Loss Landscape Geometry of Partial Differential Equation Emulators: Or, Symmetry Learning via Gradient Alignment

James Amarel JLAMAREL@LANL.GOV
Robyn Miller ROBYNM@LANL.GOV
Nicolas Hengartner NICKH@LANL.GOV
Benjamin Migliori BEN.MIGLIORI@LANL.GOV
Emily Taylor ECASLETON@LANL.GOV
Alexei Skurikhin ALEXEI@LANL.GOV
Earl Lawrence EARL@LANL.GOV
Gerd J. Kunde G.J.KUNDE@LANL.GOV
Los Alamos National Laboratory, Los Alamos, NM 87545

Abstract

We study how neural emulators of partial differential equation solution operators learn physical symmetries from data by introducing a hat-matrix diagnostic that quantifies the alignment of parameter updates between symmetry related training examples. The diagnostic is a metric-weighted overlap of loss gradients evaluated across group orbits, giving a proximal influence function for symmetry-related examples. Our measurements of gradient alignment across both translations and rotations for models trained as autoregressive fluid-flow emulators suggest that equivariance arises when training updates propagate gradients coherently throughout symmetry orbits. This finding is based on an empirical correspondence between equivariance error and cross-orbit influence. Both our UNet and ViT architectures exhibit approximate translation equivariance, yet their gradient alignment profiles differ by uniform versus periodically concentrated influence over the orbit. On a Navier-Stokes dataset, pronounced dihedral equivariance error coincides with suppressed cross-influence, identifying the failure as due to decoupled learning of symmetry group elements. Our diagnostic is architecture-agnostic and probes a mechanism for learning symmetries from data, extending beyond forward-pass equivariance tests by directly assessing whether learning updates share information across physically equivalent configurations. We identify symmetry generalization performance with symmetry-compatible gradient transport, suggesting that evaluation of scientific machine learning models requires dynamical probes of loss landscape geometry in addition to predictive accuracy.

Keywords: Equivariance, Fluid Flow, Gradient Alignment

1. Introduction

Deep learning emulators for partial differential equation (PDE) solvers routinely achieve impressive in-distribution accuracy (Brandstetter et al., 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., 2022). 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 process (Fort et al., 2020; Zhao et al., 2024).

Symmetries of the Navier-Stokes (NS) equations, namely translations, rotations, reflections, scalings, and Galilean boosts, organize the solution space into orbits whose members are physically equivalent (Brandstetter et al., 2022). A model that has learned the solution operator will share information across these orbits: gradients of the loss with respect to parameters, evaluated on symmetry related inputs, should align, on account of equivariance; without such coherence, the resulting loss differentials do not constructively influence one another, rendering the orbit decoupled. When gradients constructively couple across examples, the respective losses are reduced sympathetically. Measuring cross-influence thus offers a diagnostic beyond standard forward-pass equivariance checks, exposing the degree to which gradient signals reflect physical expectations.

If influence across group actions presents only weakly, the model is memorizing localized patterns (Arpit et al., 2017; He and Su, 2020; Chatterjee, 2020) rather than learning physical processes. Conversely, marked gradient alignment demonstrates 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 probes a process by which models learn to generalize across orbits, providing a principled tool to assess how architectural choices, loss design, and inductive biases promote, or hinder, generalization.

We study models that have learned approximate translation equivariance, despite the data distribution of initial conditions not being stationary under translations. This data-driven attainment of generalization across the translation group is considered together with a similar setup involving the square group; the NS dataset that we consider presents initial conditions that are invariant only under the Klein four-subgroup of the square group, and it is only these symmetry transformations that are learned by our models. By considering the structural correspondence of influence profiles and equivariance error, our work contributes to better understanding the underlying process by which physical symmetries can be learned from data distributions that are not manifestly invariant to the associated symmetry group. It is important to recognize that a test example can be both out of distribution in the statistical sense and physically equivalent to an example represented by said data distribution; we use the influence to measure if gradient update signals leverage this physical equivalence.

