Explicit or Implicit? Encoding Physics at the Precision Frontier

Victor Bresó-Pla
Harvard University

Kevin Greif
University of California, Irvine

Vinicius Mikuni
Nagoya University

Benjamin Nachman
Stanford University
SLAC National Accelerator Laboratory

Tilman Plehn
Universität Heidelberg

Tanvi Wamorkar
Stanford University
SLAC National Accelerator Laboratory
twamorka@stanford.edu

Daniel Whiteson
University of California, Irvine

Abstract

High-performance machine learning tools in particle physics rest on two complementary directions: encoding symmetries explicitly in the architecture, and implicitly learning the structure of the data through large-scale (pre-) training. We compare the performance of the representative L-GATr and OMNILEARN models on three especially challenging tasks: reweighting-based unfolding, likelihood-ratio estimation, and weakly supervised anomaly detection. Across all benchmarks, both methods achieve comparable performance given the statistical precision of the finetuning datasets, suggesting that the significant efficiency gains from encoding known particle physics structures are largely method-independent.

1 Introduction

Particle physics experiments at accelerators such as the Large Hadron Collider (LHC) produce enormous quantities of data, and extracting physics from this data requires distinguishing subtle differences between distributions of collision events. Machine learning has become essential for this task [1–4]. Collision events produce showers of particles whose collective properties are governed by Lorentz symmetry, giving particle physics data a rich geometric structure. A key design question has therefore emerged: should models be *explicitly* constrained to respect this symmetry by design (*explicit* encoding), or should they be free to *implicitly* discover it through training on large datasets (*implicit* encoding)?

Previous studies have compared these approaches for top tagging [5] or multi-label jet classification [6], where the classes are quite different. In that regime, architectural symmetry constraints provide clear data-efficiency gains, but the best performance is reached by pretrained transformers [7, 8]. However, many important tasks in particle physics require distinguishing distributions that are nearly identical. These include:

- **Likelihood-ratio estimation**: quantifying how much more probable the data is under one physical model than another. We study two cases: direct estimation on electron-proton deep inelastic scattering events recorded by the H1 detector at HERA, and iterative reweighting via OmniFold for Z+jets unfolding, where the ratio is applied repeatedly to correct detector-level data back to the underlying physical distribution.

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

- **Anomaly detection**: searching for rare new phenomena hidden within a large background of known processes, where the signal may constitute only a tiny fraction of events.

All three tasks reduce to accurately estimating the ratio of two probability densities from samples, a hard problem precisely when those densities are similar. We compare two leading models on all three tasks to determine whether explicit or implicit physics encoding confers an advantage in this precision regime. A full account of the methods, datasets and results is provided in Ref. [9].

2 Two Approaches to Encoding Physics

The explicit approach produces architecturally constrained models that treat physically equivalent configurations identically by construction, reducing the learning burden at the cost of architectural complexity. The implicit approach uses large-scale pretraining on diverse data, allowing models to develop internal representations of physical structure without hard constraints.

Explicit symmetry: L-GATr. The Lorentz Geometric Algebra Transformer (L-GATr) [10, 7, 11] is a transformer architecture that is by construction equivariant under Lorentz transformations, representing inputs as elements of the spacetime geometric algebra. Additional non-geometric features can be included at the cost of partially breaking the symmetry in a controlled way. We test both, the full L-GATr model (10^6–2×10^6 parameters) and a streamlined variant, L-GATr-slim (7×10^5 parameters), which retains the key symmetry structure at lower computational cost [11].

Implicit symmetry: OmniLearn. OmniLearn [12] is a foundation model pretrained on diverse data and then adapted (fine-tuned) to specific tasks. Its backbone is the Point-Edge Transformer (PET), a hybrid architecture combining standard transformer attention with graph neural network layers that capture local geometric relationships between nearby particles. OmniLearn is pretrained on approximately 10^8 simulated particle jets from a diverse dataset spanning many jet types [6], acquiring representations of physical structure implicitly through data exposure rather than architectural constraints. Fine-tuned OmniLearn models converge 2–3.5$\times$ faster than training from scratch, though the one-time pretraining cost is large ($\sim 10^{19}$ floating-point operations, roughly 25$\times$ the cost of a full L-GATr training). For downstream users who rely on publicly released pretrained weights, this cost is not incurred.

