SEAL - A Symmetry Encouraging Loss for High Energy Physics

Pradyun Hebbar [1 2]
Thandikire Madula [3]
Vinicius Mikuni [4]
Benjamin Nachman [5 6]
Nadav Outmezguine [7 1 8]
Inbar Savoray [7 1]

Abstract

Physical symmetries provide a powerful inductive bias for machine learning models in science, improving robustness, data efficiency, and interpretability. However, building models that explicitly enforce symmetries requires specialized architectures, and real-world experiments often break these symmetries through finite detector granularity and energy thresholds. We introduce SEAL (Symmetry Encouraging Loss), a family of soft-constraint loss terms that encourage equivariance, requiring no architectural modifications. We present two complementary variants: GSEAL, which penalizes output differences under random group transformations of the input, and δSEAL, which penalizes the model's gradients along directions corresponding to infinitesimal symmetry transformations. We focus on Lorentz invariance, which is highly relevant to High Energy Physics and rarely encountered in other domains, but the formulation is general and applicable to any Lie group. Using top quark tagging as a case study, we observe that the addition of the soft constraints leads to more robust performance while requiring minimal computational costs.

[1]Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA [2]Department of Particle Physics, University of Geneva, Geneva 1205, Switzerland [3]University College London, Gower Street, London, WC1E 6BT, UK [4]Nagoya University, Kobayashi-Maskawa Institute, Aichi 464-8602, Japan [5]Department of Particle Physics and Astrophysics, Stanford University, Stanford, CA 94305, USA [6]Fundamental Physics Directorate, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA [7]Leinweber Institute for Theoretical Physics, University of California, Berkeley, CA 94720, USA [8]Microsoft. Correspondence to: Inbar Savoray <inbar.savoray@berkeley.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Machine learning methods are now essential tools in the search for new, fundamental interactions at colliders and elsewhere. While particle physics is benefiting immensely from advances in the broader machine learning community, we also have unique challenges that require dedicated solutions. One set of challenges is related to the structure of particle physics data. Unlike natural images and language, particle physics data are naturally represented as variable-length point clouds that transform under the Poincaré group. Developing machine learning tools that accommodate this structure is thus an essential research topic in particle physics.

In this paper, we will focus on the Lorentz group, since it is highly relevant for particle physics and not relevant in most other applications. However, the methods we develop here are not specific to the Lorentz group and could have widespread utility.

One well-studied approach to accommodate the Lorentz covariance (called *equivariance* in the machine learning literature) is to design machine learning models that explicitly respect the symmetry (Gong et al., 2022; Qiu et al., 2023; Bogatskiy et al., 2024; Batatia et al., 2023; Spinner et al., 2024; 2025). Such models have been shown to be highly data efficient and perform well on a variety of tasks. However, equivariant networks also come with notable challenges. Typically, embedding symmetries into the network architecture requires using a limited set of operations that lead to highly constrained layers and increased complexity. Consequently, these models often have higher computational costs. Additionally, specialized networks may not observe the same scaling properties as general-purpose architectures, limiting their usability to the regime of smaller datasets.

Furthermore, equivariant networks assume that symmetries manifest perfectly in the data, which is often not true with experimental observations. For example, detectors have finite energy thresholds. Additionally, the direction of the particle beams contribute to the broken symmetry. Previous works (Spinner et al., 2024; Bogatskiy et al., 2024)

try to account for broken symmetries in their equivariant networks by including handcrafted symmetry breaking effects as inputs into the network. This approach, requires the exact knowledge of the symmetry-breaking mechanism to be included. An alternative is to construct the model from both equivariant and non-equivariant components, as done for the Lorentz group in Refs. (Murnane et al., 2023; Nabat et al., 2025).

Equivariant models are then considered "hard" constraints as the networks can only output functions that are strictly equivariant with respect to a group transformation. In this work we explore encouraging equivariance via "soft" constraints which do not require changing the architecture of the model. To apply the constraint, we propose a symmetry encouraging loss (SEAL) – a term added to the loss function which is minimized when the symmetry is respected. We introduce two variations for SEAL; a group-level SEAL (GSEAL), which penalizes the model based on differences between the model's outputs for inputs before and after the group transformation, and an infinitesimal SEAL (δSEAL), which penalizes the gradients of the model along directions corresponding to symmetry transformations of the inputs. Implementing SEAL only requires prior knowledge of the symmetry group and how the inputs transform under the symmetry. It is otherwise completely generic, and can be applied to any architecture with no additional inference-time computational costs, and minimal additional train-time costs. A few previous works have tested penalty terms similar to SEAL, including recently Ref. (Elhag et al., 2024) in the context of E_3 invariance in several problems, and Ref. (Akhound-Sadegh et al., 2023) for enforcing the equivariance of Physics-Informed Networks.

The paper is organized as follows. Sec. 2 introduces the detailed formulation of SEAL. The experiments we have conducted are detailed in Sec. 3, with toy experiments presented in Sec. 3.1, and jet-tagging experiments discussed in Sec. 3.2. We conclude in Sec. 4.

2. Encouraging Symmetries

2.1. Equivariance

A function $f : X \to Y$ is said to be equivariant with respect to a symmetry group G if:

$$f\left(g \odot x\right) = \left(g \odot f\right)(x) \tag{1}$$

for all $x \in X$ and $g \in G$, where $g\odot$ is a group action. We note that $g \odot f$, which is determined from the group action on Y, can be different from $g \odot x$. The function is said to be invariant with respect to this group if:

$$f(g \odot x) = f(x) \tag{2}$$

for all $x \in X$ and $g \in G$. Invariance is thus a special case of equivariance, where $g \odot f$ is the identity map.

2.2. The Lorentz Group

The Lorentz group is the group of all linear transformations of four-dimensional space-time that preserve the Minkowski inner product. Where for a four-vector $u = (t, x, y, z)$ the Minkowski inner product is given by $\eta(u, u) = t^2 - x^2 - y^2 - z^2$. The Lorentz group is a Lie group denoted $O(1, 3)$, however, if we restrict the transformations to those that preserve both the orientation of space and direction of time, then we obtain the restricted Lorentz group denoted by $SO(1, 3)^+$. The transformations included in this group are 3D spatial rotations and Lorentz boosts.

