appear in tension with the G2$\approx$0 zero-ablation result. The keep-only intervention does not isolate standalone bivector capacity: the scalar branch remains available to the readout, so the remaining performance can still use scalar information. We therefore interpret the near-zero G2 zero-ablation as the cleaner evidence that bivectors are not load-bearing here.

6.2. Layer-Resolved Ablation

Figure 4 shows the global and layer-resolved grade ablations for L-GATr on TopTagging. Several patterns are clear.

Vector-like ablation impact is largest at the input stage (mean $\Delta\mathrm{AUC} = 0.422$) and decays monotonically with depth. Scalar-like impact is smaller throughout ($\Delta\mathrm{AUC} = 0.091$ at the input stage, near zero after the first block). Bivector ablation is negligible at every individual layer ($\Delta\mathrm{AUC} < 0.002$ at all depths). All grade-group ablations converge to $\Delta\mathrm{AUC} \approx 0$ at the final layer by architectural necessity: the grade-preserving output projection can only read G0 signal from the final layer, regardless of what other grades encode. Seed 1001 shows substantially larger input-stage vector-like impact (0.809) than seeds 1002/1003 ($\approx$0.28), consistent with the seed-variable global ablation result and the multiple-pathway hypothesis.

6.3. Training without Bivectors

To verify that the G2 ≈ 0 finding is not an artifact of the trained network suppressing bivectors post-hoc, we also study a model trained with bivectors architecturally disabled (G2 channel zeroed during training). This model achieves baseline AUC 0.9865 ± 0.0002, statistically indistinguishable from the standard L-GATr (0.9867 ± 0.0003). Its vector-like $\Delta\mathrm{AUC}$ (0.216) is comparable to seeds 1002/1003 of the standard model. Thus, architecturally removing G2 during training has negligible effect on task performance. Full grade ablations for this model are in Appendix E.1. A complementary check with the alternative subgroup setting (five independent grades rather than three mixed groups) likewise finds G2 negligible and further shows that parity-odd grades (G3, G4) are not load-bearing; full tables are in Appendix E.3.

6.4. L-GATr-slim Contrast

L-GATr-slim retains only G0 scalars and G1 4-vectors and has no G2, G3, or G4 channels. Its ablation pattern differs sharply from L-GATr: zeroing all scalar channels reduces AUC to 0.500 ($\Delta\mathrm{AUC} = 0.486$), a complete loss of discriminative power, while zeroing the 4-vector channels gives $\Delta\mathrm{AUC} = 0.585$. Full component-level ablations are in Appendix E.2. The scalar branch is indispensable for L-GATr-slim because the readout is the global scalar token directly.

In contrast, L-GATr shows a more distributed pattern: vector-like channels are the dominant but seed-variable ablation target, scalar-like channels remain competitive, and the higher-order grades omitted by L-GATr-slim do not appear necessary for TopTagging. Both architectures achieve comparable task AUC, suggesting that grade structure reflects representational strategy rather than task requirement.

7. Discussion and Conclusion

We have shown that equivariance tests, scalar-channel probes, grade decomposition and geometric-algebra-invariant probes expose physically structured internal representations in Lorentz-equivariant jet taggers. The output-level checks validate that the symmetry constraints are active in the trained models, while the probe results show how this symmetry appears inside the representations: symmetry-aware models suppress linearly accessible pseudorapidity while retaining strong access to jet mass and N-subjettiness.

Probe accessibility and task use are distinct. LLoCa-T achieves the highest linear probe scores for most physics targets, despite larger measured numerical errors in the equivariance test. This suggests that linear decodability is best

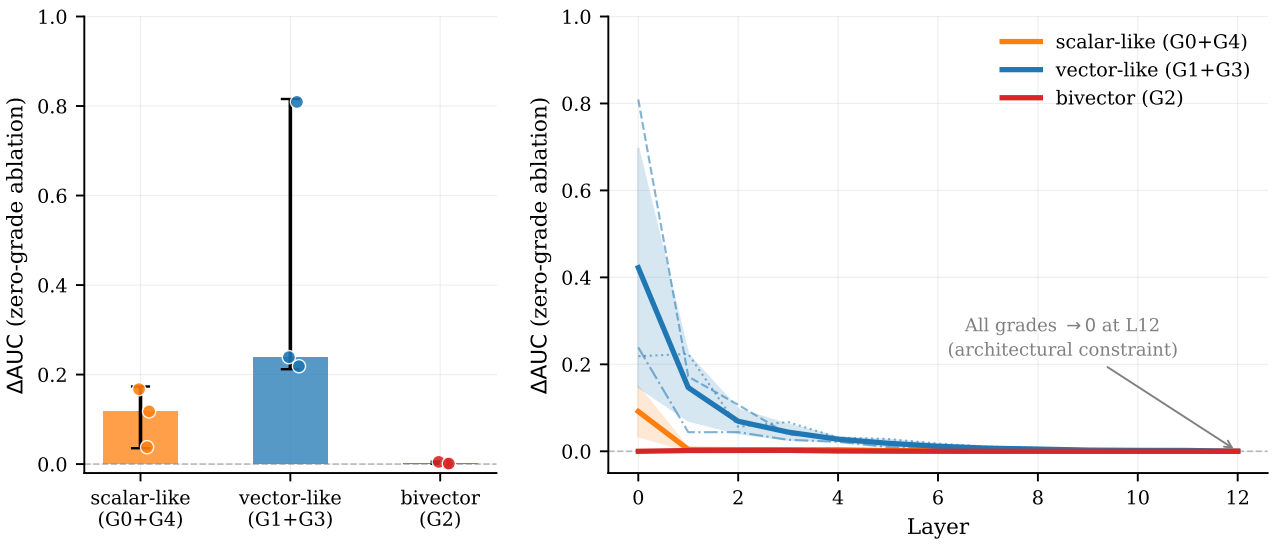

Figure 4. Grade decomposition for L-GATr on TopTagging. **Left**: global zero-grade ablation $\Delta\mathrm{AUC}$ for each grade group. Bars show combined 95% bootstrap CIs pooled across 3 seeds; dots show per-seed values. Bivector (G2) is consistently negligible; vector-like (G1+G3) is dominant but highly seed-variable, reflecting multiple usable pathways. **Right**: layer-resolved zero-grade ablation (seed mean $\pm$ std; thin lines show per-seed vector-like curves). Vector-like impact is largest at the input stage and decays monotonically; bivector is negligible at every layer. All grade groups converge to $\Delta\mathrm{AUC} \approx 0$ at the final layer by architectural necessity (grade-preserving output projection).

interpreted as a property of representation organization, not as a direct measure of task utility. The contrast between L-GATr and LLoCa-T is therefore informative: both suppress η, but their internal representations expose different physics observables to simple linear readouts.

Grade structure reflects representational strategy. For L-GATr, the clearest grade result is negative but useful: G2 bivectors are not load-bearing for these TopTagging checkpoints. They are negligible under zero-ablation, remain negligible in the five-grade side study and can be removed during training without degrading AUC. The more open result is the seed-variable importance of vector-like channels. Together with the invariant-probe evidence, this suggests that independently trained L-GATr models may distribute task information differently across scalar-like and vector-like pathways, although more seeds are needed to characterize that preference.

Limitations. Linear probes measure accessibility of information, not necessarily how the model uses it internally. Grade ablations are interventions on trained networks and may introduce distribution shift, especially for keep-only settings. The bivector result is consistent across 3 independent runs and the training-time G2-disabled check, but the seed-variable vector-like pathway still warrants additional trainings. Our analysis is limited to TopTagging.

Future directions. A fuller JetClass analysis, including per-class grade structure and h_s probes, would test whether grade patterns are task-dependent. Adapting mechanistic interpretability methods such as activation patching, circuit analysis or sparse autoencoders to geometric-algebra representations is an open opportunity. Probing whether LLoCa-T's high linear-probe accessibility translates to more interpretable intermediate representations is another promising direction.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. We gratefully acknowledge the computational resources provided by BITS Pilani that made this work possible.

Impact Statement

This paper presents work whose goal is to advance the interpretability of symmetry-constrained neural networks for particle physics applications. We do not anticipate specific negative societal consequences from this work.

References

Brehmer, J., de Haan, P., Behrends, S., and Cohen, T. Geometric Algebra Transformer. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2305.18415.

Butter, A., Kasieczka, G., Plehn, T., and Russell, M. Deep-learned Top Tagging with a Lorentz Layer. *SciPost Physics*, 5:028, 2018. arXiv:1707.08966.

Cheng, T. Interpretability Study on Deep Learning for Jet Physics at the Large Hadron Collider. In *Machine Learning and the Physical Sciences Workshop, NeurIPS*, 2019. arXiv:1911.01872.