2. 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 et al., 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 et al., 2022) and relaxed (Finzi et al., 2021)

attainment of equivariance. Furthermore, many contemporary large-scale models (Nguyen et al., 2023; Bodnar et al., 2024a) rely on data-driven learning of symmetry. Empirical and theoretical results substantiate this strategy: softening hard constraints can improve optimization and accuracy (Wang et al., 2022). Moreover, 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).

While a constructive route for guaranteed equivariance exists, this is not the regime targeted by modern large-scale scientific models. Indeed, contemporary state-of-the-art models for PDE emulation and weather forecasting largely use architectures comprised of transformer and convolutional layers that are not equivariant by design (Takamoto et al., 2024; Lippe et al., 2023; Gupta and Brandstetter, 2022; Ohana et al., 2025; Herde et al., 2024; McCabe et al., 2026; Bodnar et al., 2024b; Nguyen et al., 2023). Our focus is therefore on settings in which symmetry learning is data-driven and contingent on local update geometry rather than guaranteed a-priori. The restriction to exactly equivariant models is not necessarily optimal when realistic data and optimization difficulties are considered (Wang et al., 2022). Moreover, for architectures that enforce exact equivariance by construction, influence across symmetry orbits would be uniform by design. The purpose of this work is expressly to better understand how models learn symmetry when the inductive bias is not explicitly given to the model through, e.g., architecture design, training losses, or data augmentation.

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 the extent to which a model has learned symmetry, e.g., forward-pass checks to evaluate equivariance error under group actions (Canez et al., 2024; Xie and Smidt, 2025), characterizations of internal representations (Godfrey et al., 2022), 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. Such diagnostics characterize equivariance of the learned map, but do not probe processes underlying data-driven symmetry-learning. In contrast, our work relates observed equivariance behavior to architectural choices and training dynamics by measuring the propagation of information across symmetry related states.

A central approach in explainable artificial intelligence 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). Euclidean influence functions trace predictions to training data but are delicate in deep, non-convex regimes, motivating curvature-aware variants (Koh and Liang, 2020; Basu et al., 2020; Jacot et al., 2020; Fort and Ganguli, 2019).

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

training data, we selected 3 classes of Riemann-type initial conditions (CE-RP, CE-RPUI, CE-CRP), each with 6,500 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; for more details see our companion paper that considers time-translation symmetry, coherence over initial condition classes, and influence across physics-informed objectives (Amarel et al., 2026). To complement the compressible Euler results, we also apply our analysis to models trained on velocity fields generated by the Navier-Stokes (NS) equations using, NS-BB, NS-Gauss, and NS-Sines initial conditions. These flows occupy a qualitatively different feature space, characterized by smoother, viscosity-regularized dynamics and vorticity-dominated structure.

Optimization trajectories for CE and NS are shown in Figure 8; training is halted in the late-stage learning regime, so the hat-matrix diagnostics probe local gradient geometry after the models have entered low-test-loss basins but before learning has completely saturated. We used Adam with learning rate 5×10^{-4} and weight decay $\lambda = 10^{-4}$ on mini-batches of $N = 48$ transitions. The cost function was a scaled mean-squared error (SMSE), that normalizes errors by channel RMS to balance large and small-amplitude features, supporting the capture of shocks and wavefronts 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, 2023b,a) and Zygote.jl (Innes, 2018), with 3 seeds controlling initialization and dataset splits. Results are reported with quantile range bars to capture variability across seeds.