2.3. Soft Penalty Formulation

We introduce soft symmetry constraints by modifying the training procedure and including a penalty term to the loss function. We have data pairs (x, y) where x is the input and y is the target. We train a neural network with parameters θ to minimize the difference between the network output and data target, captured by the loss function $\mathcal{L}$:

$$\min_\theta \mathbb{E}[\mathcal{L}(f_\theta(x), y)]\,. \tag{3}$$

Loss functions such as cross entropy are often used for classification tasks, whereas regression is often done with the mean squared error (MSE). To encourage symmetries, we modify Eq. 3 such that the new loss function is:

$$\min_\theta \mathbb{E}[\mathcal{L}(f_\theta(x), y) + \lambda\Gamma(f_\theta(x), y)]\,. \tag{4}$$

The function Γ is the SEAL, introduced to penalize the model when the symmetry is violated. The tunable hyperparameter λ determines the relative weight of the two loss components.

A general form of Γ is given by Equation 5. If the symmetry is strictly conserved, Γ vanishes for every term in both the integrand and sum and therefore no penalty is applied.

$$\Gamma = \sum_{x\in X} \int_{g\in G} dg\, \left|f(g \odot x) - \left(g \odot f\right)(x)\right|^2\,. \tag{5}$$

To utilize this formulation of the penalty term, two practical adjustments to the general function Γ are required. First, we address the computational challenge of minimizing Γ over the entire space X. Rather than evaluating every point, we employ a Monte Carlo summation over a specific dataset. While we use the training data for X in this work, the choice of X is flexible and can be tailored to any input

region where a symmetry is either expected or desired. Second, as the Lorentz group is continuous and non-compact, integration over the group is intractable[1]. Consequently, for the Lorentz group, we consider two approximate forms for Γ: one that considers Lorentz invariance on a global scale and another that considers it on a local scale.

The global penalty $\Gamma_{\rm G}$ (GSEAL) can be implemented as a stochastic penalty where for each mini-batch of x we apply a random Lorentz transformation and penalize the network deviation:

$$\Gamma_G = \frac{1}{N}\sum_{i=1}^{N} |f(g_i \odot x_i) - (g_i \odot f)(x_i)|^2 \qquad (6)$$

To implement the transformations g_i, during training data point x_i in the mini-batch is randomly boosted by a 3D Lorentz boost:

$$\Lambda = \begin{bmatrix} \gamma & -\gamma\vec{\beta}^T \\ -\gamma\vec{\beta} & I + (\gamma-1)\frac{\vec{\beta}\vec{\beta}^T}{\beta^2} \end{bmatrix}, \qquad (7)$$

where $\vec{\beta}^T = [v_x, v_y, v_z]/c$, and $\gamma = 1/\sqrt{1-\|\vec{\beta}\|^2}$. In our studies here, the boost direction $\hat{\vec{\beta}}$ is uniformly sampled from the unit sphere, while $\|\vec{\beta}\|^2$ is uniformly distributed between 0 and $\beta^2_{\rm max} = 0.95$ to avoid instabilities. We sample a different transformation g for every training data point x and for every training epoch.

The second form of Γ, Γ_δ (δSEAL) is based on a local, differential, Lorentz transformation. For a continuous group, an infinitesimal transformation is generated by the group generators L^a which span its algebra. For example, under an infinitesimal Lorentz transformation, a Lorenz 4-vector is transformed as:

$$x \to x + \sum_a \epsilon^a \cdot L^a_x \cdot x\,, \qquad (8)$$

where $a = 1, ..., 6$ indexes the six generators of the Lorentz group (three rotations and three boosts), ϵ is the infinitesimal transformation parameter, and L^a_x are 4×4 generators in the (1/2,1/2) representation of the Lorentz algebra. Inputs in different representations of the Lorentz group (such as scalars, vectors, tensors etc.) will be transformed by the corresponding representation of L^a. The infinitesimal penalty is then obtained by Taylor expanding the difference:

$$|f(g\odot x) - (g\odot f)(x)|^2 \simeq$$
$$\sum_{a,b} \epsilon^a\epsilon^b (\nabla f \cdot L^a_x \cdot x - L^a_f \cdot f)(\nabla f \cdot L^b_x \cdot x - L^b_f \cdot f(x))\,, \qquad (9)$$

[1]Since the Lorentz group is non-compact, no finite and invariant measure exists and so the integration is ill-defined.

where L^a_f are the six generators in the representation of the Lorentz algebra corresponding to the desired representation of f. To eliminate the ϵ dependence, we divide by $\|\epsilon\|^2$ and maximize over ϵ, resulting in:

$$\Gamma_\delta = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{a=1}^{n} |\nabla f \cdot L^a_x \cdot x_i - L^a_f \cdot f(x_i)|^2\,, \qquad (10)$$

where n is the number of generators, which for the Lorentz group is $n = 6$.

In the next sections, we describe experiments in which the SEALs Γ_G and Γ_δ are added to the task loss. In our examples, the input x were always in the 4-vector representation of the Lorentz group, and the symmetry penalties were aimed at encouraging a Lorentz-invariant function, for which $g \odot f = f$ for Γ_G in eq. (6) and $L^a_f \cdot f = 0$ for Γ_δ in Eq. (10). For both losses, if f is multi-dimensional and/or chosen to be in a multi-dimensional representation one would apply the Euclidean norm over its components when calculating the loss.

3. Experiments

3.1. Toy Experiments

The first set of experiments is aimed at studying the performance of the symmetry penalties in a simple regression task. For this purpose, we have used a sample of 10^5 randomly generated four-vectors $\{p_i\}$, where each of the four components were uniformly distributed between -0.5 and 0.5 (on-shellness was not enforced). The function to be regressed f was a second degree polynomial in Lorentz-invariant scalar products. In one case, $f(p_i)$ depended only on $m_i^2 = p_i^\alpha \eta_{\alpha\beta} p_i^\beta$, and we denote this case as the exactly symmetric case. In the second case, a constant small "spurion" four-vector $s = \left(0, 0, 0, 10^{-3}\right)$ was introduced such that $f(p_i) = f\left(m_i^2, p_i^\alpha \eta_{\alpha\beta} s^\beta\right)$. This case represents a breaking of Lorentz invariance, as f is no longer invariant under arbitrary Lorentz transformations of the original four-vector p_i.

The regression model for this toy example was a multilayer perceptron (MLP) with three hidden layers, each of width 300 with a Gaussian Error Linear Unit (GELU) activation (Hendrycks & Gimpel, 2016). We used the standard MSE loss for the regression, and tested different values of the coefficient λ, determining the relative impact of the added SEAL.