Esmail, W., Hammad, A., and Nojiri, M. IAFormer: Interaction-Aware Transformer network for collider data analysis. *SciPost Physics*, 20:108, 2026. arXiv:2505.03258.

Kasieczka, G., Plehn, T., Butter, A., Cranmer, K., Debnath, D., Dillon, B. M., et al. The Machine Learning Landscape of Top Taggers. *SciPost Physics*, 7:014, 2019. arXiv:1902.09914.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019. arXiv:1905.00414.

Legge, T., Wang, A., Ortiz, J., Limouzi, V., Zhao, Z., Gandrakota, A., Khoda, E. E., Ngadiuba, J., Duarte, J., and Cavanaugh, R. Why Is Attention Sparse In Particle Transformer? In *Machine Learning and the Physical Sciences Workshop, NeurIPS*, 2025. arXiv:2512.00210.

Petitjean, A., Plehn, T., Spinner, J., and Köthe, U. Economical Jet Taggers – Equivariant, Slim, and Quantized. *arXiv preprint arXiv:2512.17011*, 2025. arXiv:2512.17011.

Qu, H., Li, C., and Qian, S. Particle Transformer for Jet Tagging. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18281–18292, 2022. arXiv:2202.03772.

Spinner, J., Bresó, V., de Haan, P., Plehn, T., Thaler, J., and Brehmer, J. Lorentz-Equivariant Geometric Algebra Transformers for High-Energy Physics. In *Advances in Neural Information Processing Systems*, volume 38, 2024. arXiv:2405.14806.

Spinner, J., Favaro, L., Lippmann, P., Pitz, S., Gerhartz, G., Plehn, T., and Hamprecht, F. A. Lorentz Local Canonicalization: How to Make Any Network Lorentz-Equivariant. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.20280.

Thaler, J. and Van Tilburg, K. Identifying Boosted Objects with N-subjettiness. *Journal of High Energy Physics*, 2011:015, 2011. arXiv:1011.2268.

Wang, A., Gandrakota, A., Ngadiuba, J., Sahu, V., Bhatnagar, P., Khoda, E. E., and Duarte, J. Interpreting Transformers for Jet Tagging. In *Machine Learning and the Physical Sciences Workshop, NeurIPS*, 2024. arXiv:2412.03673.

Appendix

A. Attention Maps

Figures 5–8 show mean final-layer attention maps on TopTagging for all three Lorentz-aware models (L-GATr, L-GATr-slim, LLoCa-T), truncated to the top-50 constituents by p_T. Figure 5 gives an overview of the head-averaged maps across all three models and all three training seeds. Figures 6, 7 and 8 break out all 8 individual attention heads per model.

Both L-GATr and L-GATr-slim show a strong left-column pattern at position $j = 0$ in most heads, indicating that many heads attend heavily to the highest-p_T constituent. L-GATr tends to have sharper particle-0 focus in several heads (Head 2 across all seeds), while L-GATr-slim shows a somewhat broader spread. LLoCa-T displays a qualitatively different pattern: the head-averaged map is more diffuse and lacks the sharp left-column feature, reflecting the different representational strategy of canonicalization-based equivariance. These patterns are consistent across all three seeds.

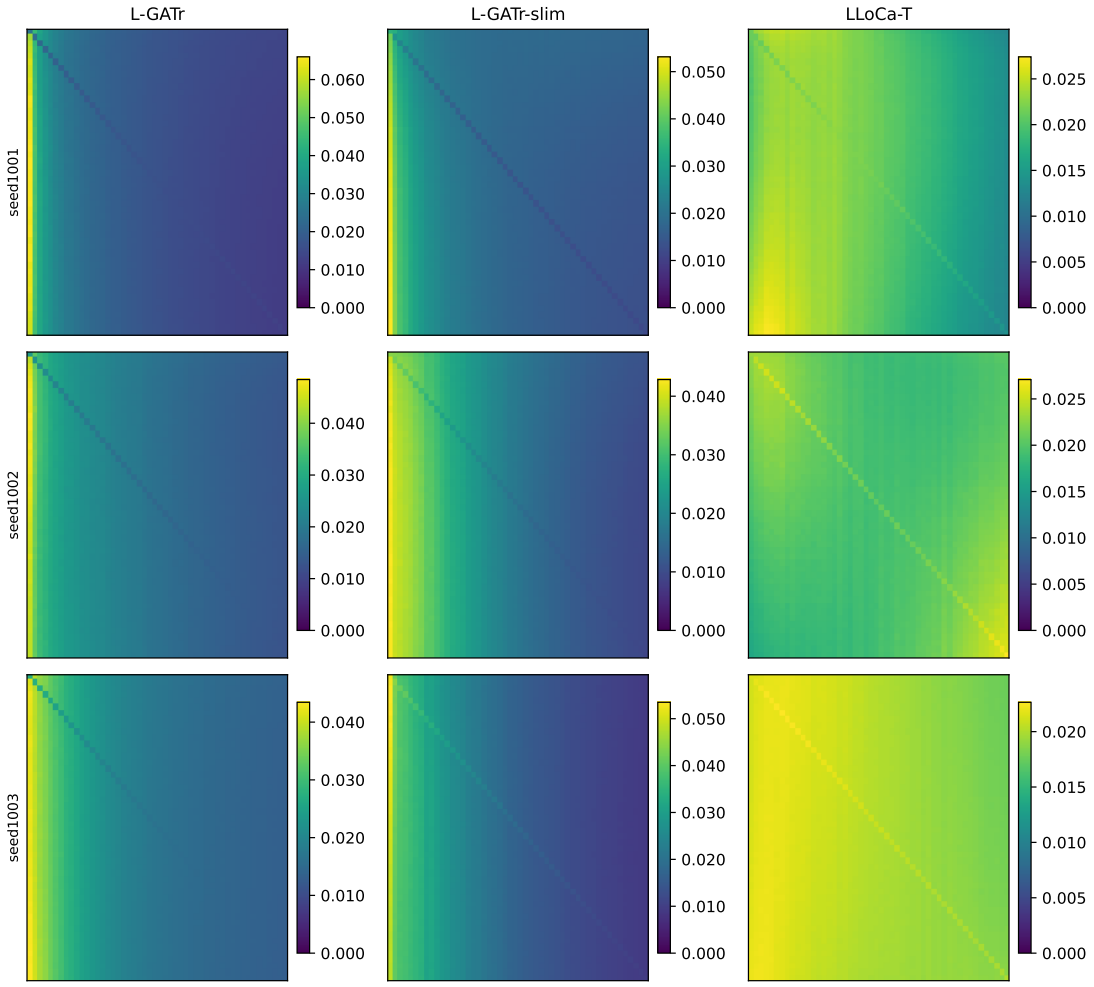

Figure 5. Overview: head-averaged mean final-layer attention maps for L-GATr (left), L-GATr-slim (centre) and LLoCa-T (right), across all three training seeds (rows). L-GATr and L-GATr-slim both show strong attention to the leading-p_T constituent; LLoCa-T displays a more diffuse pattern.

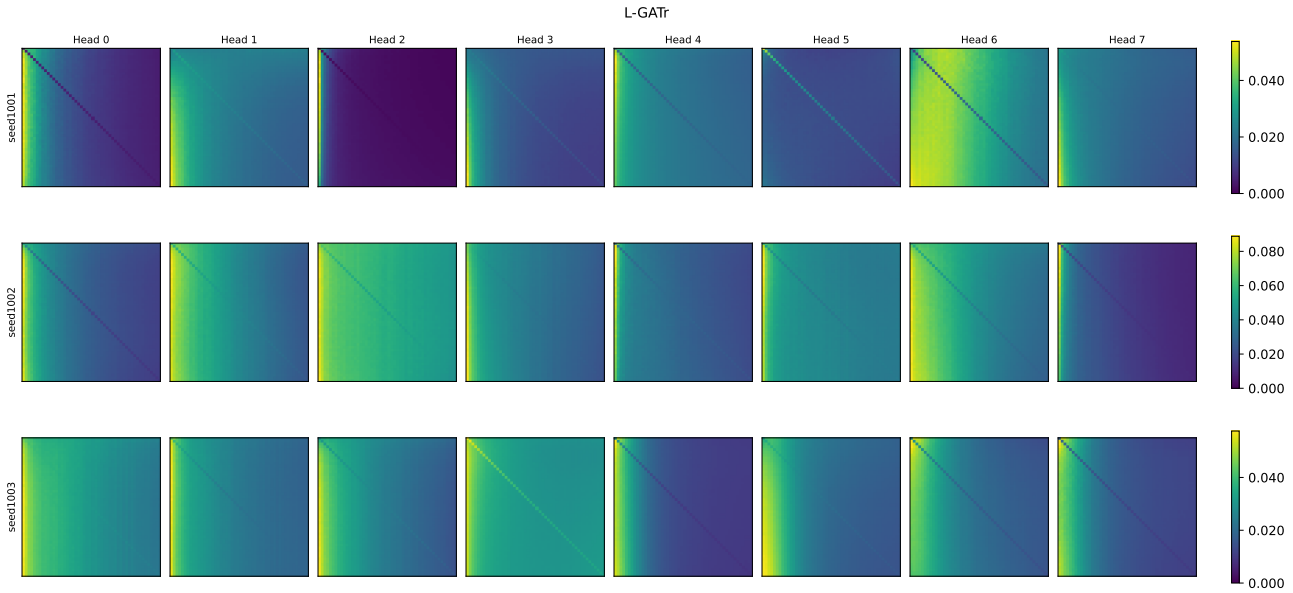

Figure 6. Individual attention heads (0–7) for L-GATr, across all three training seeds (rows). Head 2 shows an especially sharp particle-0 focus; other heads display varying degrees of leading-p_T attention.