To evaluate our models, we compute the influence function, which can be expressed as the Lie derivative of the cost along gradient flow induced by individual test examples. By formulating the influence through the Lie derivative, we elevate the diagnostic from its traditional role as a static, final-state, measurement, to a dynamical quantity that can be meaningfully observed during training. However, due to computational expense, the present paper primarily measures influence at a single parameter point in the late-stage learning regime [see Figure 8]; only Figure 7 tracks influence at checkpointed epochs. Let $V^\mu = -\chi^{\mu\nu}\partial_\nu C_x$ denote the vector field generated by the loss evaluated on an example x. The influence of this update on the loss evaluated at the group-transformed input gx is given by[1]

$$-\mathcal{L}_V C_{gx} = (\partial_\mu C_{gx})\, \chi^{\mu\nu}\, (\partial_\nu C_x), \tag{1}$$

where the regularized Gauss-Newton metric $\chi_{\mu\nu} = \eta_{\mu\nu} + \lambda\delta_{\mu\nu}$ is comprised of both an isotropic regularization term of strength λ and η, the pullback of the Euclidean metric on function space to parameter space. Einstein summation convention is implicit and we use standard index raising notation from differential geometry; $\chi^{\mu\nu}$ denote the elements of χ^{-1} (Absil et al., 2008). Equation 1 recovers the familiar influence function definition as a metric-weighted overlap between gradients derived from the cost evaluated on an example x and the transformed counterpart gx (Fort and Ganguli, 2019). Intuitively, the influence function measures whether a gradient update induced by one example decreases (or increases) the loss of another. When applied to symmetry related inputs, it quantifies whether learning signals

1. Programmatically, the translation group action is realized via cyclic permutation of array indices. For scalar fields, the square group can be effected with standard commands for rotating and flipping images; for the velocity field one must take care to appropriately mix vector indices in addition to rotating the coordinate basis.

propagate coherently along symmetry orbits. In regression, the Gauss-Newton matrix plays the role of the Fisher-information by supplying the Jacobian-induced metric on parameter space (Martens, 2020); sandwiching the metric between Jacobians yields the (λ-regularized) hat-matrix $J_\mu \chi^{\mu\nu} J_\nu$, which reduces to the neural tangent kernel in the Euclidean limit (Huber and Ronchetti, 2009; Jacot et al., 2020). The influence function (Equation 1) can also be expressed by sandwiching the hat-matrix between prediction residuals (Amarel et al., 2026).

For each model seed, the influence function is evaluated across 18 test mini-batches comprising 3 distinct initial conditions over 16 time-steps. As a first-order sensitivity of a scalar loss to an example-induced update, the influence function has no canonical scale. Common choices include normalization by gradient magnitude, yielding the cosine similarity, normalization by trace or Frobenius norms, standardized influence functions Héritier and Ronchetti (1994); Lu et al. (1997), or alternatively leaving the response unnormalized altogether. In the experiments below, we use cosine normalization, reporting influence after division by both the perturbation and response gradient magnitudes. This normalization sets each self-response to unity, which removes the absolute scale of local sensitivities and hence isolates the cross-response magnitude, a comparison appropriate when the effects of influence are accumulated over many optimization steps.

For further discussion of data selection, model selection, an intuitive derivation of the influence function, and details on its computation, see section A, section B, section C, and section D, respectively.

4. Results

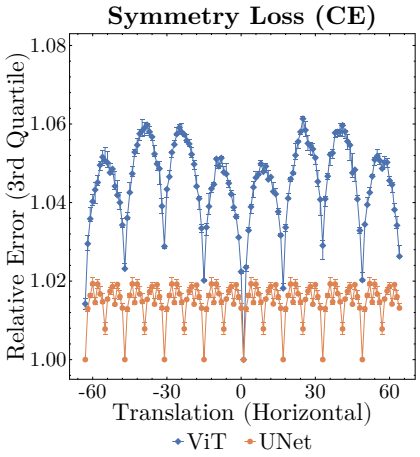
Our evaluation jointly considers equivariance error, which probes forward-pass consistency under symmetry transformations, and influence function matrix elements. The influence function reveals whether learning propagates information coherently across symmetry related states, exposing whether a model is genuinely learning symmetries or merely fitting correlations in the data. Convergence to basins that respect the symmetries of the underlying problem is both essential for generalization and a persistent challenge for current architectures. Throughout this section, we test the central hypothesis that forward-pass equivariance emerges when local update geometry propagates gradient information coherently along symmetry orbits. We find that forward-pass equivariance error is in structural correspondence with the extent to which our models and training regimes support symmetry-coherent gradients.