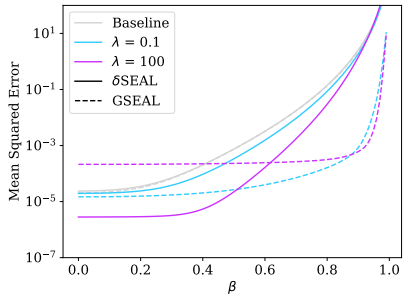
In Fig. 1, we present the MSE between the model's prediction and the true value of $f(p)$. To test the model's performance, we show the MSE obtained on test data which was sampled from the same distribution as the training data, but then boosted in the z-direction by a boost factor β. As can be seen in the plots, when the symmetry is exact, applying a

symmetry penalty can be beneficial both for improving the accuracy on in-distribution test data, and for extrapolating to boosted data. Furthermore, even when a small symmetry breaking is introduced, a SEAL with a modest coefficient can still improve the performance on the original test data compared to using the MSE loss alone, where higher values of λ may be helpful for extrapolating further out.

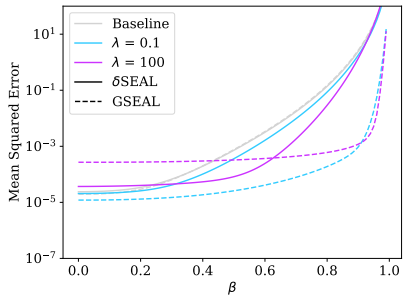
The plot shows the performance both for applying the group-sample SEAL (GSEAL) and the algebra, or infinitesimal SEAL (δSEAL). As expected, GSEAL results in an error that is flatter as a function of the test's boost. For this toy example, GSEAL allows for a modest improvement in performance in-distribution, which becomes more dramatic as the boost increases. When the symmetry is exact, δSEAL with $\lambda = 100$ yields the best accuracy for test boosts up to $\beta \approx 0.4$. This implies that the infinitesimal symmetry penalty, despite only accessing the local properties of the transformation, can be useful for generalization to a large range of boosts. However, when the symmetry is approximate, GSEAL with $\lambda = 0.1$ provides the best performance up to boosts of $\beta \approx 0.8$. GSEAL seems to be less sensitive to small symmetry-breaking effects, although imposing the symmetry more strictly. In the plot presented here, the maximal boost sampled for calculating GSEAL during training was $\beta^2_{\text{max}} = 0.95$. We discuss the effects of choosing different β_{max} values for GSEAL, as well as results for intermediate values of λ in Appendix A.

3.2. Jet Tagging

We illustrate the effect of training with soft penalty constraints in a realistic setting with the ATLAS Top Tagging dataset (ATLAS Collaboration, 2022b). In this task we want to classify jets initiated by the decays of top quarks from the ones produced through Quantum Chromodynamics (QCD). Many deep learning approaches to jet tagging have been investigated over the years such as multilayer perceptrons (MLPs) (Almeida et al., 2015), convolutional neural networks (CNNs) (de Oliveira et al., 2016; Macaluso & Shih, 2018; Bhattacharya et al., 2022; Chen et al., 2020; Lin et al., 2018; ATLAS Collaboration, 2017; Komiske et al., 2017; Chien & Elayavalli, 2018; Barnard et al., 2017; Kasieczka et al., 2017; Choi et al., 2019; Li et al., 2021) and recurrent neural networks (RNNs) (Guest et al., 2016; Fraser & Schwartz, 2018; Egan et al., 2017; Bols et al., 2020). Architectures such as DeepSets (Komiske et al., 2019), graph neural networks (GNNs) (Scarselli et al., 2009; Battaglia et al., 2018) and transformers (Vaswani et al., 2023) which respect the permutation invariance of the particles in the jet have been shown to improve the performance of ML models on jet tasks (Villadamigo et al., 2025; Aad et al., 2025; Mikuni & Nachman, 2025a; Mikuni & Canelli, 2021; Shlomi et al., 2021; ATLAS Collaboration, 2022a; Qu et al., 2024; Qu & Gouskos, 2020; Mikuni



(a)



(b)

Figure 1. The MSE score as a function of the boost applied to the training data. In Fig. 1a the symmetry is exact, while in Fig. 1b the symmetry is broken by a small spurion $s = \left(0, 0, 0, 10^{-3}\right)$. The differential symmetry penalty δSEAL is shown in solid lines, and the group-sample penalty loss GSEAL is shown in dashed lines.

& Canelli, 2020). Lorentz invariance is a natural requirement for the classifier output, as the jet's label should not depend on the spatial orientation or the boost of the jet. Indeed, further improvements have been demonstrated by using Lorentz equivariant top-taggers (Gong et al., 2022; Qiu et al., 2023; Bogatskiy et al., 2024; Batatia et al., 2023), however, the colliding beams, detector effects, imperfect reconstruction and clustering schemes introduce an effective possible breaking of the symmetry.

In this dataset, events are generated with PYTHIA 8 using the NNPDF2.3LO (Ball et al., 2013) set of parton distribution functions and the A14 (Buckley, 2014) set of tuned parameters. Additional pileup effects are simulated by overlaying inelastic interactions on top of the hard scattering process using the 2017 data taking period. Hadronic boosted top quarks are produced in simulated events containing the decay of a heavy Z' boson with mass fixed at 2 TeV. Background jets are obtained from simulations of generic dijet events. Unified Flow Objects (Aad et al., 2021) are used to combine the information of multiple detectors to provide particle reconstruction. Jets are clustered using anti-k_t algorithm (Cacciari & Salam, 2006; Cacciari et al., 2012; 2008) using R=1.0 while additional pileup mitigation algorithms (Berta et al., 2014; 2019; Cacciari et al., 2015; Larkoski et al., 2014) are applied. We use 16 million jets for training and 4 million jets used for validation.

We investigate two training procedures: a baseline classification in which we minimize the binary cross entropy loss in line with equation 3 and a soft constraint procedure where we train the classifier with Γ_G or with Γ_δ. A transformer model was used for both training procedures. It is composed of first an embedding layer of 256 units, then 3 transformer encoder layers (Paszke et al., 2019), each with a model dimension of 256 and 4 attention heads. The pooling operation following the encoder layers is the mean. The final section of the model is a feed-forward neural network with 3 hidden layers each containing 128 units. A ReLu (Agarap, 2019) activation function is used between the hidden layers. The final layer has a single unit followed by a sigmoid activation function. Overall the model has about 1.3 million trainable parameters. The inputs to the model are functions of the four-momenta of the jet constituents. The variables used for each constituent i were its absolute energy E^i and transverse momentum p_T^i through $\log\left(E^i/1\text{GeV}\right)$, $\log\left(p_T^i/1\text{GeV}\right)$, as well as five additional variables relative to the the jet's J kinematics – $\log\left(p_T^i/p_T^J\right)$, $\log\left(E^i/E^J\right)$, $\Delta\phi = \phi^i - \phi_J$, $\Delta\eta = \eta^i - \eta_J$ and $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ with η the pseudorapidity and ϕ the azimuthal angle.