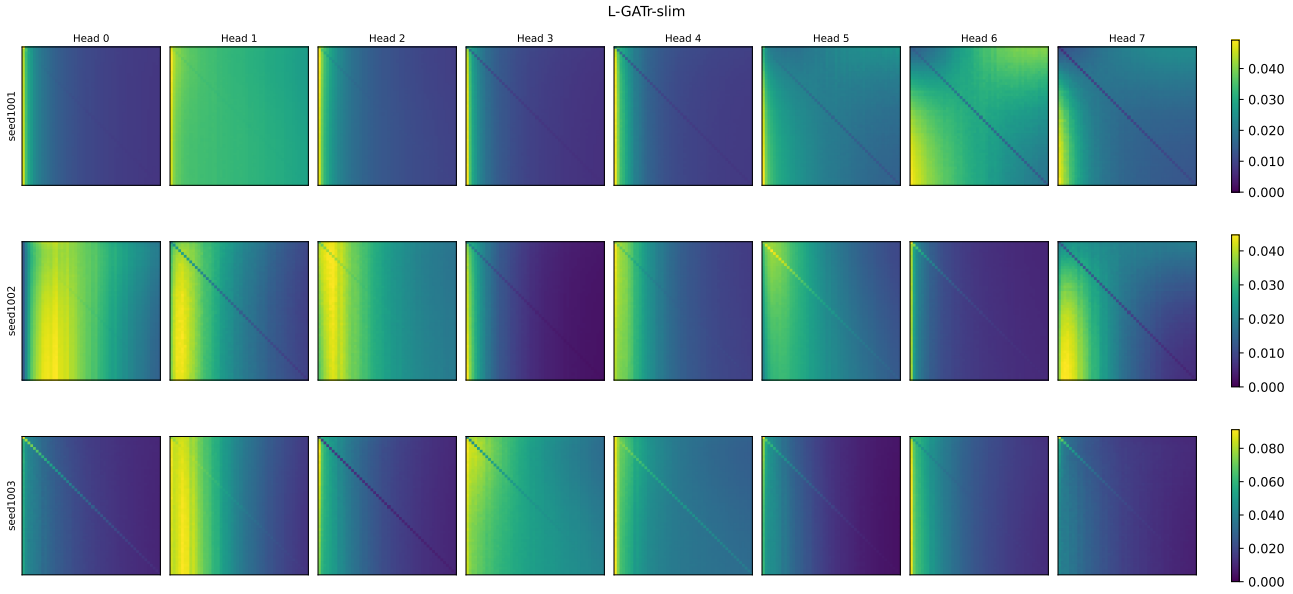

Figure 7. Individual attention heads (0–7) for L-GATr-slim, across all three training seeds (rows). Pattern is broadly similar to L-GATr but with a somewhat broader spread across constituents.

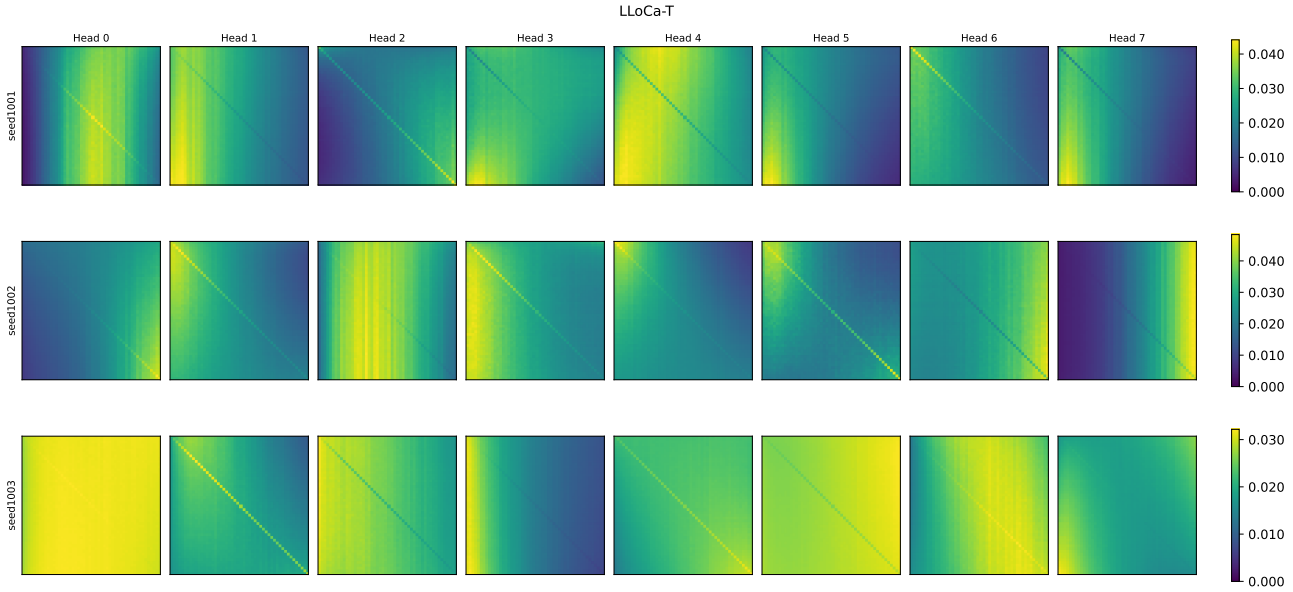

Figure 8. Individual attention heads (0–7) for LLoCa-T, across all three training seeds (rows). Attention patterns are more spatially diffuse than L-GATr/L-GATr-slim, reflecting the different internal structure of canonicalization-based equivariance.

B. CKA Representation Similarity

We compute linear Centered Kernel Alignment (CKA (Kornblith et al., 2019)) on scalar-channel representations as a supporting diagnostic for how representations evolve with depth. Unlike the probe and ablation analyses in the main text, these plots are descriptive: they show where layers become similar or reorganise, but they do not by themselves identify which physics observables are encoded.

We compute linear CKA between scalar channel representations h_{s} at all pairs of layers, for L-GATr, L-GATr-slim and LLoCa-T (5,000 jets, optimised Gram matrix approach). The matrices use transformer block outputs L0–L11 (12 layers), with axes oriented so that $(L0, L0)$ is at the bottom-left and $(L11, L11)$ at the top-right. Figure 9 shows the 12×12 CKA matrices for all three models and all three training seeds. All three models show high adjacent-layer similarity and lower, but still substantial, early–late similarity. Thus the scalar representations evolve gradually with depth rather than undergoing a sharp reorganization at a single layer; LLoCa-T and L-GATr are slightly more globally self-similar than L-GATr-slim in these CKA diagnostics.