4.1. Translation Group

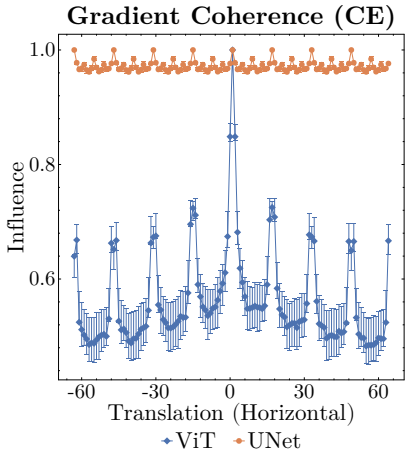
For the translation group, we first evaluate purely horizontal and vertical translations across the periodic spatial domain. Because the governing PDE applies uniformly in space, the learned dynamics should be equivariant under translations. In any given flow snapshot, nonlinear interactions occupy only a small number of localized regions, yet these structures may arise at arbitrary spatial locations. A model that learns translational equivariance will therefore treat such interactions consistently wherever they occur, a prerequisite for both capturing rare but dynamically significant events and enabling accurate long-time extrapolation.

Both architectures exhibit median relative equivariance error on the order of one percent or less, indicating that training largely confers translational symmetry generalizing capabilities. To capture the structure of those examples for which low equivariance error was not obtained, we consider the third quartile error-profile. Notably, the relative equivariance error attained by our models [see Figure 1] reflects genuine generalization over the translation group: we did not explicitly augment the training data with translated states, nor was either architecture designed to enforce exact equivariance over the entire translation group under consideration. In particular, translation stationarity is violated by the distribution from which four quadrant initial conditions for both CE-RP and CE-RPUI flows were drawn. Nevertheless, we observe nontrivial influence across translated states, indicating that training updates couple translation-related inputs, even without hard symmetry constraints.

Inspection of forward-pass equivariance error together with influence profile [see Figure 1] provides an explanatory correspondence for symmetry learning: resonant minima in the error profile coincide with structured extrema in the response, and vice-versa. Both models support cross-influence reflective of response geometry that supports nontrivial propagation across the translation orbit and thus generalization over translations is achieved, even in the absence of explicit translation augmentation or exact equivariance constraints.



(*a*) Relative equivariance error as a function of horizontal translation.



(*b*) Gradient alignment between an example and its horizontally translated state.

Figure 1: Diagnostics under horizontal translations on CE data. Markers and rangebars indicate the median and interquartile range over seeds. Lack of uniformity reflects horizontal symmetry breaking. For vertical translations, see Figure 3; for analogous results with NS data, see Figure 5.

The translation response for our ViTs [see Figure 1 and Figure 3] exhibits a pronounced axis-dependent resonance structure. Under horizontal (vertical) translations, the periodicity of wavelength 16 (64) pixels indicates that the measured influence/equivariance observables

couple anisotropically to the phase of the translation relative to the patch lattice, with different group elements emphasized in different directions. This axis-dependent anisotropy may be due to representational bias introduced by flattening the two-dimensional patch grid into a one-dimensional token sequence. While the patch embedding itself is convolutional of stride four, the subsequent permutation into tokens and channel mixing allows for a preferred ordering on spatial degrees of freedom, thereby breaking the manifest equivalence between horizontal and vertical directions, which impacts the learned response geometry. In contrast, our UNets employ convolutional operators across contracting and expanding resolutions, with four down-sampling stages. The coarsest representation has spatial extent 8×8, which promotes a comparatively smooth dependence of influence on translation, with only low-amplitude variation at shorter 8 pixel wavelengths and their subharmonics. Consistent with this architecture-level partial symmetry adherence, UNets exhibit similar influence profiles across horizontal and vertical translations.