Additionally, we present the performance of the PELICAN (Bogatskiy et al., 2024) model, a tagger with hard symmetry constraints in its architecture. For the case of top-tagging, PELICAN enforces Lorentz-invariance through creating an intermediate matrix of all possible Lorentz scalars (all pair-wise scalar products of constituent four-momenta) at the very first layer, and processing only those scalars moving forward. Since we are interested in comparing softly-constrained models to fully symmetric ones, we do not show the performance of the PELICAN-spurion variant, which supplements the inputs with symmetry-breaking constants representing the colliding beams.

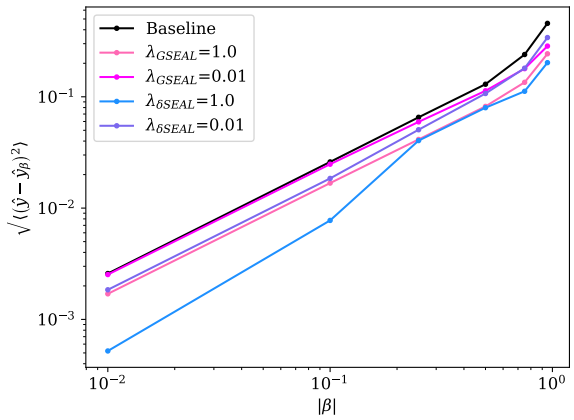


Figure 2. Tagger invariance to 3D Lorentz boosts as a function of the boost parameter evaluated on the ATLAS Top Tagging dataset.

We evaluate the results by first quantifying the invariance of the predicted outputs under Lorentz transformations. We apply a 3D Lorentz boost to the test data with different values of $|\vec{\beta}|$, and calculate the difference between the model's prediction for the original and boosted jet. The results are shown in Fig. 2. As expected, both GSEAL and δSEAL attain better invariance than the baseline model across boosts. For smaller values of the constraint strength λ, GSEAL is noticeably more effective at large boosts, while δSEAL is more effective at small boosts. While at $\lambda = 1.0$ δSEAL becomes more invariant than GSEAL even at large boosts, this may come at a greater cost to the model's accuracy on the test dataset. We therefore proceeded with $\lambda = 1.0$ for GSEAL, and $\lambda = 0.01$ for δSEAL.

In Fig. 3, we show the balanced accuracy of the taggers as a function of the jet p_T. The balanced accuracy is defined as $0.5\left(\text{TP}/\left(\text{TP}+\text{FN}\right)+\text{TN}/\left(\text{TN}+\text{FP}\right)\right)$, with TP (TN) the number of correctly identified signal (background) jets, and FN (FP) the number of wrongly identified signal (background) jets. We chose the balanced accuracy over the accuracy since the number of signal and background jets is not necessarily equal in each p_T bin. The uncertainty bars for the transformer classifiers are given by calculating the standard deviation of five trainings with different seeds. We find that the performance across all

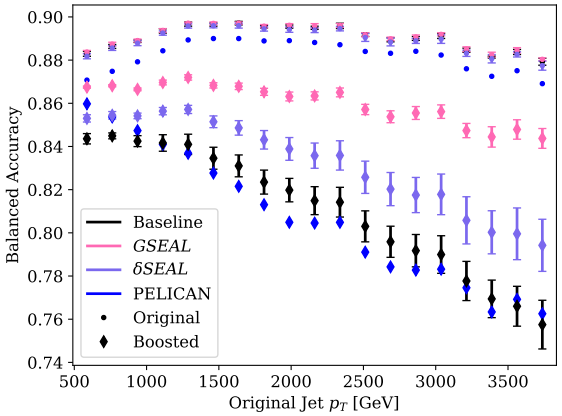


Figure 3. Balanced accuracy as a function of the original jet transverse momentum. Circular markers represent models evaluated on the original test dataset. Diamonds show the balanced accuracy evaluated on the boosted test data set. Also shown is the performance of PELICAN.

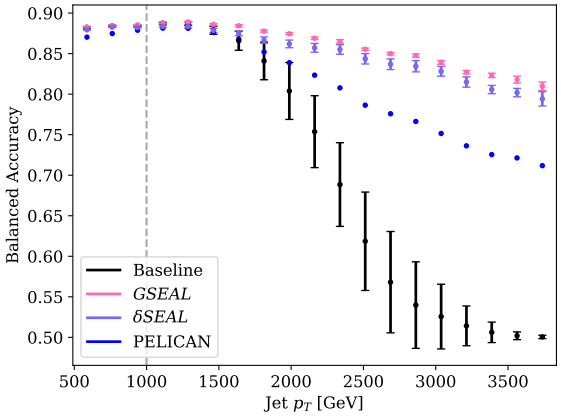


Figure 4. Tagger accuracy as a function of jet p_T for the baseline model and soft penalty model. The vertical line represents the p_T cut applied during training, values to the left were seen during training, values to the right were unseen. The SEALs weights are chosen to be $\lambda = 1.0$ both for models trained with Γ_G and for models trained with Γ_δ.

models is stable as a function of the jet p_T, and is similar between the baseline and softly-constrained models. The overall performance metrics are summarized in Table 1. For a comparison of the training time and evaluation time see Appendix B.

While our models found equally good fits to the original data, they differ in their predictions on boosted jets. This is shown via the diamond shaped markers in Fig. 3, which depict the balanced accuracy on randomly boosted jets, drawn from the same boost distribution used for GSEAL. Since the truth label of these boosted jets is unknown, the balanced accuracy for a boosted jet is calculated with respect to the truth label of the original jet. After boosting the original jets, we observe a reduction in performance for all models. Similarly to Figure 2, both SEAL variations improve the model's robustness to boosts. GSEAL achieves the highest similarity between original and boosted inputs, aligned with its training objective.

Next, we investigate the ability of the models to extrapolate to unseen regions of the phase space used during the training. We train the taggers on jets with $p_T \leq 1$ TeV, and then we evaluate their performance on the original test jets, which extend to a higher p_T range.

In Fig. 4, we show the balanced accuracy of the baseline tagger and the taggers trained with SEAL as a function of the jet p_T, where we set $\lambda = 1.0$ for both GSEAL and δSEAL. We observe that the models perform similarly up to approximately 1.5 TeV, close to the training cut. However, beyond this point the baseline model's accuracy deteriorates rapidly, while both δSEAL and GSEAL show improved levels of robustness. In addition, the baseline model's variance across trainings grows significantly with respect to those of the softly-constrained models, with GSEAL exhibiting the highest accuracy and smallest variance. [2] In Fig. 5 we show the inverse of the background acceptance rate of the taggers at a signal efficiency of 0.3. It is clear that the taggers trained with the soft symmetry constraints are able to have a higher background rejection, by a factor of 10-20 more compared to the baseline for the same signal efficiency.