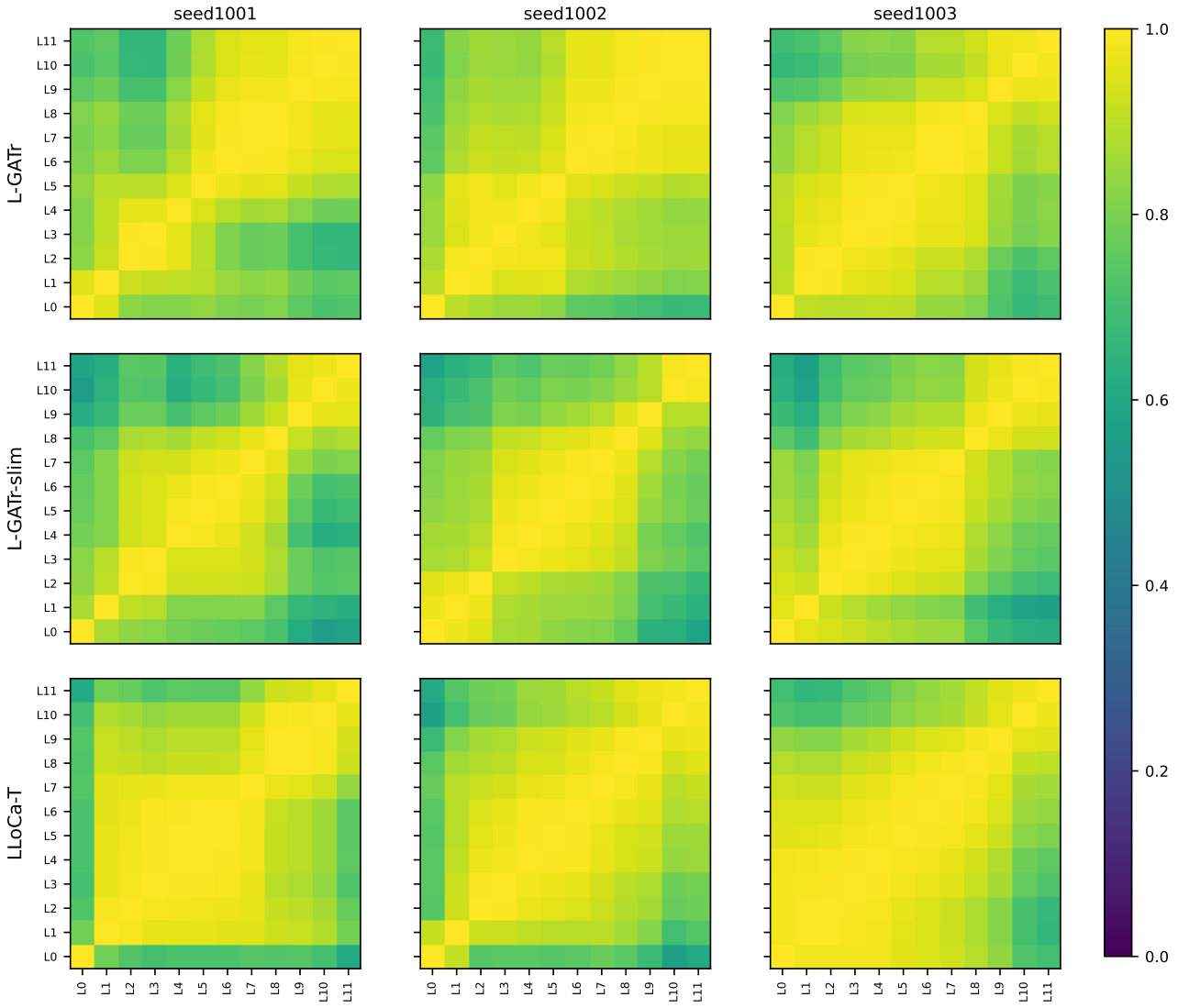

Figure 9. Linear CKA between scalar-channel representations at all layer pairs (L0–L11) for L-GATr (top row), L-GATr-slim (middle row) and LLoCa-T (bottom row), across all three training seeds (columns). Axes: L0 at bottom-left, L11 at top-right. Colorscale shared across all panels $[0, 1]$.

C. Scalar Channel Probes

Probe training details. For the TopTagging h_s probes, we work with 10,000 jets and fit probes at every layer. Regression targets use Ridge regression with $\alpha = 1.0$; binary and multi-class targets use logistic regression with L2 regularization (`lbfgs`, $C = 1.0$). For particle-level targets, train/test split is defined at the jet level.

Scalar-probe uncertainty. For scalar-channel probes, shaded bands and tabulated intervals use combined 95% bootstrap intervals obtained by pooling bootstrap draws across the 3 independently trained seeds. Each seed contributes 200 bootstrap draws per model, target and layer, for 600 pooled draws per reported interval. This convention summarizes both finite-sample variation within a checkpoint and training-seed variation across independently trained checkpoints.

Table 3. Full linear probe uncertainty summary at the final layer for all five models. Format: combined-bootstrap median [95% CI]. **Part** η row highlighted: ≈ 0 for all symmetry-aware models (L-GATr, slim, LLoCa-T), non-zero for non-equivariant models.

Probe	L-GATr	L-GATr-slim	LLoCa-T	Vanilla-T	ParT	Metric
top/QCD	0.986 [0.984, 0.989]	0.986 [0.984, 0.989]	0.986 [0.983, 0.988]	0.985 [0.982, 0.987]	0.984 [0.981, 0.987]	AUC
jet mass	0.983 [0.977, 0.984]	0.953 [0.929, 0.962]	0.993 [0.987, 0.994]	0.990 [0.978, 0.991]	0.981 [0.979, 0.983]	R^2
jet mult	0.923 [0.907, 0.957]	0.827 [0.816, 0.874]	0.971 [0.957, 0.977]	0.962 [0.933, 0.968]	0.921 [0.917, 0.925]	R^2
jet τ_1	0.960 [0.949, 0.967]	0.933 [0.916, 0.942]	0.974 [0.964, 0.977]	0.970 [0.950, 0.976]	0.954 [0.947, 0.957]	R^2
jet τ_2	0.875 [0.867, 0.888]	0.855 [0.837, 0.875]	0.924 [0.907, 0.935]	0.927 [0.912, 0.932]	0.900 [0.892, 0.905]	R^2
jet τ_3	0.809 [0.764, 0.837]	0.766 [0.714, 0.779]	0.885 [0.864, 0.894]	0.847 [0.806, 0.863]	0.811 [0.793, 0.839]	R^2
jet τ_{21}	0.504 [0.447, 0.557]	0.485 [0.436, 0.517]	0.609 [0.557, 0.637]	0.579 [0.557, 0.600]	0.563 [0.536, 0.583]	R^2
jet τ_{32}	0.572 [0.549, 0.605]	0.566 [0.545, 0.587]	0.634 [0.603, 0.653]	0.625 [0.609, 0.641]	0.610 [0.591, 0.630]	R^2
part p_T	0.669 [0.657, 0.696]	0.871 [0.853, 0.875]	0.837 [0.808, 0.852]	0.830 [0.824, 0.845]	0.750 [0.709, 0.757]	R^2
part η	-0.002 [-0.006, 0.000]	0.000 [-0.005, 0.005]	-0.006 [-0.011, -0.002]	0.151 [0.119, 0.198]	0.287 [0.267, 0.310]	R^2
part ϕ	-0.002 [-0.005, 0.000]	-0.004 [-0.008, -0.001]	-0.004 [-0.008, -0.001]	-0.006 [-0.011, -0.002]	-0.005 [-0.008, -0.003]	R^2
part E	0.557 [0.543, 0.579]	0.739 [0.722, 0.750]	0.722 [0.675, 0.734]	0.712 [0.693, 0.724]	0.639 [0.611, 0.653]	R^2
part quartile	0.763 [0.720, 0.800]	0.723 [0.688, 0.744]	0.760 [0.738, 0.807]	0.738 [0.702, 0.766]	0.803 [0.798, 0.806]	Acc.
part ΔR	0.910 [0.901, 0.922]	0.948 [0.936, 0.956]	0.935 [0.928, 0.946]	0.986 [0.981, 0.986]	0.945 [0.926, 0.947]	R^2

Figure 10 collects the remaining h_s probe trajectories not shown in the main text, including top/QCD, jet mass, jet multiplicity, τ_1, τ_2, τ_3, τ_{21}, particle p_T, particle E, particle ϕ, particle quartile and particle ΔR.

