Loss Landscape Geometry and the Learning of Symmetries: Or, What Influence Functions Reveal About Memorization and Generalization

James Amarel[1], **Robyn Miller, Nicolas Hengartner, Benjamin Migliori, Emily Casleton, Alexei Skurikhin, Earl Lawrence, Gerd J. Kunde**

Los Alamos National Laboratory, Los Alamos, NM 87545
[1]jlamarel@lanl.gov

Abstract

We introduce a diagnostic for symmetry learning in PDE surrogates: the influence function computed across symmetry-related states. On compressible Euler flows, our diagnostic reveals that a UNet exhibits partial but unstable influence across square group actions and translations, whereas a ViT reaches lower prediction error yet shows largely orthogonal updates across orbits. This exposes an optimization-symmetry tradeoff: stronger inductive biases promote data efficiency but can couple updates rigidly; flexible architectures optimize easily but ignore physical structure. Our diagnostic offers a reproducible test for whether training dynamics propagate information across symmetry orbits, a necessary ingredient for robust generalization in scientific machine learning.

1 Introduction

Deep learning emulators for partial differential equation (PDE) solvers routinely achieve impressive in-distribution accuracy (Brandstetter, Worrall, and Welling 2023; Herde et al. 2024; Takamoto et al. 2024; Lippe et al. 2023; Gupta and Brandstetter 2022; Ohana et al. 2025), yet they often fail to respect the fundamental symmetries of the governing equations (Akhound-Sadegh et al. 2023; Gregory et al. 2024; Gruver et al. 2024). This limitation undermines their ability to extrapolate and generalize, raising the question: are such models truly learning physics, or merely fitting correlations present in the training data? Explaining this gap requires probing not just the outputs, but also the learning dynamics (Fort et al. 2020; Zhao et al. 2024).

Symmetries of the Euler equations, namely translations, rotations, reflections, scalings, and Galilean boosts, organize the solution space into orbits whose members are physically equivalent (Brandstetter, Welling, and Worrall 2022). A model that has internalized the solution operator should propagate information seamlessly across these orbits: gradients of the loss with respect to parameters, evaluated on symmetry-related inputs, should align; without such coherence, the resulting loss differentials do not constructively influence one another, rendering the orbit decoupled. Measuring cross-influence offers a diagnostic beyond standard forward-pass equivariance checks, exposing the degree to which training updates are physically consistent.

If influence across group actions presents only weakly, the model is memorizing localized patterns rather than learning physical processes (Arpit et al. 2017; He and Su 2020; Chatterjee 2020). Conversely, persistent gradient coherence signals that the network has learned to couple symmetry-related states, consistent with the behavior of a true solution operator. Our symmetry-aware gradient diagnostic therefore quantifies a model's ability to generalize across orbits, providing a principled tool to assess how architectural choices, loss design, and inductive biases promote, or hinder, robust generalization.

2 Contributions

This work extends a previously developed gradient-based explainability framework (Amarel et al. 2025) to examine why data-driven PDE emulators often fail to learn and exploit physical symmetries. We introduce a geometry-aware, symmetry-conditioned gradient-influence diagnostic that probes how training updates propagate across symmetry group orbits, specifically the dihedral group of rotations and reflections, in addition to specific discrete translations, and pair this analysis with forward equivariance error tests to produce a coherent audit of symmetry learning.

As Figure 1 shows, the ViT's accuracy advantage over a UNet on compressible Euler flow vanishes once symmetry transformations are applied. This empirical collapse motivates the analyses that follows, where we quantify how gradient coherence tracks the flow of information through a symmetry orbit. Together, these diagnostics bridge predictive metrics with the underlying learning dynamics that govern generalization.

Our analysis reveals that high predictive accuracy can coexist with disrespect of symmetry not only in representation space but also in the local geometry, which may not support a coherent update structure across symmetry-related inputs. We further discuss the trade-off between optimizability and constraint, shedding light on how architectural inductive bias can improve symmetry coupling at the cost of training stability. Together, these findings provide a practical tool for interpreting and validating scientific machine learning models, advancing the explainable artificial intelligence agenda of probing and re-engineering model behavior to foster knowledge-driven development, particularly in weather and climate modeling.