4. Conclusion

Symmetries are fundamental to particle physics, and can be used to guide ML models towards specific behaviors. While enforcing symmetries through architectural choices can be useful, equivariant networks are challenged by expressivity and scalability, and require modifications to account for symmetry-breaking effects. We presented SEAL, a symmetry-enhancing loss term, to incentivize Lorentz invariance in ML models without modifying their architecture. We introduced two variations, GSEAL – penalizing differences in the model's output in response to random boosts of the input, and δSEAL – penalizing the model's gradients along symmetry transformations directions.

In a toy regression task, we found that SEAL can improve performance both when the symmetry is exact and when it is approximate. In a top-tagging task with the realistic ATLAS dataset (ATLAS Collaboration, 2022b), SEAL was

[2] Interestingly, the baseline model's variance shrink around the 0.5 accuracy mark as the model predicts that all jets in this region are QCD jets independent of the truth jet origin.

	Balanced Accuracy	AUC	$1/\epsilon_b^{0.3}$
Baseline	$0.891 \pm 1.3 \cdot 10^{-3}$	$0.959 \pm 1.0 \cdot 10^{-3}$	652 ± 37
Baseline + GSEAL	$0.891 \pm 1.1 \cdot 10^{-3}$	$0.959 \pm 7.6 \cdot 10^{-4}$	638 ± 37
Baseline + δSEAL	$0.890 \pm 1.7 \cdot 10^{-3}$	$0.959 \pm 1.2 \cdot 10^{-3}$	620 ± 48
PELICAN	0.890	0.959	630

Table 1. Performance metrics for our taggers: accuracy, area under the curve, and inverse of background acceptance rates at signal efficiency of 0.3. The errors are given by the standard deviation across 5 trainings.

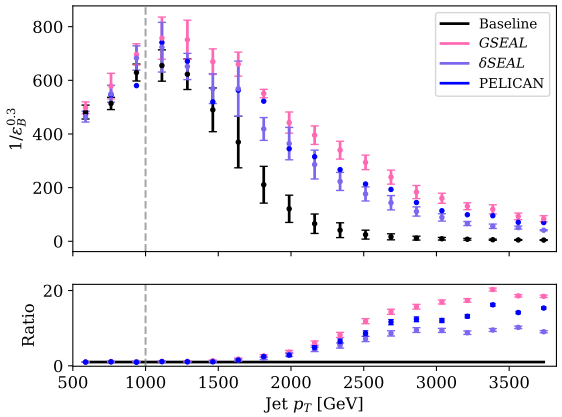


Figure 5. Tagger background rejection at a signal acceptance of 0.3 a function of jet p_T for the baseline model and soft penalty model. The vertical line represents the p_T cut applied during training, values to the left were seen during training, values to the right were unseen.

able to improve the model's invariance to Lorentz transformations without sacrificing performance. We have also observed that SEAL improved extrapolation from low-p_T jets to high-p_T jets, implying it enhances generalization to unseen kinematical regions.

SEAL can be applied in a wide range of contexts, including different tasks and physical objectives, as well as outputs that have other Lorentz transformation properties, such as particle energies or four-momenta. Additionally, since SEAL does not make use of truth labels, it may be useful beyond supervised learning. One may also test SEAL with further datasets exhibiting different levels of Lorentz invariance, and quantify its utility for different data sizes, model sizes and architectures.

There are a few further directions to explore for optimizing SEAL's implementation. For example, rotations can be included in GSEAL by sampling a random rotation matrix in addition to the boost matrix (for a discussion of a general form for Lorentz transformations see (Haber, 2024)). A more in-depth study of the SEAL hyperparameters is needed to find the ideal constraint strength λ and maximal boost $\beta_{\max}$, which could be learnable or change dynamically during training. Different distributions for sampling transformations for GSEAL may also be considered.

Here we compared softly-constrained symmetries to architectures enforcing those constraints perfectly. However, one can consider other methods accounting for approximate symmetries, such as including symmetry-breaking inputs. Another technique is data augmentation, where group orbits are assigned the same label as the data during training (Quiroga et al., 2019; Gerken et al., 2022; Iglesias et al., 2023; Chen et al., 2025).[3] Those can be studied in comparison to SEAL, as well as be combined with it.

Finally, since SEAL does not require any adjustments to current network models, it would be interesting to investigate the impact of adding SEAL to common jet taggers in HEP using different architectures. One example is to add SEAL to fine-tuning tasks from a pre-trained model without constraints. This can be accomplished foundational models (Mikuni & Nachman, 2025a;b; Bhimji et al., 2025; Feickert & Nachman, 2021; Birk et al., 2024; Harris et al., 2024; Golling et al., 2024; Leigh et al., 2024; Bardhan et al., 2025), where the pre-trained model can be loaded with SEAL added as part of the loss function.

Code Availability

For the code for this paper see https://github.com/inbarsavoray/SEAL.git.

Acknowledgments

VM is supported by JST EXPERT-J, Japan Grant Number JPMJEX2509. BN is supported by the U.S. Department of Energy (DOE), Office of Science under contract DE-AC02-76SF00515. IS and NO are supported by the U.S. Department of Energy (DOE), Office of Science under contract DE-AC02-05CH11231. IS also acknowledges support by the Weizmann Institute of Science Women's Postdoctoral Career Development Award. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported

[3] While our approach shares similarities with data augmentation, augmentation randomly transforms the inputs, we focus on directly penalizing the model through a loss term based on the symmetry group.

by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC awards HEP-ERCAP0021099 and HEP-ERCAP0028249. TM gratefully acknowledge the support of the UK's Science and Technology Facilities Council (STFC).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

A. SEAL Hyperparameters

As explained in the main text, SEAL introduces additional hyperparameters that need to be set prior to training. The first parameter is λ, which is common to both GSEAL and δSEAL, and characterizes the relative strength of the symmetry penalty compared to the data fit term. The second parameter is $\beta_{\max}$, which sets the maximal boost an input can be transformed by while calculating GSEAL during training. The effects of choosing different values for λ and $\beta_{\max}$ in the toy experiments described in Sec. 3.1 are shown in Fig. 6.

As expected, larger values of $\beta_{\max}$ correspond to flatter performance curves, even if in the expense of the data-fit. Larger values of λ are also associated with increased invariance, however are less correlated with the turning point of the curve. While $\lambda > 100$ are not seen the plots, those corresponded to worse performance than $\lambda = 100$ for all models. For small $\beta_{\max}$, we expect δSEAL and GSEAL to approach each other provided that $\lambda_{\text{GSEAL}} \approx \lambda_{\delta\text{SEAL}}/\beta^2_{\max}$, as is confirmed in plot 6c. The match is better for small λ and small boosts applied to the test inputs.