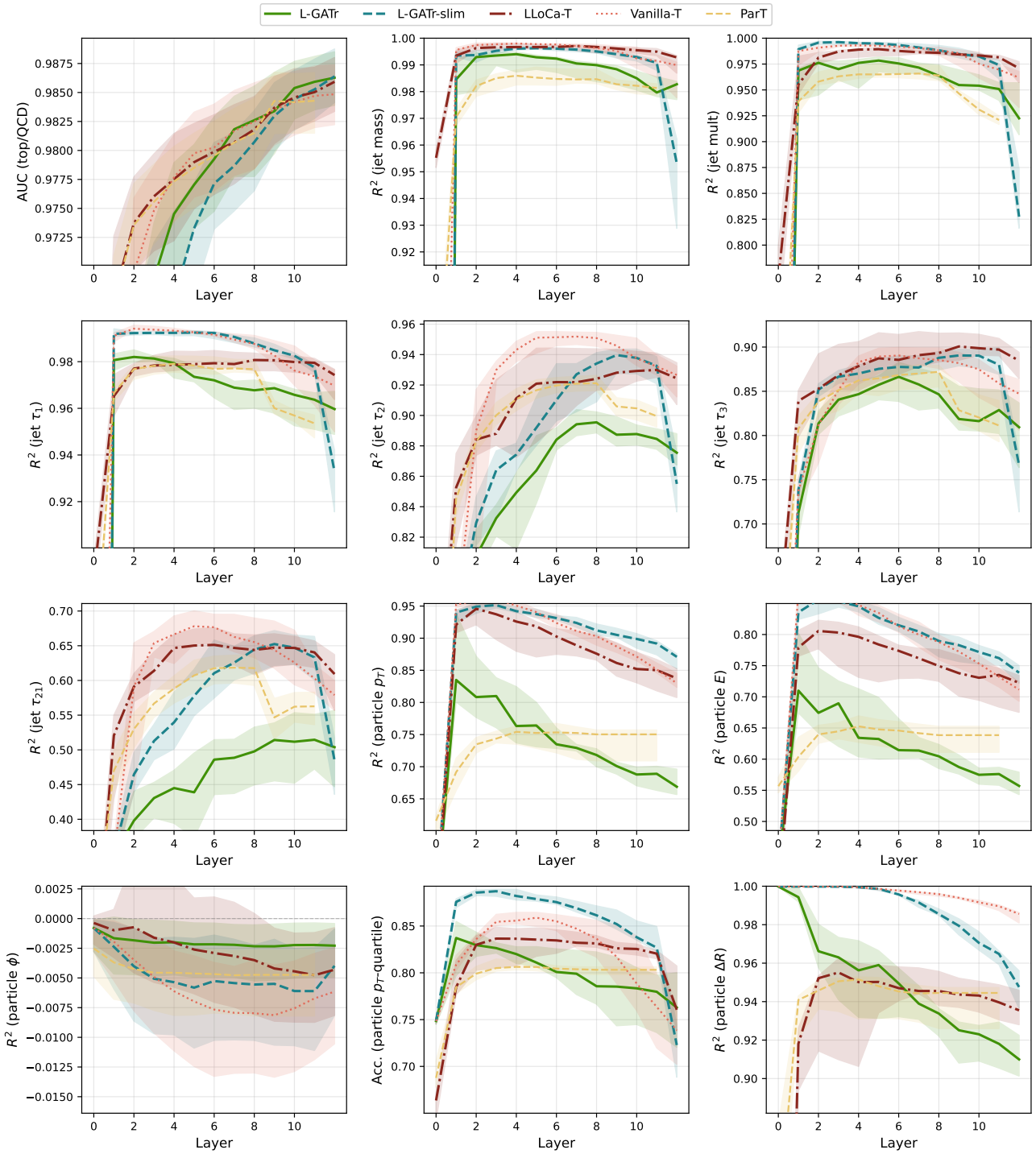

Figure 10. All remaining h_s probe trajectories across all five TopTagging models (4×3 grid; particle η and jet τ_{32} shown in the main text are omitted). **Row 1**: top/QCD AUC, jet mass, jet multiplicity. **Row 2**: τ_1, τ_2, τ_3. **Row 3**: τ_{21}, particle p_T, particle E. **Row 4**: particle ϕ, p_T-rank quartile, particle ΔR. Shaded bands: combined 95% bootstrap CIs pooled across 3 seeds for all models. Particle ϕ is near zero for all models, consistent with approximate azimuthal symmetry of the dataset.

D. Multivector Invariant Probes

Tables 4 and 5 report linear probe scores for all 832 Lorentz-invariant features at the final layer, broken down by grade group: scalar-like (G0+G4), vector-like (G1+G3), bivector (G2), and all 832 features combined. Jet-level entries use combined-bootstrap medians with 95% CIs pooled across 3 seeds (600 draws); h_{s} uses seed mean $\pm$ std. Particle-level entries use seed mean $\pm$ std throughout (3 seeds).

Jet-level probes (Table 4). The "All" column (all 832 features combined) matches or exceeds h_{s} for regression targets; classification AUC is essentially identical (All $= 0.985\,[0.982,\,0.987]$ vs. $h_{\mathrm{s}} = 0.987 \pm 0.000$). For τ_{21}, the full invariant set (All $= 0.648\,[0.581,\,0.680]$) notably exceeds h_{s} (0.494 ± 0.040), suggesting that τ_{21} information is encoded in the inter-grade structure of h_{mv} but less linearly accessible from h_{s} alone. The scalar-like and bivector groups have similar scores for classification (G0+G4 $= 0.986\,[0.984,\,0.989]$; from the combined bootstrap for individual grades: G0 $= 0.986\,[0.984,\,0.988]$, G4 $= 0.986\,[0.983,\,0.989]$), confirming that the model concentrates discriminative invariants into the G0 and G4 blade values.

Table 4. Geometric-algebra-invariant probe scores at the final layer for L-GATr on TopTagging, broken down by grade group and target. Metric: R^2 for regression, AUC for classification. Multivector columns: combined-bootstrap median [95% CI] (600 pooled draws across 3 seeds). h_{s}: seed mean $\pm$ std.

Target	h_{s}	scalar-like	vector-like	bivector	All	Metric
top/QCD	0.987 ± 0.000	$0.986\,[0.984,\,0.989]$	$0.982\,[0.979,\,0.985]$	$0.966\,[0.961,\,0.971]$	$0.985\,[0.982,\,0.987]$	AUC
jet mass	0.980 ± 0.004	$0.943\,[0.933,\,0.957]$	$0.983\,[0.955,\,0.989]$	$0.883\,[0.858,\,0.923]$	$0.989\,[0.980,\,0.993]$	R^2
jet mult	0.913 ± 0.017	$0.905\,[0.897,\,0.926]$	$0.950\,[0.840,\,0.958]$	$0.802\,[0.714,\,0.822]$	$0.974\,[0.921,\,0.977]$	R^2
jet τ_1	0.958 ± 0.006	$0.924\,[0.917,\,0.933]$	$0.958\,[0.937,\,0.962]$	$0.878\,[0.845,\,0.897]$	$0.971\,[0.964,\,0.974]$	R^2
jet τ_2	0.872 ± 0.007	$0.832\,[0.821,\,0.863]$	$0.914\,[0.865,\,0.922]$	$0.761\,[0.721,\,0.824]$	$0.925\,[0.906,\,0.938]$	R^2
jet τ_3	0.802 ± 0.022	$0.773\,[0.748,\,0.802]$	$0.861\,[0.802,\,0.875]$	$0.668\,[0.613,\,0.749]$	$0.893\,[0.864,\,0.899]$	R^2
jet τ_{21}	0.494 ± 0.040	$0.466\,[0.434,\,0.488]$	$0.597\,[0.510,\,0.626]$	$0.473\,[0.417,\,0.501]$	$0.648\,[0.581,\,0.680]$	R^2
jet τ_{32}	0.558 ± 0.010	$0.587\,[0.547,\,0.607]$	$0.616\,[0.548,\,0.643]$	$0.493\,[0.471,\,0.528]$	$0.658\,[0.626,\,0.677]$	R^2

Particle-level probes (Table 5). The η probe is near zero for all grade groups across all 3 seeds (All: -0.025 ± 0.006; h_{s}: -0.003 ± 0.002), confirming that the Lorentz-invariant features carry no substantial frame-dependent rapidity information. The ϕ probe is similarly near zero for all groups (All: -0.021 ± 0.002).

Table 5. Particle-level geometric-algebra-invariant probe scores at the final layer for L-GATr on TopTagging. Format: seed mean $\pm$ std (3 seeds). The η row is near zero for all grade groups (see Section 5.2).

Target	h_{s}	scalar-like	vector-like	bivector	All	Metric
part p_T	0.667 ± 0.019	0.544 ± 0.049	0.904 ± 0.028	0.481 ± 0.064	0.916 ± 0.021	R^2
part η	$\mathbf{-0.003\pm0.002}$	$\mathbf{-0.004\pm0.002}$	$\mathbf{-0.020\pm0.006}$	$\mathbf{-0.009\pm0.002}$	$\mathbf{-0.025\pm0.006}$	R^2
part ϕ	-0.002 ± 0.001	-0.002 ± 0.001	-0.013 ± 0.001	-0.006 ± 0.000	-0.021 ± 0.002	R^2
part E	0.588 ± 0.013	0.482 ± 0.068	0.849 ± 0.021	0.410 ± 0.058	0.866 ± 0.016	R^2
part quartile	0.761 ± 0.028	0.660 ± 0.027	0.790 ± 0.006	0.653 ± 0.030	0.819 ± 0.005	Acc.
part ΔR	0.912 ± 0.008	0.719 ± 0.018	0.864 ± 0.011	0.668 ± 0.071	0.926 ± 0.002	R^2

Figure 12 shows the corresponding particle-level invariant-probe trajectories; as in the main text, particle η and ϕ stay near zero across grade groups.