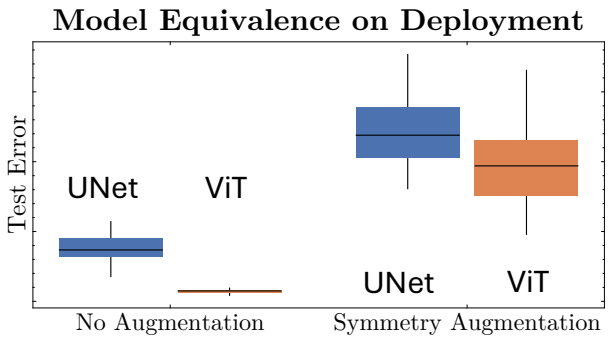

Figure 1: Test error distributions for a UNet and a ViT trained on compressible Euler flows (lower is better). Although the ViT achieves markedly lower error on untransformed data, its performance degenerates to that of a nominally weaker UNet when evaluated on physical symmetry-transformed inputs. The "No Augmentation" column corresponds to identity inputs, while "Symmetry Augmented" denotes the dihedral group-averaged error (Figure 2). Our work seeks to explain this symmetry-blind generalization gap in the ViT through curvature-adjusted gradient influence analysis.

3 Method

We compare a UNet (13M parameters, 4 down-sampling blocks, 24 embedding channels) and a Vision Transformer (ViT; 5M parameters, 6 layers, 256 channels) trained as emulators for two-dimensional compressible Euler flows from PDEGym (Herde et al. 2024). For data, we selected three classes of Riemann-type initial conditions (CE-RP, CE-RPUI, CE-CRP), each with $5{,}000$ trajectories of 16 time steps. Each state snapshot is a 128×128 grid of mass density, Cartesian momentum density, and energy density. Models were trained autoregressively to emulate the Euler evolution operator.

Optimization used Adam with learning rate 5×10^{-4} and weight decay $\lambda = 10^{-6}$ on mini-batches of $N = 64$ transitions. The cost function was a scaled mean-squared error (SMSE), that normalizes errors by channel RMS to balance large and small-amplitude features, ensuring shocks and wavefronts are captured while retaining sensitivity to quiescent flows, in addition to rendering dimensionless the influence matrix of interest. Both models were trained in distributed mode on two 40GB A100 GPUs using Lux.jl (Pal 2023a,b) and Zygote.jl (Innes 2018), with three seeds controlling initialization and dataset splits. Results are reported with quantile range bars to capture variability across seeds and mini-batches. Despite having fewer parameters, the ViT consistently outperforms the UNet after 90 epochs (Amarel et al. 2025).

To evaluate our models, we compute the influence function, which can be expressed as the Lie derivative of the cost along gradient directions induced by individual training examples. Let $X^\mu = \chi^{\mu\nu}\partial_\nu C_x$ denotes the vector field generated by the loss evaluated on an example x. The influence of this update on the loss evaluated at the transformed input gx is given by

$$L_X C_{gx} = (\partial_\mu C_x)\,\chi^{\mu\nu}\,(\partial_\nu C_{gx}), \tag{1}$$

where $\chi_{\mu\nu} = \eta_{\mu\nu} + \lambda\delta_{\mu\nu}$ is the regularized neural tangent kernel metric (Jacot, Gabriel, and Hongler 2020), and $\chi^{\mu\nu}$ denote the elements of χ^{-1}; Equivalently, the influence function can be defined as a metric-weighted overlap between gradients derived from the cost evaluated on an example x and the transformed counterpart gx. Einstein summation convention is implicit and we use standard index raising notation from differential geometry. In regression, the neural tangent kernel plays the role of a Fisher-information analog by supplying the Jacobian-induced metric on parameter space (Martens 2020). For each model seed, the influence function is evaluated across three test mini-batches comprising full trajectories for each of the three training-time classes of initial conditions. The resulting influence matrices are scaled by their Frobenius norms and subsequently averaged. In practice, χ is applied via a Krylov.jl (Montoison and Orban 2023) matrix-free solver, yielding a measure of gradient alignment sensitive to the local geometry of the loss surface (Fort and Ganguli 2019; Zielinski, Krishnan, and Chatterjee 2020)