3 Experiments

We evaluate both models on three tasks requiring precision likelihood-ratio estimation, spanning different collision systems and detectors.

3.1 Correcting simulations of proton collisions

Our first task is unfolding: using machine learning to statistically correct a simulation to match real data, accounting for detector effects [13–17]. The specific method we use, OmniFold [13, 16], trains a classifier to iteratively reweight simulated events at detector level so that their distribution matches the data, then propagates this correction back to the underlying particle-level distribution. We run this procedure for 5 iterations.

We use simulated proton-proton collisions producing a Z boson plus hadronic jets at 14 TeV center-of-mass energy [18, 13]. Two different event generators: Herwig [19–21] and Pythia [22], play the roles of "data" and "simulation"; they produce similar but not identical distributions, mimicking the real-world situation where simulation does not perfectly describe data.

We measure unfolding quality using the triangular discriminator on several jet observables: jet mass m, jet width w, jet multiplicity N, groomed mass $\log\rho$, groomed momentum fraction z_g, and N-subjettiness ratio τ_{21}. Lower traingular discriminator values indicate better agreement between the corrected simulation and the target. Results are shown in Table 1 and representative unfolded distributions in Figure 1. Both L-GATr and OmniLearn substantially outperform the PET architecture trained from scratch, confirming that both explicit and implicit physics priors help. L-GATr achieves the best performance on three of the six observables, suggesting that explicit Lorentz symmetry is a useful inductive bias for this task. The slim variant L-GATr-slim matches or exceeds the full

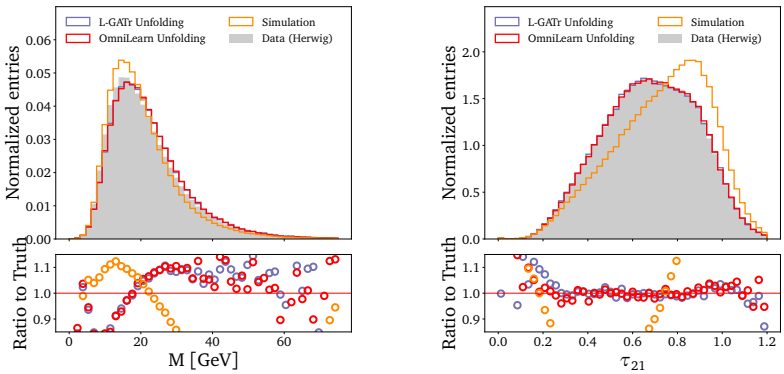

Figure 1: Unfolded distributions for two observables obtained from L-GATr, PET and OMNILEARN after 5 reweighting iterations. The data events are generated with Herwig, and the simulation events are generated with Pythia. The L-GATr architecture consists of 10^6 learnable parameters.

Method (Params.)	m	w	N	$\log \rho$	z_g	τ_{21}
PET (2×10^6)	2.35 ± 0.06	0.22 ± 0.02	0.49 ± 0.03	0.17 ± 0.02	0.35 ± 0.02	0.12 ± 0.02
OMNILEARN (2×10^6)	2.06 ± 0.05	0.18 ± 0.02	0.33 ± 0.03	$\mathbf{0.12 \pm 0.01}$	$\mathbf{0.27 \pm 0.02}$	0.09 ± 0.01
L-GATr-slim (7×10^5)	2.63 ± 0.07	0.14 ± 0.02	0.54 ± 0.04	0.14 ± 0.02	0.28 ± 0.02	$\mathbf{0.07 \pm 0.01}$
L-GATr (10^6)	2.05 ± 0.06	$\mathbf{0.12 \pm 0.02}$	$\mathbf{0.24 \pm 0.02}$	0.23 ± 0.02	0.28 ± 0.02	0.08 ± 0.01
L-GATr (2×10^6)	$\mathbf{2.01 \pm 0.06}$	0.17 ± 0.02	0.47 ± 0.04	0.13 ± 0.02	0.37 ± 0.03	0.08 ± 0.01

Table 1: Particle-level triangular discriminators ($\times 1000$) between the Herwig and reweighted Pythia distributions after 5 iterations of OMNIFOLD using the L-GATr, L-GATr-slim and OMNILEARN networks as classifiers. The best results are in bold.

model on several shape variables despite having roughly $7 \times$ fewer parameters, while the larger 2×10^6-parameter L-GATr offers only marginal gains over the 10^6 version, indicating diminishing returns from scaling at this dataset size.