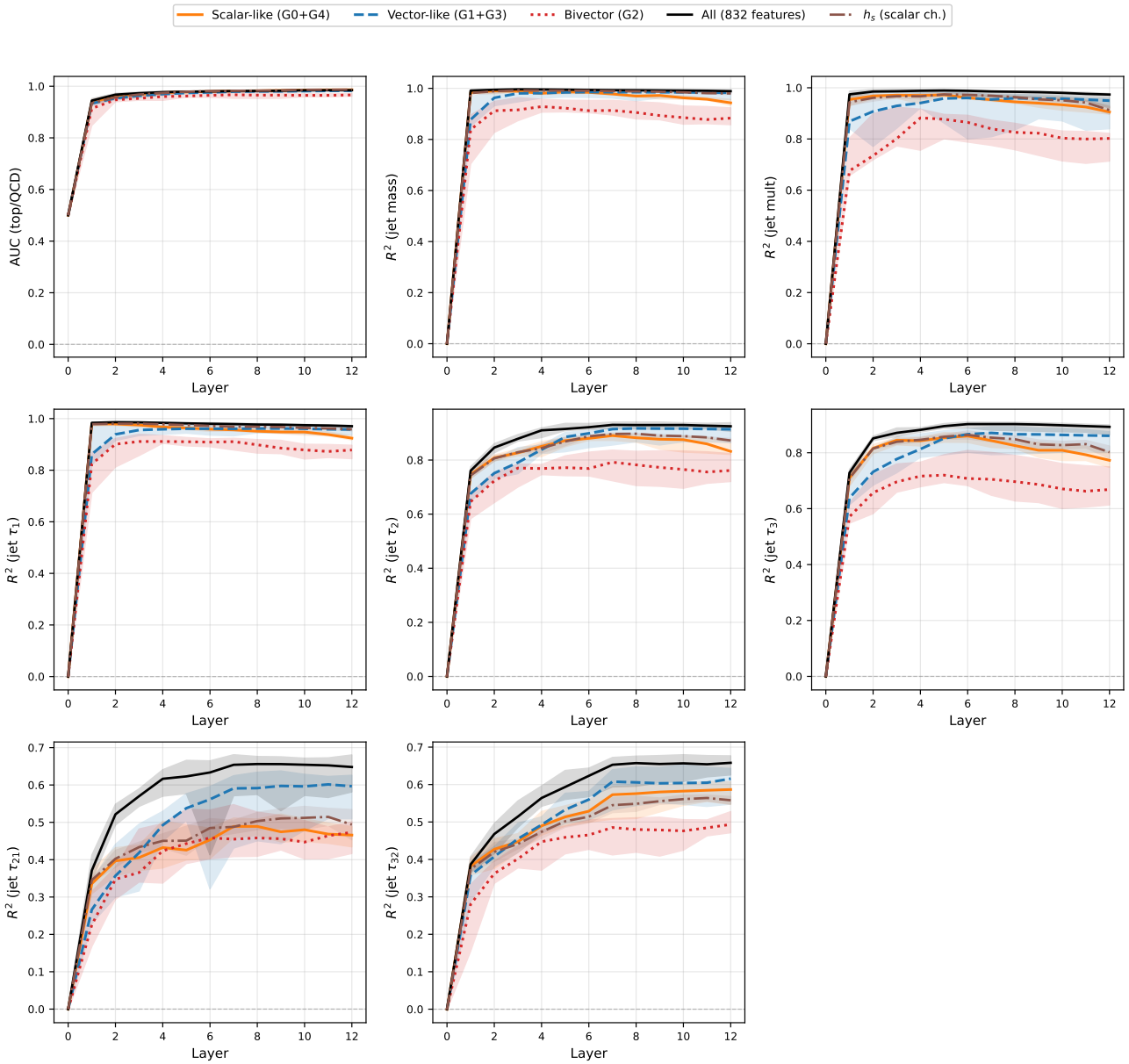

Figure 11. Per-layer geometric-algebra-invariant probe trajectories for all 8 jet-level targets (3×3 grid), broken down by grade group. Lines: scalar-like (G0+G4), vector-like (G1+G3), bivector (G2), all 832 features combined and h_s scalar-channel reference. Shaded bands: combined 95% bootstrap CIs for the multivector feature sets (600 pooled draws for jet-level targets); h_s uses seed mean $\pm$ std.

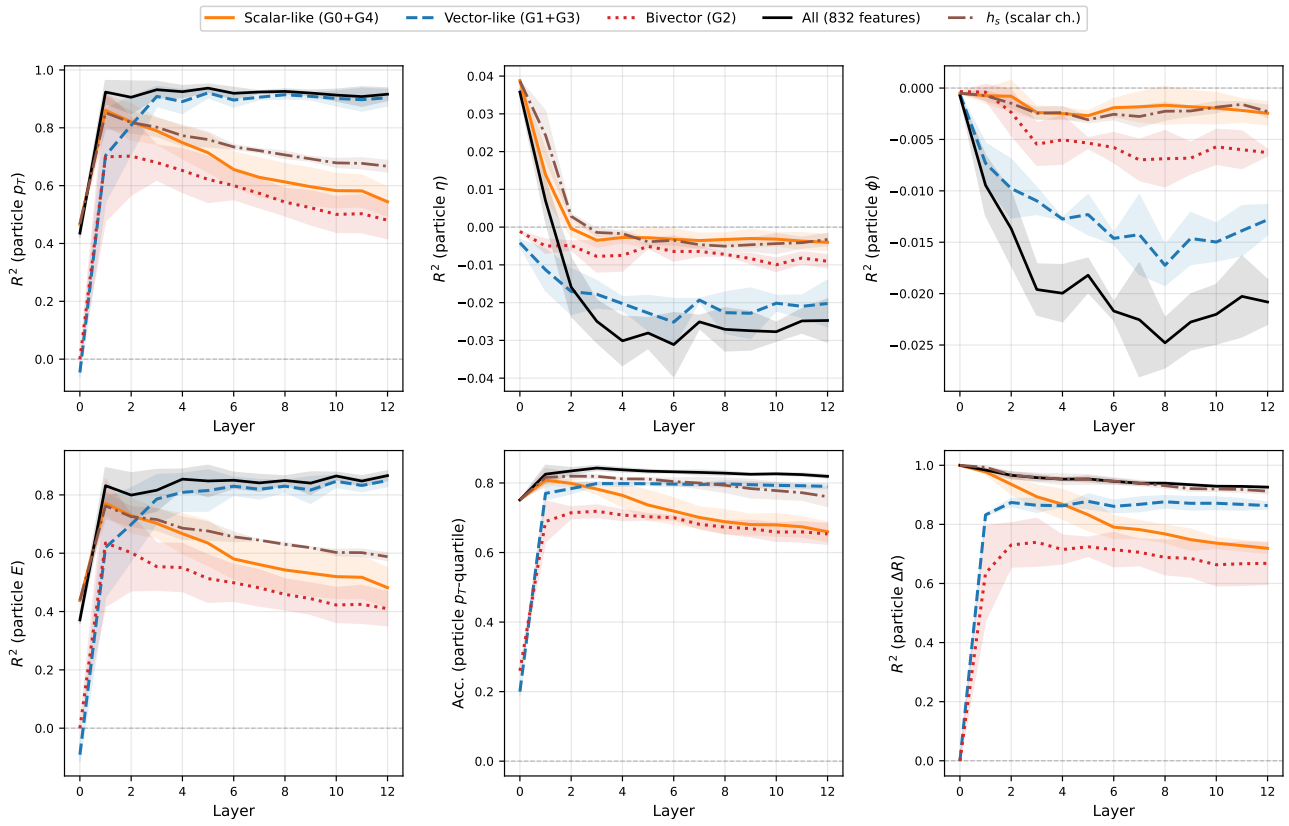

Figure 12. Per-layer geometric-algebra-invariant probe trajectories for all 6 particle-level targets (2×3 grid), broken down by grade group. Same line and color conventions as Figure 11. Shaded bands: seed mean $\pm$ std (3 seeds). Particle η and ϕ are near zero for all grade groups, confirming no frame-dependent information is encoded.