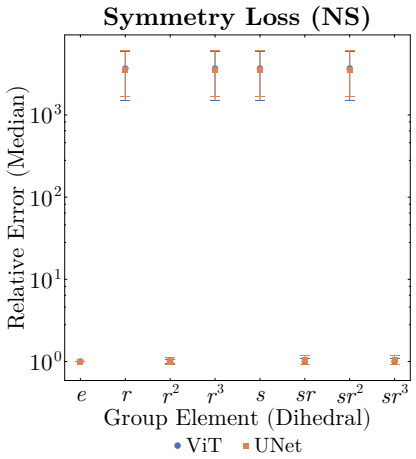
That our UNets exhibit nearly uniform gradient coherence, while our ViTs distribute influence non-uniformly yet support larger influence on privileged group elements, suggests a trade-off between inductive bias and optimization flexibility. Non-uniform coupling over the group allows our ViTs to concentrate each learning signal on a subset of translation phases, producing stronger local responses at the cost of more heterogeneous orbit-wise coupling. Less rigid coupling between translation elements for the ViT may be advantageous when paired with an overall larger peak response, as its updates are flexibly dispersed throughout the group. Indeed, low cross-influence in the Gauss-Newton geometry indicates that a gradient update derived from a given example does not appreciably change the function over a mini-batch. Hence the model can specialize on a per-example basis. Enforcing uniform coupling across all group elements may constrain the space of admissible update directions and, in some settings, impede optimization (Zhang, 2019; Azulay and Weiss, 2019; Kayhan and van Gemert, 2020) by inducing a mean gradient direction that is inconsistent with individual update directions (Wang et al., 2025). In our translation group related experiments, we observe exclusively positive median influence values, suggesting that optimization is not frustrated by conflicting updates; instead, the distinction lies in how learning signal is distributed: architectures with weaker inductive bias are free to concentrate influence on a subset of symmetry transformations, enabling rapid training loss reduction via a strong gradient response, but limiting generalization capabilities across the group.

4.2. Dihedral Group

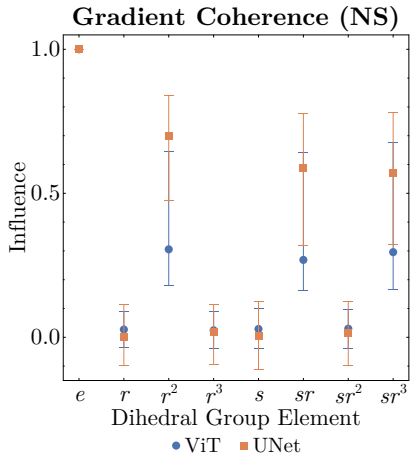
Next, we analyze symmetry learning with respect to the dihedral group D_4, the set of discrete rotations and reflections admissible under the datasets used for training. Our models generalize on those group transformations for which gradients are propagated coherently through the loss landscape. Where generalization fails, we show that the local update geometry does not support full orbit-wise coupling and visualize how data-induced anisotropies are absorbed into the learned loss landscape geometry.

Our Navier-Stokes trained models fail to learn the dihedral group D_4, a breakdown that presents across random seeds and mini-batches despite otherwise excellent pointwise accuracy and low untransformed test loss [see Figure 2]. Both architectures achieve competitive performance on the identity element e, 180 degree rotations r^2, 90 degree rotations

followed by a reflection about the vertical axis sr, and 270 degree rotations followed by such a reflection sr^3, yet their response under the remaining group actions reveals a sharp and systematic lack of symmetry adherence. In this case, the relative SMSE increases catastrophically, by nearly 10^4, indicating that the learned predictors are not even approximately equivariant when all group actions are considered. Such behavior is not entirely unexpected. The Navier-Stokes dataset carries a pronounced directional bias in feature space, inherited from the choice of initial conditions and the anisotropic, vorticity-dominated structure of the flows that is not invariant under reflections or 90 degree rotations; the directional bias renders the data distribution (our models) invariant (equivariant) only under the Klein four-subgroup V_4 of the square group.[2] What is striking, however, is that the symmetry breaking is not merely a property of the forward map. The influence structure associated with these unlearned group actions is distinguished in Figure 2: the group elements with catastrophic equivariance error are precisely those with suppressed cross-influence.