3.2 Likelihood ratio estimation for electron-proton collisions

Our second task involves a different type of collider experiment: deep inelastic scattering (DIS), in which an electron probes the internal structure of a proton. We use data from the H1 detector at the HERA collider [23–25], reweighting events from one generator (Rapgap 3.1 [26]) to match another (Djangoh 1.4 [27]).

Method (Params.)	AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.5$)
PET (2×10^6)	0.5691	0.547	2.467 ± 0.002
OMNILEARN (2×10^6)	0.5695	0.547	2.470 ± 0.003
L-GATr (10^6)	0.5603	0.541	2.396 ± 0.003
L-GATr (2×10^6)	0.5630	0.543	2.408 ± 0.002

Table 2: Performance comparison of classifiers on the H1 dataset.

The network performance is evaluated using the accuracy, AUC and rejection rate at 50% signal efficiency on the test dataset, and is listed in Table 2. The resulting reweighted distributions are shown in Figure 2. L-GATr consistently underperforms both OMNILEARN and PET on this task, and increasing the L-GATr size to 2×10^6 parameters, our closest match to OMNILEARN, yields only marginal improvements. In particular, the limited improvement when increasing the L-GATr capacity suggests that scaling the number of model parameters does not match the performance gains provided

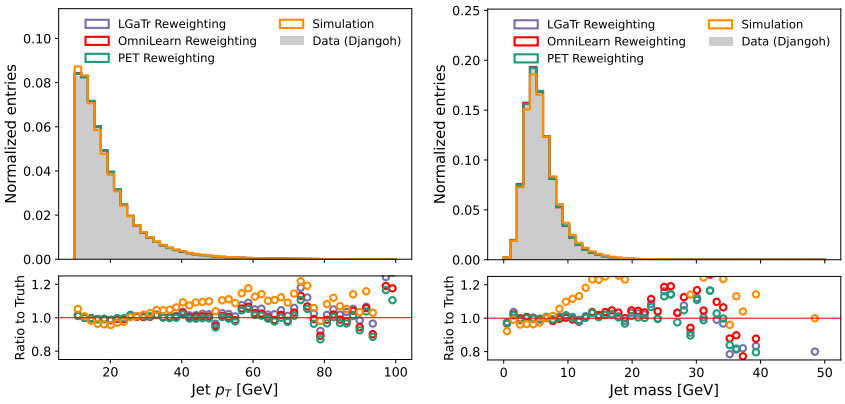

Figure 2: Reweighted distributions for four observables obtained from L-GATr, PET and OMNILEARN after 5 classifier trainings. The data events are generated with Djangoh, and the simulation events are generated with Rapgap. The L-GATr architecture consists of 10^6 learnable parameters.

by the design of the PET architecture and to a lesser degree OMNILEARN's large-scale pretraining on the data side.

3.3 Searching for new physics with anomaly detection

Our third task is weakly supervised anomaly detection [28–30]: training a classifier to separate data, possibly containing a small fraction of signal events, from a pure-background reference sample. We use the LHC Olympics benchmark [31], in which a hypothetical heavy resonance decays to pairs of lighter particles, producing a localized excess in the distribution of the total invariant mass of the event.