Overall, the test performance in-distribution depends very weakly on the particular choices of λ and β. This implies that similarly good fits to the train and test data can be found at various levels of invariance. More dramatic differences are only apparent for $\lambda > 10$ for GSEAL with $\beta_{\max} = 0.95$, and δSEAL with $\lambda = 100$, which is interestingly better than its weaker counterparts when the symmetry is exact.

B. Timing Comparison

The training time and evaluation time for our transformer model and for PELICAN are shown in Table 2. These are measured for a single batch processing step containing 256 jets on a single A100 GPU, as averaged over one epoch.

	Train Step [s]	Evaluation Step [s]
Baseline	0.03	0.005
Baseline + GSEAL	0.06	0.005
Baseline + δSEAL	0.06	0.005
PELICAN	0.4	0.2

Table 2. Time for training and evaluation per batch of 256 jets on a single GPU.

References

Aad, G. et al. Optimisation of large-radius jet reconstruction for the ATLAS detector in 13 TeV proton–proton collisions. *Eur. Phys. J. C*, 81(4):334, 2021. doi: 10.1140/epjc/s10052-021-09054-3.

Aad, G. et al. Transforming jet flavour tagging at ATLAS. 2025.

Agarap, A. F. Deep learning using rectified linear units (relu), 2019. URL https://arxiv.org/abs/1803.08375.

Akhound-Sadegh, T., Perreault-Levasseur, L., Brandstetter, J., Welling, M., and Ravanbakhsh, S. Lie point symmetry and physics-informed networks. *Advances in Neural Information Processing Systems*, 36:42468–42481, 2023.

Almeida, L. G., Backović, M., Cliche, M., Lee, S. J., and Perelstein, M. Playing tag with ANN: boosted top identification with pattern recognition. *Journal of High Energy Physics*, 2015(7):86, jul 2015. ISSN 1029-8479. doi: 10.1007/JHEP07(2015)086. URL https://doi.org/10.1007/JHEP07(2015)086.

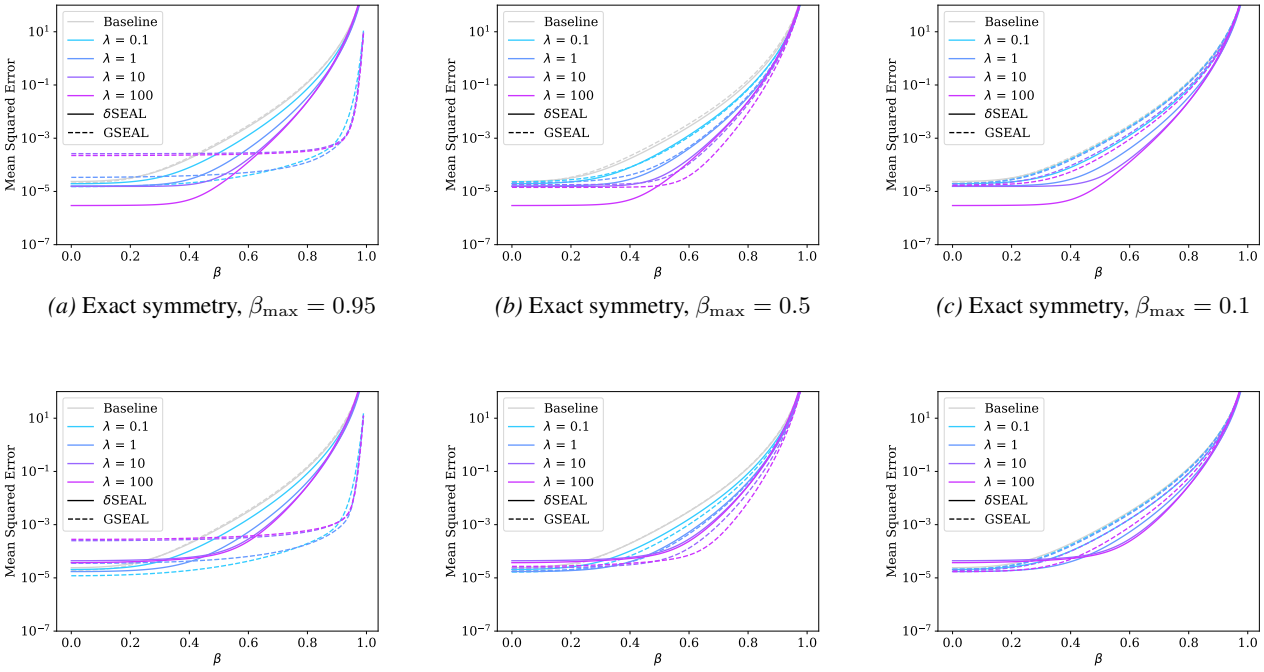
ATLAS Collaboration. Quark versus Gluon Jet Tagging Using Jet Images with the ATLAS Detector. Technical report, CERN, Geneva, 2017. URL https://cds.cern.ch/record/2275641.

ATLAS Collaboration. Graph Neural Network Jet Flavour Tagging with the ATLAS Detector. Technical report, CERN, Geneva, 2022a. URL https://cds.cern.ch/record/2811135.

ATLAS Collaboration. Constituent-Based Top-Quark Tagging with the ATLAS Detector. Technical report, CERN, Geneva, 2022b. URL https://cds.cern.ch/record/2825328.

Ball, R. D. et al. Parton distributions with LHC data. *Nucl. Phys. B*, 867:244–289, 2013. doi: 10.1016/j.nuclphysb.2012.10.003.

Bardhan, J., Agrawal, R., Tilak, A., Neeraj, C., and Mitra, S. HEP-JEPA: A foundation model for collider physics using joint embedding predictive architecture. 2 2025.



(a) Exact symmetry, $\beta_{\max} = 0.95$ *(b)* Exact symmetry, $\beta_{\max} = 0.5$ *(c)* Exact symmetry, $\beta_{\max} = 0.1$

(d) Broken symmetry, $\beta_{\max} = 0.95$ *(e)* Broken symmetry, $\beta_{\max} = 0.5$ *(f)* Broken symmetry, $\beta_{\max} = 0.1$

Figure 6. The MSE score as a function of the boost applied to the training data. In the top row the symmetry is exact (Fig. 6a–6c), while in the bottom row the symmetry is broken by a small spurion $s = (0, 0, 0, 10^{-3})$ (Fig. 6d–6f). The differential symmetry penalty is shown in solid lines, and the group-symmetry loss in dashed lines.