(*a*) Relative equivariance error for each D_4 group action.



(*b*) Gradient alignment between an example and its D_4-transformed state.

Figure 2: Diagnostics under D_4 symmetry on Navier-Stokes (NS) data. Markers and rangebars indicate the median and interquartile range over seeds. Nonuniformity across group elements reflects dihedral symmetry breaking; group actions with large equivariance error exhibit suppressed gradient alignment. For the corresponding CE diagnostics, see Figure 6.

Rather than facilitating a democratic sharing of coupling across the dihedral orbit, optimization drove our models into a region of the loss landscape that displays a clear de-

2. We thank Reviewer q6u2 for naming the Klein four-subgroup and sharpening our discussion on the role of the initial condition invariance group. Our results show that the invariance group is not the sole factor in determining generalization capabilities. Indeed, while our models fail to generalize across D_4, they attain approximate equivariance over translations. In both cases the initial condition distribution is not invariant to respective symmetry group; a distinguishing observable across these two scenarios is the cross-influence profile.

marcation between well-learned and challenging group elements, with the latter receiving subdominant influence. These challenging group elements receive near-zero cross-influence, indicating that training has converged to a basin with symmetry-incompatible local geometry. The checkpointed influence profiles in Figure 7 reveal how this orbit-wise separation develops throughout optimization. Although the two sectors begin with comparable cross-influence, both UNet and ViT rapidly establish and maintain stronger transfer within $V_4 \setminus e$ than into $D_4 \setminus V_4$, yielding training dynamics that support coherent gradient propagation over the learned Klein four-subgroup but not the full dihedral orbit.

We emphasize that although the dihedral symmetry is physically valid, gradient signals do not accumulate coherently over symmetry operations. Instead, data-induced anisotropies are absorbed into the local loss geometry, inducing local updates that reinforce symmetry-breaking. Such rotations are in-distribution under reasonable deployment assumptions, meaning that the observed decoupling reflects a generalization gap between model performance on the training distribution and the operational distribution, indicating limited ability to generalize over the full data manifold. Our influence probe interrogates directions that are poorly aligned with the primary drift direction traced by the training flow. Hence, symmetry related gradient signals enter only as subleading perturbations and do not accumulate coherently, explaining how models achieve excellent pointwise accuracy on in-training-distribution data while converging to basins whose local geometry explicitly breaks dihedral symmetry. Indeed, Figure 2 explains the apparent tension between low in-training-distribution test error and severe dihedral failure: under symmetry-agnostic training, the induced local update geometry does not propagate information coherently across the full D_4 orbit, so neither model class learns the D_4-equivariant extension expected of the underlying solution operator.

5. Conclusion

Our analysis sharpens recent insights on the interplay between inductive bias, optimization, and symmetry learning (Canez et al., 2024). We show that test time equivariance error tracks how the local update geometry distributes influence across group orbits. While symmetry-agnostic architectures can attain low test error, producing solutions that interpolate accurately, such models need not allocate cross-influence across symmetry groups, which inhibits their ability to learn the underlying solution operator.

By explicitly measuring cross-influence between symmetry related states, our framework reveals whether learning propagates information coherently along an orbit or instead distribute influence in a way that precludes convergence to generalizing solutions. This provides a novel diagnostic for distinguishing models that learn to exploit shared structure from those that effectively assemble collections of local estimators. Tying symmetry generalization directly to the geometry of the learned loss landscape complements standard equivariance accuracy metrics and provides a clearer criterion for when apparent performance reflects genuine physics learning. Such diagnostics are essential for building trust in scientific machine learning systems, particularly in applications requiring robustness under symmetry transformations. Looking forward, this perspective motivates the development of approximate or relaxed symmetry mechanisms that retain sufficient structure to guide generalization while preserving the flexibility needed for efficient optimization.