The primary limitation of our analysis is computational. Low error tolerance solutions for a mini-batch of influence matrices in the χ-metric require significant compute resources, making it impractical to compute the full normalizing denominator required to obtain the cosine angle between gradients. Because of this limitation, prior studies have relied on uncontrolled approximations (TransferLab 2024; George 2021; Martens and Grosse 2020). In contrast, we computed the action of χ with relative error tolerances of 5×10^{-3} and 5×10^{-2} for the UNet and ViT influence matrices, respectively. A further limitation is that our study considers only UNet and ViT baselines, without exploring the full symmetry group of the Euler equations, particularly geometric convolutional and Lie-symmetry-aware architectures. We focused on UNet and ViT because they represent two predominant architectural backbones, and influence functions for exactly equivariant models would remain constant over an orbit.

4 Results and Discussion

Our evaluation considers both equivariance error (forward-pass consistency under symmetry) and influence function matrix elements (alignment of parameter updates between symmetry-related inputs). The influence function reveals whether learning dynamics propagate information coherently across symmetry-related states, exposing whether a model is genuinely learning physics or merely fitting data. We find that forward error metrics alone are insufficient to characterize the extent to which our models have internalized symmetry: the UNet enforces coupling inconsistently, while the ViT converges to a symmetry-breaking solution despite superior predictive accuracy. Convergence to basins that respect the symmetries of the underlying problem is both essential for generalization and a persistent challenge for current architectures.

The results of our evaluation underscores the trade-off between inductive bias and optimization ease. The ViT readily achieves low test error, but remains symmetry-blind, whereas the UNet partially encodes symmetry at the cost of relatively frustrated late-stage training dynamics (Zhang 2019; Azulay and Weiss 2019; Kayhan and van Gemert 2020). Mini-batches exhibiting large variance in their influence function indicate a type-II gradient misalignment, in which the mean gradient direction is inconsistent with individual update directions (Wang et al. 2025). Conflicting influence not only slows optimization but also disrupts the coherent parameter updates required for symmetry learning. This contrast illustrates why hard equivariant constraints can hinder convergence, while unconstrained models converge rapidly but fail to generalize beyond the training distribution, emphasizing the utility of approximately constrained modeling (Wang, Walters, and Yu 2022).

4.1 Dihedral Group

We analyze model behavior under the action of the dihedral group D_4 at the first step in the autoregressive evolution of compressible Euler flow. At this time, we expect a trained model to perform equivalently on D_4-rotated inputs because the data generating process for Riemann initial conditions is itself D_4 symmetric. Failure to capture the governing physics at the initial step is especially detrimental, as early inconsistencies propagate and amplify across subsequent rollout steps.

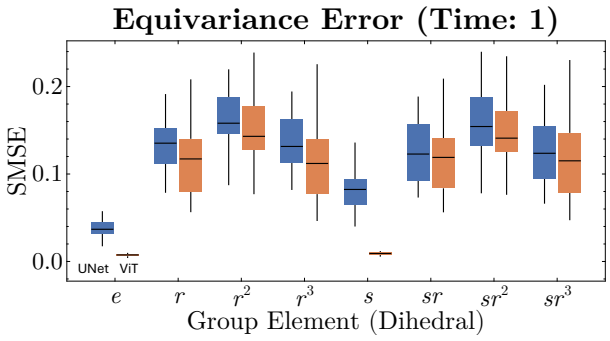

Figure 2: Box plot of forward-pass equivariance error under D_4 rotations and reflections. Despite markedly different test accuracy on untransformed data, both models show similar error on dihedral group rotated inputs (left box: UNet; right box: ViT). Medians and whiskers denote the Tukey box-and-whisker summary.

Figure 2 shows the SMSE of the UNet and ViT when their outputs are transformed along the D_4 orbit, which is generated by counter-clockwise $\pi/2$ rotations r and reflections s about the vertical axis; e denotes the identity transformation. Despite the ViT's superior performance on untransformed test data, both models exhibit comparable equivariance errors once inputs are rotated or reflected. The ViT does not show a substantial rise in error under reflections, but overall neither model demonstrates consistent dihedral symmetry. This failure reflects their lack of inductive biases; awareness