Barnard, J., Dawe, E. N., Dolan, M. J., and Rajcic, N. Parton shower uncertainties in jet substructure analyses with deep neural networks. *Physical Review D*, 95 (1), January 2017. ISSN 2470-0029. doi: 10.1103/physrevd.95.014018. URL http://dx.doi.org/10.1103/PhysRevD.95.014018.

Batatia, I., Geiger, M., Munoz, J., Smidt, T., Silberman, L., and Ortner, C. A general framework for equivariant neural networks on reductive lie groups. *Advances in Neural Information Processing Systems*, 36:55260–55284, 2023.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., and Pascanu, R. Relational inductive biases, deep learning, and graph networks, 2018. URL https://arxiv.org/abs/1806.01261.

Berta, P., Spousta, M., Miller, D. W., and Leitner, R. Particle-level pileup subtraction for jets and jet shapes. *JHEP*, 06:092, 2014. doi: 10.1007/JHEP06(2014)092.

Berta, P., Masetti, L., Miller, D. W., and Spousta, M. Pileup and Underlying Event Mitigation with Iterative Constituent Subtraction. *JHEP*, 08:175, 2019. doi: 10.1007/JHEP08(2019)175.

Bhattacharya, S., Guchait, M., and Vijay, A. H. Boosted top quark tagging and polarization measurement using machine learning. *Physical Review D*, 105(4), February 2022. ISSN 2470-0029. doi: 10.1103/physrevd.105.042005. URL http://dx.doi.org/10.1103/PhysRevD.105.042005.

Bhimji, W., Harris, C., Mikuni, V., and Nachman, B. OmniLearned: A Foundation Model Framework for All Tasks Involving Jet Physics. 10 2025.

Birk, J., Hallin, A., and Kasieczka, G. OmniJet-α: The first cross-task foundation model for particle physics. 3 2024.

Bogatskiy, A., Hoffman, T., Miller, D. W., Offermann, J. T., and Liu, X. Explainable equivariant neural networks for particle physics: PELICAN. *JHEP*, 03:113, 2024. doi: 10.1007/JHEP03(2024)113.

Bols, E., Kieseler, J., Verzetti, M., Stoye, M., and Stakia, A. Jet flavour classification using deepjet. *Journal of Instrumentation*, 15(12):P12012–P12012, December 2020. ISSN 1748-0221. doi: 10.1088/1748-0221/15/12/

p12012. URL http://dx.doi.org/10.1088/1748-0221/15/12/P12012.

Buckley, A. ATLAS Pythia 8 tunes to 7 TeV data. In *6th International Workshop on Multiple Partonic Interactions at the LHC*, pp. 29, 12 2014.

Cacciari, M. and Salam, G. P. Dispelling the N^3 myth for the k_t jet-finder. *Phys. Lett.*, B641:57–61, 2006. doi: 10.1016/j.physletb.2006.08.037.

Cacciari, M., Salam, G. P., and Soyez, G. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008. doi: 10.1088/1126-6708/2008/04/063.

Cacciari, M., Salam, G. P., and Soyez, G. FastJet User Manual. *Eur. Phys. J.*, C72:1896, 2012. doi: 10.1140/epjc/s10052-012-1896-2.

Cacciari, M., Salam, G. P., and Soyez, G. SoftKiller, a particle-level pileup removal method. *Eur. Phys. J. C*, 75 (2):59, 2015. doi: 10.1140/epjc/s10052-015-3267-2.

Chen, Y.-C. J., Chiang, C.-W., Cottin, G., and Shih, D. Boosted W and Z tagging with jet charge and deep learning. *Phys. Rev. D*, 101(5):053001, 2020. doi: 10.1103/PhysRevD.101.053001.

Chen, Z.-E., Chiang, C.-W., and Hsieh, F.-Y. Improving the performance of weak supervision searches using data augmentation. *JHEP*, 09:169, 2025. doi: 10.1007/JHEP09(2025)169.

Chien, Y.-T. and Elayavalli, R. K. Probing heavy ion collisions using quark and gluon jet substructure, 2018. URL https://arxiv.org/abs/1803.03589.

Choi, S., Lee, S. J., and Perelstein, M. Infrared safety of a neural-net top tagging algorithm. *Journal of High Energy Physics*, 2019(2), February 2019. ISSN 1029-8479. doi: 10.1007/jhep02(2019)132. URL http://dx.doi.org/10.1007/JHEP02(2019)132.

de Oliveira, L., Kagan, M., Mackey, L., Nachman, B., and Schwartzman, A. Jet-images — deep learning edition. *Journal of High Energy Physics*, 2016 (7), July 2016. ISSN 1029-8479. doi: 10.1007/jhep07(2016)069. URL http://dx.doi.org/10.1007/JHEP07(2016)069.

Egan, S., Fedorko, W., Lister, A., Pearkes, J., and Gay, C. Long short-term memory (lstm) networks with jet constituents for boosted top tagging at the lhc, 2017. URL https://arxiv.org/abs/1711.09059.

Elhag, A. A., Rusch, T. K., Giovanni, F. D., and Bronstein, M. Relaxed equivariance via multitask learning, 2024. URL https://arxiv.org/abs/2410.17878.

Feickert, M. and Nachman, B. A Living Review of Machine Learning for Particle Physics. 2 2021.

Fraser, K. and Schwartz, M. D. Jet charge and machine learning. *Journal of High Energy Physics*, 2018 (10), October 2018. ISSN 1029-8479. doi: 10.1007/jhep10(2018)093. URL http://dx.doi.org/10.1007/JHEP10(2018)093.

Gerken, J. E., Carlsson, O., Linander, H., Ohlsson, F., Petersson, C., and Persson, D. Equivariance versus augmentation for spherical images, 2022. URL https://arxiv.org/abs/2202.03990.

Golling, T., Heinrich, L., Kagan, M., Klein, S., Leigh, M., Osadchy, M., and Raine, J. A. Masked particle modeling on sets: towards self-supervised high energy physics foundation models. *Mach. Learn. Sci. Tech.*, 5 (3):035074, 2024. doi: 10.1088/2632-2153/ad64a8.

Gong, S., Meng, Q., Zhang, J., Qu, H., Li, C., Qian, S., Du, W., Ma, Z.-M., and Liu, T.-Y. An efficient lorentz equivariant graph neural network for jet tagging. *Journal of High Energy Physics*, 2022(7), July 2022. ISSN 1029-8479. doi: 10.1007/jhep07(2022)030. URL http://dx.doi.org/10.1007/JHEP07(2022)030.