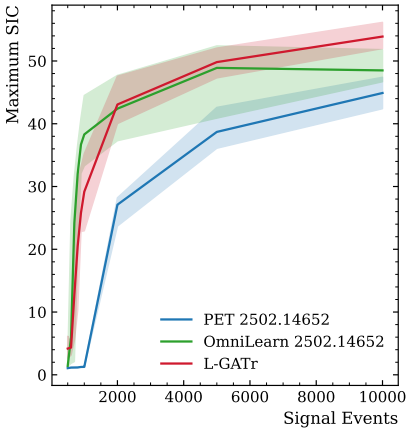

Figure 3: Maximum SIC against the number of injected signal events for the randomly initialized PET and L-GATr models, as well as the pretrained OMNILEARN. Solid lines plot the mean maximum SIC across an ensemble of 10 independent trainings, and shaded bands represent the 68% confidence band over the ensemble.

Performance is measured by the significance improvement characteristic (SIC), which quantifies how much a classifier enhances the signal-to-background ratio relative to a random selection. Figure 3 shows the maximum SIC as a function of the number of injected signal events, where both L-GATr and OmniLearn substantially outperform a randomly initialized baseline and achieve comparable sensitivity across the full range of signal fractions tested.

4 Conclusion

Across all three benchmarks, both methods achieve comparable performance given the statistical precision of the fine-tuning datasets, suggesting that the efficiency gains from encoding known particle physics structures are largely method-independent. It is not clear whether the two approaches learn exactly the same underlying physics, and there is no practical impediment to combining them — a systematic study of whether leveraging both strategies simultaneously yields further gains is left to future work.

References

[1] Alexander Radovic, Mike Williams, David Rousseau, Michael Kagan, Daniele Bonacorsi, Alexander Himmel, Adam Aurisano, Kazuhiro Terao, and Taritree Wongjirad. Machine learning at the energy and intensity frontiers of particle physics. *Nature*, 560(7716):41–48, 2018.

[2] Georgia Karagiorgi, Gregor Kasieczka, Scott Kravitz, Benjamin Nachman, and David Shih. Machine Learning in the Search for New Fundamental Physics. 12 2021.

[3] Matthew Feickert and Benjamin Nachman. A Living Review of Machine Learning for Particle Physics. 2 2021.

[4] Tilman Plehn, Anja Butter, Barry Dillon, Theo Heimel, Claudius Krause, and Ramon Winterhalder. Modern Machine Learning for LHC Physicists. 11 2022.

[5] Anja Butter et al. The Machine Learning landscape of top taggers. *SciPost Phys.*, 7:014, 2019.

[6] Huilin Qu, Congqiao Li, and Sitian Qian. Particle Transformer for Jet Tagging. 2 2022.

[7] Johann Brehmer, Víctor Bresó, Pim de Haan, Tilman Plehn, Huilin Qu, Jonas Spinner, and Jesse Thaler. A Lorentz-equivariant transformer for all of the LHC. *SciPost Phys.*, 19(4):108, 2025.

[8] Wahid Bhimji, Chris Harris, Vinicius Mikuni, and Benjamin Nachman. OmniLearned: A Foundation Model Framework for All Tasks Involving Jet Physics. 2025.

[9] Anonymous authors. Explicit or Implicit? Encoding Physics at the Precision Frontier. 3 2026.

[10] Johann Brehmer, Pim de Haan, Sönke Behrends, and Taco Cohen. Geometric algebra transformer, 2023.

[11] Antoine Petitjean, Tilman Plehn, Jonas Spinner, and Ullrich Köthe. Economical Jet Taggers – Equivariant, Slim, and Quantized. 12 2025.

[12] Vinicius Mikuni and Benjamin Nachman. Solving key challenges in collider physics with foundation models. *Phys. Rev. D*, 111(5):L051504, 2025.

[13] Anders Andreassen, Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. OmniFold: A Method to Simultaneously Unfold All Observables. *Phys. Rev. Lett.*, 124(18):182001, 2020.

[14] Marco Bellagente, Anja Butter, Gregor Kasieczka, Tilman Plehn, and Ramon Winterhalder. How to GAN away Detector Effects. *SciPost Phys.*, 8(4):070, 2020.

[15] Marco Bellagente, Anja Butter, Gregor Kasieczka, Tilman Plehn, Armand Rousselot, Ramon Winterhalder, Lynton Ardizzone, and Ullrich Köthe. Invertible Networks or Partons to Detector and Back Again. *SciPost Phys.*, 9:074, 2020.