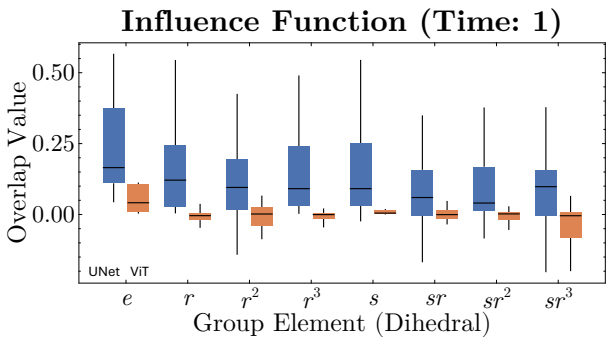

Figure 3: Box plot of influence across x and gx for $g \in D_4$. ViT overlaps concentrate near zero, which demonstrates that gradient updates do not propagate across symmetry orbits, whereas UNet shows larger but highly variable overlap, indicating partial yet inconsistent coupling. Medians and whiskers denote the Tukey box-and-whisker summary.

of rotation and reflection must be inferred from data. Figure 3 probes the learning dynamics by examining the influence across dihedral transformations. The ViT shows consistently minimal influence across gradients computed on x and gx, with $g \in D_4$. This demonstrates that ViT late-stage training dynamics treat symmetry-related states as unrelated problems. In other words, while the ViT fits the untransformed distribution well, its parameter updates fail to propagate information across the orbit, explaining its poor equivariance generalization. In contrast, the UNet shows larger but highly variable influence values. This variance indicates inconsistent gradient alignment across rotations and reflections. Such instability reflects the rigidity of convolutional features. This unstable coupling slows convergence and frustrates optimization.

Together, Figure 2 and Figure 3 explain poor dihedral generalization despite good raw accuracy. Neither architecture internalizes dihedral symmetry in its learning dynamics. For the ViT, positional encodings create a fundamental mismatch with rotation; for the UNet, convolutional filters confer only translation symmetry, leaving rotations to be learned opportunistically from local data. Furthermore, the UNet architecture can respect only a subset of the translation group on account of downsampling layers.

4.2 Translation Group

For the translation group, we evaluate purely horizontal and vertical translations at the latest time that is not on the boundary of the time domain seen during training. At this stage, it is most reasonable to expect the model to have learned translation equivariance: prolonged mixing tends to render the dataset statistically homogeneous in feature space. Because the governing PDE applies uniformly across space, learned dynamics should be equivariant under translations. In any given flow snapshot, wave interactions occupy only a few localized regions, yet they may arise at arbitrary spatial locations. A model that learns translational equivariance will therefore treat such interactions consistently wher-

ever they occur, ensuring that these rare but critical events are captured and enabling accurate long-time extrapolation.

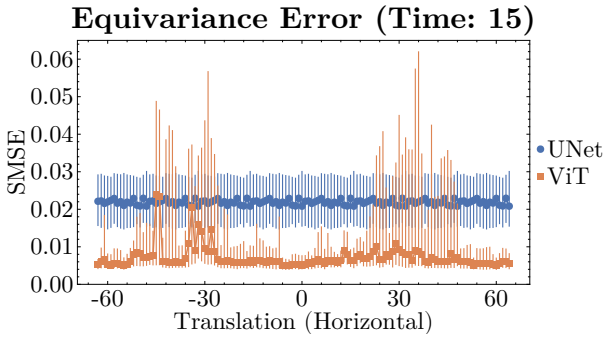

Figure 4: Forward-pass equivariance error under horizontal translations. The UNet, which is partially translation equivariant by design, maintains relatively consistent error across shifts. The ViT achieves lower average error overall but exhibits sharp spikes at specific translation values where its learned representation fails to preserve translation symmetry. Markers represent median values and the whiskers indicate neighboring quantiles.

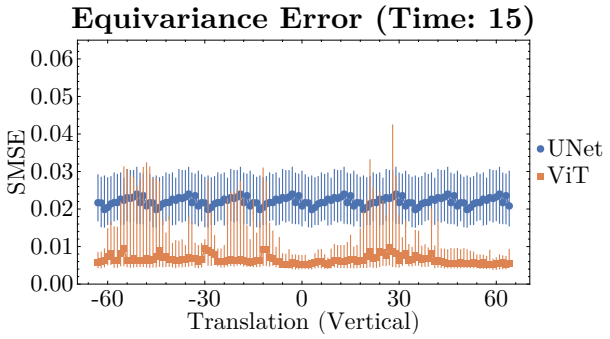

Figure 5: Forward-pass equivariance error under vertical translations. Both models exhibit incomplete translation equivariance, as in Figure 4, but comparison with the horizontal case reveals horizontal-vertical symmetry breaking, i.e., anisotropic treatment of spatial directions. Markers represent median values and the whiskers indicate neighboring quantiles.