Guest, D., Collado, J., Baldi, P., Hsu, S.-C., Urban, G., and Whiteson, D. Jet flavor classification in high-energy physics with deep neural networks. *Physical Review D*, 94(11), December 2016. ISSN 2470-0029. doi: 10.1103/physrevd.94.112002. URL http://dx.doi.org/10.1103/PhysRevD.94.112002.

Haber, H. E. Explicit Form for the Most General Lorentz Transformation Revisited. *Symmetry*, 16(9):1155, 2024. doi: 10.3390/sym16091155.

Harris, P., Kagan, M., Krupa, J., Maier, B., and Woodward, N. Re-Simulation-based Self-Supervised Learning for Pre-Training Foundation Models. 3 2024.

Hendrycks, D. and Gimpel, K. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL http://arxiv.org/abs/1606.08415.

Iglesias, G., Talavera, E., Gonzalez-Prieto, A., Mozo, A., and Gomez-Canaval, S. Data augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing and Applications*, 35(14): 10123–10145, March 2023. ISSN 1433-3058. doi: 10.1007/s00521-023-08459-3. URL http://dx.doi.org/10.1007/s00521-023-08459-3.

Kasieczka, G., Plehn, T., Russell, M., and Schell, T. Deep-learning top taggers or the end of qcd? *Journal of*

High Energy Physics, 2017(5), May 2017. ISSN 1029-8479. doi: 10.1007/jhep05(2017)006. URL http://dx.doi.org/10.1007/JHEP05(2017)006.

Komiske, P. T., Metodiev, E. M., and Schwartz, M. D. Deep learning in color: towards automated quark/gluon jet discrimination. *Journal of High Energy Physics*, 2017 (1), January 2017. ISSN 1029-8479. doi: 10.1007/jhep01(2017)110. URL http://dx.doi.org/10.1007/JHEP01(2017)110.

Komiske, P. T., Metodiev, E. M., and Thaler, J. Energy flow networks: deep sets for particle jets. *Journal of High Energy Physics*, 2019(1), January 2019. ISSN 1029-8479. doi: 10.1007/jhep01(2019)121. URL http://dx.doi.org/10.1007/JHEP01(2019)121.

Larkoski, A. J., Marzani, S., Soyez, G., and Thaler, J. Soft Drop. *JHEP*, 05:146, 2014. doi: 10.1007/JHEP05(2014)146.

Leigh, M., Klein, S., Charton, F., Golling, T., Heinrich, L., Kagan, M., Ochoa, I., and Osadchy, M. Is Tokenization Needed for Masked Particle Modelling? 9 2024.

Li, J., Li, T., and Xu, F.-Z. Reconstructing boosted higgs jets from event image segmentation. *Journal of High Energy Physics*, 2021(4), April 2021. ISSN 1029-8479. doi: 10.1007/jhep04(2021)156. URL http://dx.doi.org/10.1007/JHEP04(2021)156.

Lin, J., Freytsis, M., Moult, I., and Nachman, B. Boosting $h \to b\bar{b}$ with machine learning. *Journal of High Energy Physics*, 2018(10), October 2018. ISSN 1029-8479. doi: 10.1007/jhep10(2018)101. URL http://dx.doi.org/10.1007/JHEP10(2018)101.

Macaluso, S. and Shih, D. Pulling out all the tops with computer vision and deep learning. *Journal of High Energy Physics*, 2018(10), October 2018. ISSN 1029-8479. doi: 10.1007/jhep10(2018)121. URL http://dx.doi.org/10.1007/JHEP10(2018)121.

Mikuni, V. and Canelli, F. ABCNet: An attention-based method for particle tagging. *Eur. Phys. J. Plus*, 135(6): 463, 2020. doi: 10.1140/epjp/s13360-020-00497-3.

Mikuni, V. and Canelli, F. Point cloud transformers applied to collider physics. *Mach. Learn. Sci. Tech.*, 2(3): 035027, 2021. doi: 10.1088/2632-2153/ac07f6.

Mikuni, V. and Nachman, B. Solving key challenges in collider physics with foundation models. *Phys. Rev. D*, 111(5):L051504, 2025a. doi: 10.1103/PhysRevD.111.L051504.

Mikuni, V. and Nachman, B. Method to simultaneously facilitate all jet physics tasks. *Phys. Rev. D*, 111(5): 054015, 2025b. doi: 10.1103/PhysRevD.111.054015.

Murnane, D., Thais, S., and Wong, J. Semi-Equivariant GNN Architectures for Jet Tagging. *J. Phys. Conf. Ser.*, 2438(1):012121, 2023. doi: 10.1088/1742-6596/2438/1/012121.

Nabat, S., Ghosh, A., Witkowski, E., Kasieczka, G., and Whiteson, D. Learning broken symmetries with approximate invariance. *Phys. Rev. D*, 111(7):072002, 2025. doi: 10.1103/PhysRevD.111.072002.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Qiu, S., Han, S., Ju, X., Nachman, B., and Wang, H. Holistic approach to predicting top quark kinematic properties with the covariant particle transformer. *Phys. Rev. D*, 107(11):114029, 2023. doi: 10.1103/PhysRevD.107.114029.

Qu, H. and Gouskos, L. Jet tagging via particle clouds. *Physical Review D*, 101(5), March 2020. ISSN 2470-0029. doi: 10.1103/physrevd.101.056019. URL http://dx.doi.org/10.1103/PhysRevD.101.056019.

Qu, H., Li, C., and Qian, S. Particle transformer for jet tagging, 2024. URL https://arxiv.org/abs/2202.03772.

Quiroga, F., Ronchetti, F., Lanzarini, L., and Bariviera, A. F. *Revisiting Data Augmentation for Rotational Invariance in Convolutional Neural Networks*, pp. 127–141. Springer International Publishing, March 2019. ISBN 9783030154134. doi: 10.1007/978-3-030-15413-4_10. URL http://dx.doi.org/10.1007/978-3-030-15413-4_10.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

Shlomi, J., Battaglia, P., and Vlimant, J.-R. Graph neural networks in particle physics. *Machine Learning: Science and Technology*, 2(2):021001, January 2021. ISSN 2632-2153. doi: 10.1088/2632-2153/abbf9a. URL http://dx.doi.org/10.1088/2632-2153/abbf9a.

Spinner, J., Bresó, V., de Haan, P., Plehn, T., Thaler, J., and Brehmer, J. Lorentz-Equivariant Geometric Algebra Transformers for High-Energy Physics. 10 2024.

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. 5 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Villadamigo, J. M., Frederix, R., Plehn, T., Vitos, T., and Winterhalder, R. FASTColor – Full-color Amplitude Surrogate Toolkit for QCD. 9 2025.