E. Grade Ablations

Ablation methodology. Zero-grade ablation zeros the selected multivector grade or grade group at the outputs of all 13 network stages (the input linear layer plus 12 transformer blocks). Keep-only ablation zeros all other grades or grade groups at those same hidden outputs. Layer-resolved ablation zeros one grade or grade group at exactly one stage output. These hooks operate on hidden multivectors rather than on the raw input embedding. In all figures below, bars show combined 95% bootstrap CIs pooled across 3 seeds (3 000 draws); overlaid dots show per-seed ΔAUC; the right panel shows seed mean $\pm$ std with per-seed thin lines for the dominant channel.

E.1. Bivector=False Confirmation

Figure 13 shows the grade ablation results for L-GATr trained with bivectors architecturally disabled (G2 channel zeroed during training). The model achieves baseline AUC 0.9865 ± 0.0002, indistinguishable from the standard L-GATr (0.9867 ± 0.0003). The bivector bar is trivially zero (G2 was never trained), while the vector-like ΔAUC per seed (0.216, 0.082, 0.399; mean 0.232 ± 0.131) shows the same high cross-seed variance seen in the standard model. The layer-resolved panel confirms that vector-like information again concentrates at the early layers and decays with depth.

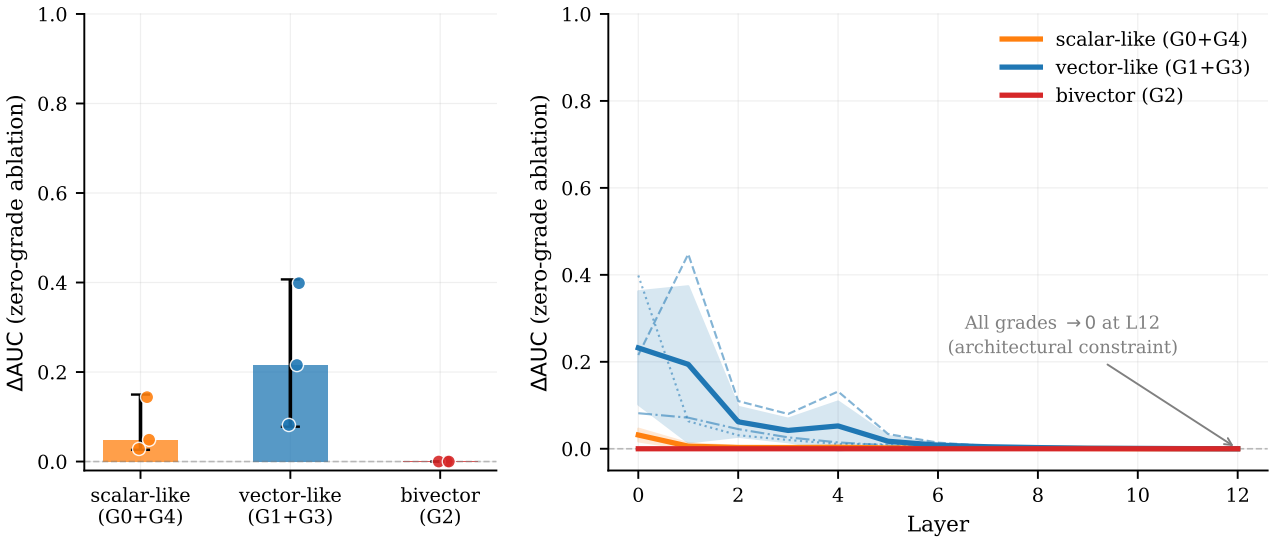

Figure 13. Grade ablations for L-GATr trained with bivectors architecturally disabled (3-group, G2 zeroed during training). **Left:** global zero-ablation ΔAUC with 95% combined bootstrap CIs (bars) and per-seed values (dots). **Right:** layer-resolved ΔAUC (seed mean $\pm$ std); thin lines show per-seed vector-like trajectories. Bivector bar is trivially zero; vector-like pattern mirrors the standard L-GATr seeds 1002/1003.

E.2. L-GATr-slim Channel Ablations

L-GATr-slim uses a different internal channel structure from L-GATr: it retains only G0 scalars (h_s) and G1 4-vectors (h_v), and its readout is the scalar channels at a global token. Table 6 reports channel ablations for L-GATr-slim on TopTagging (combined-bootstrap median [95% CI]; baseline AUC $= 0.986\,[0.985,\ 0.988]$). Zeroing all scalar channels reduces L-GATr-slim's AUC to 0.500 (chance, $\Delta\text{AUC} = 0.486\,[0.485,\ 0.488]$) because the readout is the global scalar token directly. Zeroing the 4-vector channels gives a much larger effect than previously reported ($\Delta\text{AUC} = 0.585\,[0.559,\ 0.695]$), indicating that the vector channels provide the geometric structure from which scalars are computed via the Minkowski inner product in each block.

The component-level ablations reveal which 4-vector components matter most: the energy component e_t is most critical ($\Delta\text{AUC} = 0.400\,[0.330,\ 0.727]$, with a wide seed band), followed by the full 3-momentum e_{xyz} ($\Delta\text{AUC} = 0.299\,[0.257,\ 0.381]$). The beam-axis e_z alone ($\Delta\text{AUC} = 0.240\,[0.211,\ 0.277]$) is more important than transverse components e_{xy} ($\Delta\text{AUC} = 0.190\,[0.164,\ 0.249]$), consistent with longitudinal momentum being more discriminative for TopTagging. Figure 14 visualizes the same interventions globally and layer-by-layer.

Table 6. Channel ablations for L-GATr-slim on TopTagging. Format: combined-bootstrap median [95% CI]; baseline AUC $= 0.986\,[0.985,\ 0.988]$.

CHANNEL ZEROED	AUC (MEAN)	ΔAUC
h_s **(ALL SCALARS)**	**0.500**	$0.486\,[0.485,\ 0.488]$
h_v (ALL VECTORS)	0.401	$0.585\,[0.559,\ 0.695]$
e_t (ENERGY COMPONENT)	0.586	$0.400\,[0.330,\ 0.727]$
e_z (BEAM AXIS)	0.746	$0.240\,[0.211,\ 0.277]$
e_x, e_y (TRANSVERSE)	0.796	$0.190\,[0.164,\ 0.249]$
e_x, e_y, e_z (3-MOMENTUM)	0.687	$0.299\,[0.257,\ 0.381]$

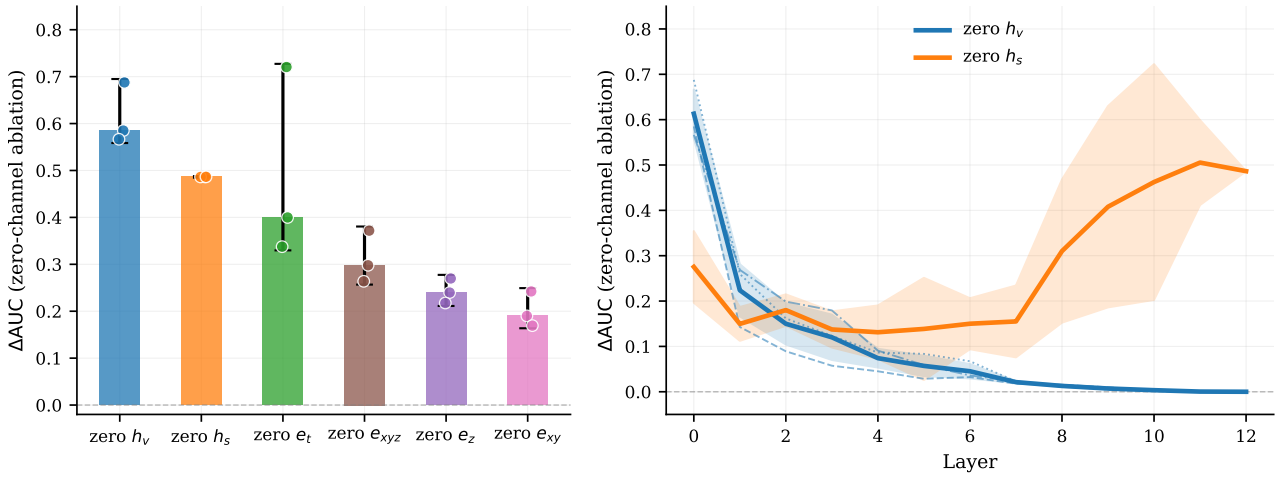

Figure 14. Channel ablations for L-GATr-slim on TopTagging. **Left:** global ΔAUC for all six interventions with combined 95% bootstrap CIs (bars) and per-seed dots. Interventions ordered by decreasing impact: h_v (all vectors), h_s (all scalars), e_t (energy), e_{xyz} (3-momentum), e_z (beam axis), e_{xy} (transverse). **Right:** layer-resolved ΔAUC for h_s and h_v only (component ablations were not run per-layer); thin lines show per-seed h_v trajectories.