Figure 4 and Figure 5 report equivariance error under horizontal and vertical translations, respectively. Across both directions, the ViT typically attains lower SMSE than the UNet, apart from exceptional inputs that produce isolated spikes. These spikes indicate sharp variations in the local loss landscape near the converged ViT parameters, indicating failure to learn translation equivariance.

Influence values shown in Figure 6 and Figure 7 further distinguish the two architectures. The UNet exhibits consistently larger, though variable, overlap values, indicating partial but variable coupling of parameter updates across translated states. The ViT, by contrast, exhibits negligible influence across shifted inputs, with a systematic asymmetry: vertical translations exhibit a greater typical overlap than horizontal translations, likely reflecting biases in patch embeddings and positional encodings. Notably, some equivariance error peaks coincide with influence resonances, consistent with the fact that the influence function matrix elements encode the local geometry of the loss surface. Other error peaks, however, occur without gradient coherence, revealing cases where the model simply fails to propagate information provided by symmetry.

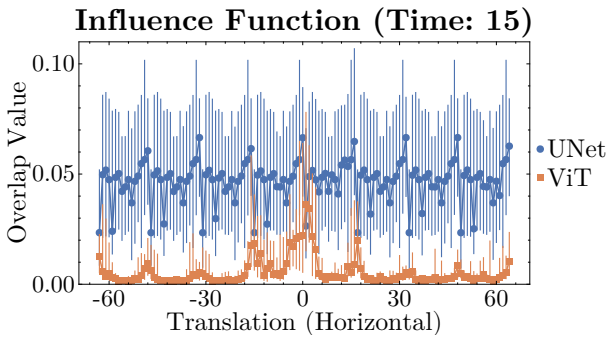

Figure 6: Influence values across horizontal translations. The UNet exhibits larger but highly variable gradient overlap, indicating partial update propagation across shifted inputs, with periodicity inherited from its convolutional and down-sampling layers. In contrast, ViT overlaps decay rapidly away from the origin, apart from a supressed periodic contribution, revealing minimal cross-shift coherence in its late-stage training dynamics.

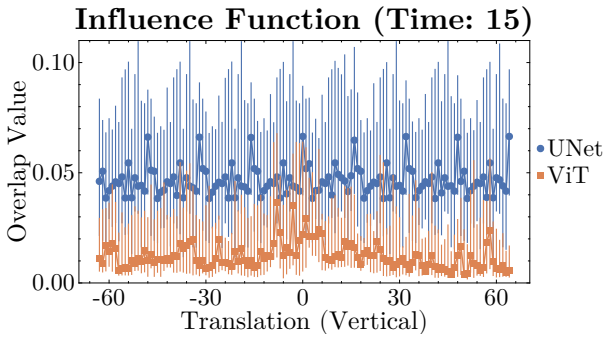

Figure 7: Influence values across vertical translations. Patterns mirror the horizontal case of Figure 6, except for an overall increase in both UNet and ViT influence values. This suggests that the ViT's reduced susceptibility to equivariance error spikes under horizontal translations, relative to vertical translations, is a result of greater gradient coherence across translations in the vertical direction.

4.3 Galilean Boosts

Galilean boosts impose a particularly restrictive symmetry, as exact equivariance is possible only for affine layers, while generic nonlinear architectures can at best satisfy this constraint approximately (Wang, Walters, and Yu 2022); in future work, we will consider the response to small boosts.

4.4 Scaling

We do not consider the scaling symmetry of continuum Navier-Stokes, because it is generically broken by discretization, numerical regularization, and coarse graining of the data.