[16] Anders Andreassen, Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, Adi Suresh, and Jesse Thaler. Scaffolding Simulations with Deep Learning for High-dimensional Deconvolution. 5 2021.

[17] Nathan Huetsch et al. The landscape of unfolding with machine learning. *SciPost Phys.*, 18(2):070, 2025.

[18] Anders Andreassen, Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. Pythia/herwig + delphes jet datasets for omnifold unfolding, 2019.

[19] M. Bahr et al. Herwig++ Physics and Manual. *Eur. Phys. J. C*, 58:639–707, 2008.

[20] Johannes Bellm et al. Herwig 7.0/Herwig++ 3.0 release note. *Eur. Phys. J. C*, 76(4):196, 2016.

[21] Johannes Bellm et al. Herwig 7.1 Release Note. 5 2017.

[22] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015.

[23] Miguel Arratia, Daniel Britzger, Owen Long, and Benjamin Nachman. Reconstructing the Kinematics of Deep Inelastic Scattering with Deep Learning. 10 2021.

[24] Miguel Arratia, Daniel Britzger, Owen Long, and Benjamin Nachman. Optimizing observables with machine learning for better unfolding. *JINST*, 17(07):P07009, 2022.

[25] Owen Long and Benjamin Nachman. Designing observables for measurements with deep learning. *Eur. Phys. J. C*, 84(8):776, 2024.

[26] Hannes Jung. Hard diffractive scattering in high-energy e p collisions and the Monte Carlo generator RAPGAP. *Comput. Phys. Commun.*, 86:147–161, 1995.

[27] K. Charchula, G. A. Schuler, and H. Spiesberger. Combined QED and QCD radiative effects in deep inelastic lepton - proton scattering: The Monte Carlo generator DJANGO6. *Comput. Phys. Commun.*, 81:381–402, 1994.

[28] Theo Heimel, Gregor Kasieczka, Tilman Plehn, and Jennifer M. Thompson. QCD or What? *SciPost Phys.*, 6(3):030, 2019.

[29] Benjamin Nachman and David Shih. Anomaly Detection with Density Estimation. *Phys. Rev. D*, 101:075042, 2020.

[30] Gregor Kasieczka et al. The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics. 1 2021.

[31] Gregor Kasieczka, Benjamin Nachman, and David Shih. Official Datasets for LHC Olympics 2020 Anomaly Detection Challenge (Version v6) [Data set]., 2019. https://doi.org/10.5281/zenodo.4536624.

PAI26 Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract clearly states the main contribution—comparing L-GATr and OmniLearn on three precision tasks and finding comparable performance—which matches the experimental results in Tables 1–2 and Figure 3.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper briefly mentions a limitation in Section 4, which is a systematic study of combining the explicit and implicit approach. This will be addresses in future work.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper is purely empirical; it contains no theoretical results or proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The models, datasets, evaluation metrics, and number of iterations are described in the text. Full experimental details are provided in the paper Ref. [9]. Public datasets (LHC Olympics, H1/HERA, Pythia/Herwig jet datasets) are cited with DOIs or references.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: No code repository is provided with this submission. Code details will be provided if the paper is selected to the workshop.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [No]

   Justification: Training details such as optimizer, learning rate, and data splits are not reported in this short paper; they are deferred to Ref. [9].

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 1 and 2 report mean $\pm$ standard deviation. Figure 3 shows 68% confidence bands computed over ensembles of 10 independent trainings, which is clearly stated in the caption.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: A full computational analysis is presented in Ref. [9].

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses standard physics simulations and publicly available datasets. No human subjects, sensitive data, or ethically problematic applications are involved.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is foundational methodology research in particle physics. The techniques developed are specific to collider data analysis and have no direct path to societal harms.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The models are physics-specific classifiers. No pretrained general-purpose models, image generators, or scraped internet datasets are released, so no safeguards are required.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models (Herwig, Pythia, H1/HERA data, LHC Olympics dataset with Zenodo DOI) are properly cited. The models being studied: L-GATr and OmniLearn are also appropriately cited and each have their own license.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: No new datasets or models are explicitly released with this submission.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are not a component of the core methodology in this research.