6. 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

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3.

Tara Akhound-Sadegh, Laurence Perreault-Levasseur, Johannes Brandstetter, Max Welling, and Siamak Ravanbakhsh. Lie point symmetry and physics informed networks, 2023. URL https://arxiv.org/abs/2311.04293.

James Amarel, Nicolas Hengartner, Robyn Miller, Kamaljeet Singh, Siddharth Mansingh, Arvind Mohan, Benjamin Migliori, Emily Casleton, Alexei Skurikhin, Earl Lawrence, and Gerd J. Kunde. Generalization vs. memorization in autoregressive deep learning: Or, examining temporal decay of gradient coherence, 2026. URL https://arxiv.org/abs/2509.00024.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks, 2017. URL https://arxiv.org/abs/1706.05394.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019. URL http://jmlr.org/papers/v20/19-519.html.

Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pages 715–724, 2020. URL https://proceedings.mlr.press/v119/basu20b.html.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Vaughan, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system, 2024a. URL https://arxiv.org/abs/2405.13063.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Vaughan, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system, 2024b. URL https://arxiv.org/abs/2405.13063.

Johannes Brandstetter, Max Welling, and Daniel E Worrall. Lie point symmetry data augmentation for neural PDE solvers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2241–2256. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/brandstetter22a.html.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural pde solvers, 2023. URL https://arxiv.org/abs/2202.03376.

Diego Canez, Nesta Midavaine, Thijs Stessen, Jiapeng Fan, Sebastian Arias, and Alejandro Garcia. Effect of equivariance on training dynamics, 2024. URL https://gram-blogposts.github.io/blog/2024/relaxed-equivariance/.

Satrajit Chatterjee. Coherent gradients: An approach to understanding generalization in gradient descent-based optimization, 2020. URL https://arxiv.org/abs/2002.10657.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, pages 2990–2999, 2016. URL https://proceedings.mlr.press/v48/cohenc16.pdf.

Taco S. Cohen and Max Welling. Steerable CNNs. In *ICLR*, 2017. URL https://arxiv.org/abs/1612.08498.

Simon Danisch and Julius Krumbiegel. Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software*, 6(65):3349, 2021. doi: 10.21105/joss.03349. URL https://doi.org/10.21105/joss.03349.

Marc Finzi, Gregory Benton, and Andrew Gordon Wilson. Residual pathway priors for soft equivariance constraints, 2021. URL https://arxiv.org/abs/2112.01388.

Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes, 2019. URL https://arxiv.org/abs/1910.05929.

Stanislav Fort, Paweł Krzysztof Nowak, Stanislaw Jastrzebski, and Srini Narayanan. Stiffness: A new perspective on generalization in neural networks, 2020. URL https://arxiv.org/abs/1901.09491.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pages 9323–9332, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Thomas George. NNGeometry: Easy and Fast Fisher Information Matrices and Neural Tangent Kernels in PyTorch, 2021. URL https://doi.org/10.5281/zenodo.4532597.

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 11893–11905. Curran Associates, Inc., 2022.

Wilson G. Gregory, David W. Hogg, Ben Blum-Smith, Maria Teresa Arias, Kaze W. K. Wong, and Soledad Villar. Equivariant geometric convolutions for emulation of dynamical systems, 2024. URL https://arxiv.org/abs/2305.12585.

Nate Gruver, Marc Finzi, Micah Goldblum, and Andrew Gordon Wilson. The lie derivative for measuring learned equivariance. *arXiv:2210.02984*, 2022. URL https://arxiv.org/abs/2210.02984.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized pde modeling, 2022. URL https://arxiv.org/abs/2209.15616.

Hangfeng He and Weijie J. Su. The local elasticity of neural networks, 2020. URL https://arxiv.org/abs/1910.06943.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for pdes, 2024. URL https://arxiv.org/abs/2405.19101.