E.3. Subgroup=False Side Study

Table 7 and Figure 15 report zero-grade ablations for L-GATr trained with the alternative subgroup setting, which uses five independent grades rather than the three mixed groups used in the main text. This side study confirms that the G2 ≈ 0 finding is not an artifact of the subgroup mixing in the default setting.

G2 bivector $\Delta\text{AUC} = 0.003\,[0.001,\ 0.005]$, essentially zero, confirming the main-text 3-group result. G1 vector is again dominant with high variance ($\Delta\text{AUC} = 0.120\,[0.097,\ 0.451]$), consistent with the multiple-pathway interpretation. G3 trivector and G4 pseudoscalar are both negligible ($\Delta\text{AUC} < 0.001$), indicating parity-odd components carry no load-bearing information for TopTagging.

Table 7. Zero-grade ablations for L-GATr on TopTagging with the alternative subgroup setting (5 independent grades). Format: combined-bootstrap median [95% CI]; per-seed values in brackets. Baseline AUC: $0.986\,[0.984,\ 0.988]$.

GRADE	ZERO ΔAUC
G0 SCALAR	$0.039\,[0.035,\ 0.047]\ [0.044, 0.038, 0.039]$
G1 VECTOR	$0.120\,[0.097,\ 0.451]\ [0.119, 0.442, 0.101]$
G2 BIVECTOR	$0.003\,[0.001,\ 0.005]\ [0.004, 0.003, 0.001]$
G3 TRIVECTOR	$0.000\,[0.000,\ 0.001]$
G4 PSEUDOSCALAR	$0.000\,[0.000,\ 0.000]$

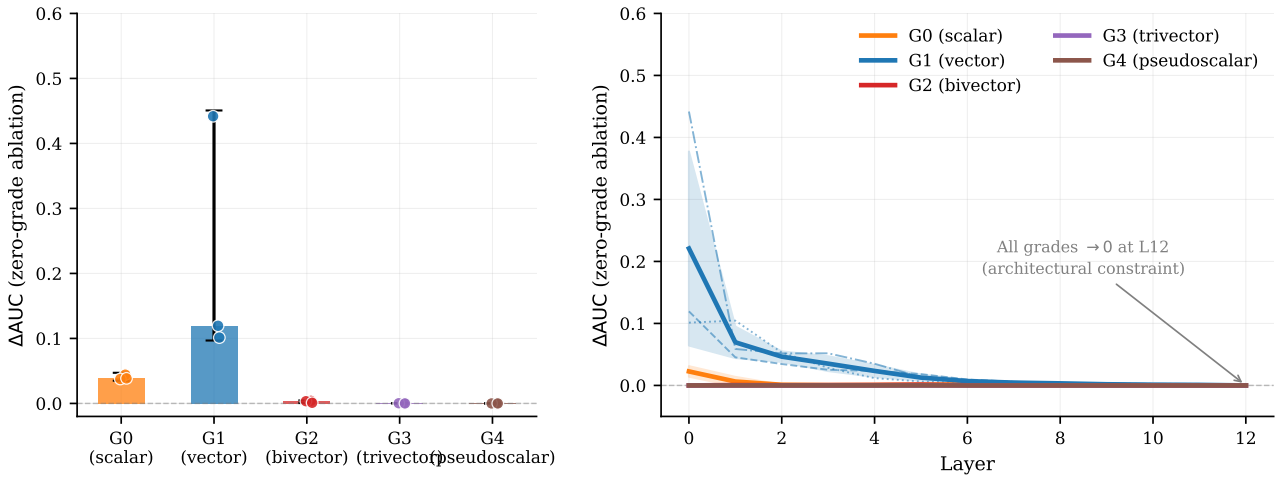

Figure 15. Grade ablations for L-GATr with the alternative subgroup setting (5 independent grades G0–G4). **Left:** global zero-ablation ΔAUC with combined 95% bootstrap CIs and per-seed dots. **Right:** layer-resolved ΔAUC (seed mean $\pm$ std); thin lines show per-seed G1 trajectories. G2, G3, and G4 are all negligible; G1 is again dominant with high cross-seed variance, consistent with the 3-group main result.

F. Equivariance Details

Equivariance uncertainty. For the random-transform equivariance summary, each model is evaluated on 200 test jets with 5 random Lorentz transforms per seed. The tabulated intervals pool 1,000 bootstrap draws from each of the 3 independently trained seeds, for 3,000 pooled draws per model. Table 8 reports the full uncertainty summary for the equivariance experiment in Section 4.

Table 8. Mean relative logit change under random Lorentz transforms (200 jets $\times$ 5 random transforms per seed). Values are combined-bootstrap medians with 95% CIs; seed ranges are shown separately. Lower is more equivariant. †LLoCa-T is equivariant via canonicalization; larger measured errors reflect numerical sensitivity of the canonicalization step.

Model	Bootstrap median [95% CI]	Seed range	vs. L-GATr
L-GATr	$5.2\,[3.0,\ 8.7] \times 10^{-3}$	$[0.31,\ 12] \times 10^{-3}$	$1\times$
L-GATr-slim	$3.9\,[2.6,\ 6.3] \times 10^{-3}$	$[2.9,\ 4.9] \times 10^{-3}$	$0.75\times$
LLoCa-T†	$0.131\,[0.060,\ 0.251]$	$[0.071,\ 0.26]$	$\sim 25\times$
Vanilla-T	$1.20\,[0.98,\ 1.63]$	$[1.03,\ 1.56]$	$\sim 230\times$
ParT	$3.34\,[2.57,\ 4.29]$	$[2.04,\ 4.79]$	$\sim 650\times$

F.1. Per-Layer Equivariance

Table 9 reports the per-layer equivariance errors for L-GATr on TopTagging (200 jets, 5 random Lorentz transforms each).

At layer 0 (the input projection), h_{s} invariance is exact to machine precision (0) and h_{mv} equivariance error is 5×10^{-8} (float32 machine epsilon), so the input projection is exactly equivariant. Across layers 1–12, errors accumulate to 10^{-4}–5×10^{-3} for h_{s} and 7×10^{-3}–10^{-2} for h_{mv}, consistent with floating-point drift across 12 GATrBlock operations. The naïve baseline (treating a boosted input as unchanged) gives errors of ~ 1.3 throughout, confirming that the equivariant architecture reduces the effective error by approximately $100\times$ at all layers.

We cap the main-text sweep at $\gamma \leq 5$ because stronger boosts are poorly supported by the TopTagging rapidity range and are more sensitive to implementation-level numerical effects.

Table 9. Per-layer equivariance errors for L-GATr on TopTagging (200 jets $\times$ 5 random Lorentz transforms). h_{s} error: relative change in scalar channels (should be 0). h_{mv} equivariance: relative change after applying the correct group action. Naïve: error if no transformation applied.

Layer	h_{s} error	h_{mv} equiv.	Naïve
L0 (input projection)	0 (exact)	5×10^{-8}	1.36
L1–L12	10^{-4}–5×10^{-3}	7×10^{-3}–10^{-2}	~ 1.3
Output logit	5.17×10^{-3} (0.52%)		

Table 10 reports the boost sweep for L-GATr.

Table 10. Boost sweep for L-GATr on TopTagging (pure boosts in a random direction, 50 jets per seed). Values are combined-bootstrap medians with 95% CIs pooled across 3 seeds. h_{mv} error: relative change in multivector hidden states. Logit error: relative change in output logit.

γ	h_{mv} ERROR	LOGIT ERROR
1.0	0 [0, 0]	0 [0, 0]
1.25	3.5 [2.7, 4.5] $\times 10^{-3}$	1.9 [1.3, 2.6] $\times 10^{-3}$
1.5	5.8 [4.4, 7.5] $\times 10^{-3}$	3.3 [2.1, 4.8] $\times 10^{-3}$
1.75	7.4 [5.8, 9.4] $\times 10^{-3}$	3.8 [2.5, 5.5] $\times 10^{-3}$
2.0	8.9 [7.0, 12] $\times 10^{-3}$	4.8 [3.2, 6.9] $\times 10^{-3}$
2.5	1.4 [1.1, 1.7] $\times 10^{-2}$	8.9 [5.1, 15] $\times 10^{-3}$
3.0	2.1 [1.7, 2.6] $\times 10^{-2}$	1.2 [0.77, 1.7] $\times 10^{-2}$
4.0	3.1 [2.5, 3.7] $\times 10^{-2}$	1.8 [1.1, 2.9] $\times 10^{-2}$
5.0	4.7 [3.9, 5.6] $\times 10^{-2}$	3.0 [1.9, 4.3] $\times 10^{-2}$