5 Related Work

Encoding symmetries as inductive biases has a long tradition in geometric deep learning. Group-equivariant CNNs generalize convolution to arbitrary groups with weight sharing across orbits (Cohen and Welling 2016). Steerable CNNs make these ideas explicit by parameterizing kernels in group-steerable bases, including variants for volumetric data (Cohen and Welling 2017; Weiler et al. 2018). For non-grid domains, tensor field networks (Thomas et al. 2018) and E(n)-equivariant graph neural networks (Garcia Satorras, Hoogeboom, and Welling 2021) extend manifest symmetry compliance to point clouds and molecules. In practice, scientific data rarely obey exact symmetries; boundary conditions, material inhomogeneities, grid discretization, and measurement noise introduce systematic symmetry breaking. This motivates research into approximate (Wang, Walters, and Yu 2022) and relaxed (Finzi, Benton, and Wilson 2021) attainment of equivariance. Furthermore, contemporary large-scale weather and climate models such as Aurora (Bodnar et al. 2024) and ClimaX (Nguyen et al. 2023) rely on data-driven learning of symmetry. Empirical and theoretical results substantiate this strategy: softening hard constraints can improve optimization and accuracy (Wang, Walters, and Yu 2022). Relatedly, classic studies show that modern CNNs can be brittle to small shifts and rotations (Azulay and Weiss 2019; Zhang 2019; Kayhan and van Gemert 2020).

In cases where a model lacking manifest symmetry is selected, probes are needed to both quantify and explain the degree of equivariance achieved after training. Several diagnostics test whether models learn symmetry, e.g., forward-pass checks to evaluate equivariance error under group actions (Canez et al. 2024; Xie and Smidt 2025) and the Lie derivative metric, which quantifies infinitesimal equivariance with layerwise decomposition (Gruver et al. 2022). Yet, spot tests with these metrics are insufficient to explain the mechanism underlying symmetry learning. To better understand the degree to which a given model exhibits equivariance, one can relate observed behavior to architectural choices and training dynamics. A central approach in explainable AI is to interrogate the loss and its derivatives to connect predictive behavior with training signals. Related work on gradient geometry links cross-example parameter update structure to out-of-sample performance: stiffness and coherent gradients capture alignment, and local elasticity studies stability under SGD updates on distant samples (Fort et al. 2020; Chatterjee 2020; He and Su 2020). Influence functions trace predictions to training data but are delicate in deep, non-convex regimes, motivating curvature-aware variants (Koh and Liang 2020; Basu, You, and Feizi 2020; Jacot, Gabriel, and Hongler 2020; Fort and Ganguli 2019).

To our knowledge, we present the first measurements of the influence between inputs and their symmetry transformed counterparts, which allows us to assess whether training updates propagate across symmetry orbits. In this work, generalization refers to the standard notion of test-risk on unseen data; our focus is on equivariance consistency and its mechanistic underpinnings. The proposed orbit-wise gradient coherence is a local property of the trained model's loss landscape. Exact equivariance implies strong coherence, but the converse need not hold. We therefore use coherence as a diagnostic that training updates couple symmetry-related states, and we relate it empirically to forward equivariance error across dihedral and translational transformations. Together, these components provide a concise framework for evaluating symmetry-consistent behavior via both forward-pass consistency and probes of the learning dynamics, including settings where equivariance is only approximate or is learned implicitly by flexible backbones.

6 Conclusion

Our analysis complements recent (Canez et al. 2024) lessons on the role of inductive biases in deep learning. Hard constraints such as equivariance offer data efficiency and principled generalization, yet our gradient diagnostics reveal that they often impose optimization bottlenecks: parameter updates become rigidly coupled across symmetry orbits, slowing convergence and frustrating training. By contrast, unconstrained architectures such as transformers converge rapidly by freely specializing their gradients, even if this means disregarding underlying symmetries. The result is high predictive accuracy but limited physical consistency; powerful interpolators rather than genuine physics-aware models.

Furthermore, our influence function analysis reframes this trade-off by exposing whether a model propagates learning coherently across symmetry-related states or merely memorizes surface correlations. From this perspective, apparent accuracy without gradient coherence signals fragile generalization. Looking forward, these diagnostics motivate the development of approximate or relaxed symmetry methods that preserve enough structure to guide generalization while retaining the flexibility needed for efficient optimization, potentially reconciling the scalability of transformers with the principled construction of equivariant models.

Beyond technical contributions, our framework aims to strengthen trust in scientific machine learning by clarifying when models truly learn physics, while also underscoring the risks of misuse if surrogate predictions are deployed without such diagnostic safeguards.

7 Acknowledgments

Research presented in this report was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number(s) 20250637DI, 20250638DI, and 20250639DI. This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001. It is published under LA-UR-25-29466.

References

Akhound-Sadegh, T.; Perreault-Levasseur, L.; Brandstetter, J.; Welling, M.; and Ravanbakhsh, S. 2023. Lie Point Symmetry and Physics Informed Networks. arXiv:2311.04293.

Amarel, J.; Hengartner, N.; Miller, R.; Singh, K.; Mansingh, S.; Mohan, A.; Migliori, B.; Casleton, E.; Skurikhin, A.; Lawrence, E.; and Kunde, G. J. 2025. Generalization vs. Memorization in Autoregressive Deep Learning: Or, Examining Temporal Decay of Gradient Coherence. arXiv:2509.00024.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; and Lacoste-Julien, S. 2017. A Closer Look at Memorization in Deep Networks. arXiv:1706.05394.

Azulay, A.; and Weiss, Y. 2019. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184): 1–25.

Basu, S.; You, X.; and Feizi, S. 2020. On Second-Order Group Influence Functions for Black-Box Predictions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, 715–724.

Bodnar, C.; Bruinsma, W. P.; Lucic, A.; Stanley, M.; Vaughan, A.; Brandstetter, J.; Garvan, P.; Riechert, M.; Weyn, J. A.; Dong, H.; Gupta, J. K.; Thambiratnam, K.; Archibald, A. T.; Wu, C.-C.; Heider, E.; Welling, M.; Turner, R. E.; and Perdikaris, P. 2024. A Foundation Model for the Earth System. arXiv:2405.13063.

Brandstetter, J.; Welling, M.; and Worrall, D. E. 2022. Lie Point Symmetry Data Augmentation for Neural PDE Solvers. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2241–2256. PMLR.

Brandstetter, J.; Worrall, D.; and Welling, M. 2023. Message Passing Neural PDE Solvers. arXiv:2202.03376.

Canez, D.; Midavaine, N.; Stessen, T.; Fan, J.; Arias, S.; and Garcia, A. 2024. Effect of equivariance on training dynamics.

Chatterjee, S. 2020. Coherent Gradients: An Approach to Understanding Generalization in Gradient Descent-based Optimization. arXiv:2002.10657.

Cohen, T. S.; and Welling, M. 2016. Group Equivariant Convolutional Networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, 2990–2999.

Cohen, T. S.; and Welling, M. 2017. Steerable CNNs. In *ICLR*.

Finzi, M.; Benton, G.; and Wilson, A. G. 2021. Residual Pathway Priors for Soft Equivariance Constraints. arXiv:2112.01388.

Fort, S.; and Ganguli, S. 2019. Emergent properties of the local geometry of neural loss landscapes. arXiv:1910.05929.

Fort, S.; Nowak, P. K.; Jastrzebski, S.; and Narayanan, S. 2020. Stiffness: A New Perspective on Generalization in Neural Networks. arXiv:1901.09491.

Garcia Satorras, V.; Hoogeboom, E.; and Welling, M. 2021. E(n) Equivariant Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, 9323–9332.

George, T. 2021. NNGeometry: Easy and Fast Fisher Information Matrices and Neural Tangent Kernels in PyTorch.

Gregory, W. G.; Hogg, D. W.; Blum-Smith, B.; Arias, M. T.; Wong, K. W. K.; and Villar, S. 2024. Equivariant geometric convolutions for emulation of dynamical systems. arXiv:2305.12585.

Gruver, N.; Finzi, M.; Goldblum, M.; and Wilson, A. G. 2022. The Lie Derivative for Measuring Learned Equivariance. *arXiv:2210.02984*.

Gruver, N.; Finzi, M.; Goldblum, M.; and Wilson, A. G. 2024. The Lie Derivative for Measuring Learned Equivariance. arXiv:2210.02984.

Gupta, J. K.; and Brandstetter, J. 2022. Towards Multi-spatiotemporal-scale Generalized PDE Modeling. arXiv:2209.15616.

He, H.; and Su, W. J. 2020. The Local Elasticity of Neural Networks. arXiv:1910.06943.

Herde, M.; Raonić, B.; Rohner, T.; Käppeli, R.; Molinaro, R.; de Bézenac, E.; and Mishra, S. 2024. Poseidon: Efficient Foundation Models for PDEs. arXiv:2405.19101.

Innes, M. 2018. Don't Unroll Adjoint: Differentiating SSA-Form Programs. *CoRR*, abs/1810.07951.

Jacot, A.; Gabriel, F.; and Hongler, C. 2020. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. arXiv:1806.07572.

Kayhan, O. S.; and van Gemert, J. C. 2020. On Translation Invariance in CNNs: Convolutional Layers can Exploit Absolute Spatial Location. arXiv:2003.07064.

Koh, P. W.; and Liang, P. 2020. Understanding Black-box Predictions via Influence Functions. arXiv:1703.04730.

Lippe, P.; Veeling, B. S.; Perdikaris, P.; Turner, R. E.; and Brandstetter, J. 2023. PDE-Refiner: Achieving Accurate Long Rollouts with Neural PDE Solvers. arXiv:2308.05732.

Martens, J. 2020. New Insights and Perspectives on the Natural Gradient Method. *Journal of Machine Learning Research*, 21(146): 1–76.

Martens, J.; and Grosse, R. 2020. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. arXiv:1503.05671.

Montoison, A.; and Orban, D. 2023. Krylov.jl: A Julia basket of hand-picked Krylov methods. *Journal of Open Source Software*, 8(89): 5187.

Nguyen, T.; Brandstetter, J.; Kapoor, A.; Gupta, J. K.; and Grover, A. 2023. ClimaX: A foundation model for weather and climate. arXiv:2301.10343.

Ohana, R.; McCabe, M.; Meyer, L.; Morel, R.; Agocs, F. J.; Beneitez, M.; Berger, M.; Burkhart, B.; Burns, K.; Dalziel,

S. B.; Fielding, D. B.; Fortunato, D.; Goldberg, J. A.; Hirashima, K.; Jiang, Y.-F.; Kerswell, R. R.; Maddu, S.; Miller, J.; Mukhopadhyay, P.; Nixon, S. S.; Shen, J.; Watteaux, R.; Blancard, B. R.-S.; Rozet, F.; Parker, L. H.; Cranmer, M.; and Ho, S. 2025. The Well: a Large-Scale Collection of Diverse Physics Simulations for Machine Learning. arXiv:2412.00568.

Pal, A. 2023a. Lux: Explicit Parameterization of Deep Neural Networks in Julia.

Pal, A. 2023b. On Efficient Training & Inference of Neural Differential Equations.

Takamoto, M.; Praditia, T.; Leiteritz, R.; MacKinlay, D.; Alesiani, F.; Pflüger, D.; and Niepert, M. 2024. PDEBENCH: An Extensive Benchmark for Scientific Machine Learning. arXiv:2210.07182.

Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; and Riley, P. 2018. Tensor Field Networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds. *arXiv:1802.08219*.

TransferLab. 2024. pyDVL.

Wang, R.; Walters, R.; and Yu, R. 2022. Approximately Equivariant Networks for Imperfectly Symmetric Dynamics. arXiv:2201.11969.

Wang, S.; Bhartari, A. K.; Li, B.; and Perdikaris, P. 2025. Gradient Alignment in Physics-informed Neural Networks: A Second-Order Optimization Perspective. arXiv:2502.00604.

Weiler, M.; Geiger, M.; Welling, M.; Boomsma, W.; and Cohen, T. 2018. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. *arXiv:1807.02547*.

Xie, Y.; and Smidt, T. 2025. A Tale of Two Symmetries: Exploring the Loss Landscape of Equivariant Models. arXiv:2506.02269.

Zhang, R. 2019. Making Convolutional Networks Shift-Invariant Again. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 7324–7334. PMLR.

Zhao, B.; Gower, R. M.; Walters, R.; and Yu, R. 2024. Improving Convergence and Generalization Using Parameter Symmetries. arXiv:2305.13404.

Zielinski, P.; Krishnan, S.; and Chatterjee, S. 2020. Weak and Strong Gradient Directions: Explaining Memorization, Generalization, and Hardness of Examples at Scale. arXiv:2003.07422.