Stéphane Héritier and Elvezio Ronchetti. Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association*, 89(427):897–904, 1994. ISSN 0162-1459. URL https://www.jstor.org/stable/2290914.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2009. ISBN 9780470129906. doi: 10.1002/9780470434697. URL https://doi.org/10.1002/9780470434697.

Michael Innes. Don't unroll adjoint: Differentiating ssa-form programs. *CoRR*, abs/1810.07951, 2018. URL http://arxiv.org/abs/1810.07951.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL https://arxiv.org/abs/1806.07572.

Osman Semih Kayhan and Jan C. van Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location, 2020. URL https://arxiv.org/abs/2003.07064.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2020. URL https://arxiv.org/abs/1703.04730.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations, 2021. URL https://arxiv.org/abs/2010.08895.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. Pde-refiner: Achieving accurate long rollouts with neural pde solvers, 2023. URL https://arxiv.org/abs/2308.05732.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL https://arxiv.org/abs/2103.14030.

Jiandong Lu, Daijin Ko, and Ted Chang. The standardized influence matrix and its applications. *Journal of the American Statistical Association*, 92(440):1572–1580, 1997. ISSN 0162-1459. URL https://www.jstor.org/stable/2965428.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00302-5. URL http://dx.doi.org/10.1038/s42256-021-00302-5.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL http://jmlr.org/papers/v21/17-678.html.

Michael McCabe, Payel Mukhopadhyay, Tanya Marwah, Bruno Regaldo-Saint Blancard, Francois Rozet, Cristiana Diaconu, Lucas Meyer, Kaze W. K. Wong, Hadi Sotoudeh, Alberto Bietti, Irina Espejo, Rio Fear, Siavash Golkar, Tom Hehir, Keiya Hirashima, Geraud Krawezik, Francois Lanusse, Rudy Morel, Ruben Ohana, Liam Parker, Mariel Pettee, Jeff Shen, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. Walrus: A cross-domain foundation model for continuum dynamics, 2026. URL https://arxiv.org/abs/2511.15684.

Alexis Montoison and Dominique Orban. Krylov.jl: A Julia basket of hand-picked Krylov methods. *Journal of Open Source Software*, 8(89):5187, 2023. doi: 10.21105/joss.05187.

Sepehr Mousavi, Shizheng Wen, Levi Lingsch, Maximilian Herde, Bogdan Raonić, and Siddhartha Mishra. Rigno: A graph-based framework for robust and accurate operator learning for pdes on arbitrary domains, 2025. URL https://arxiv.org/abs/2501.19205.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate, 2023. URL https://arxiv.org/abs/2301.10343.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Keaton Burns, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning, 2025. URL https://arxiv.org/abs/2412.00568.

Dominique Orban and Mario Arioli. *Iterative Solution of Symmetric Quasi-Definite Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

doi: 10.1137/1.9781611974737. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974737.

Avik Pal. On Efficient Training & Inference of Neural Differential Equations, 2023a.

Avik Pal. Lux: Explicit Parameterization of Deep Neural Networks in Julia, 2023b. URL https://doi.org/10.5281/zenodo.7808903.

Mahindra Singh Rautela, Alexander Most, Siddharth Mansingh, Bradley C. Love, Alexander Scheinker, Diane Oyen, Nathan Debardeleben, Earl Lawrence, and Ayan Biswas. Morph: Pde foundation models with arbitrary data modality, 2026. URL https://arxiv.org/abs/2509.21670.

Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior, 2023. URL https://arxiv.org/abs/2306.00258.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning, 2024. URL https://arxiv.org/abs/2210.07182.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. *arXiv:1802.08219*, 2018. URL https://arxiv.org/abs/1802.08219.

TransferLab. pyDVL, 2024. URL https://github.com/aai-institute/pyDVL.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics, 2022. URL https://arxiv.org/abs/2201.11969.

Sifan Wang, Ananyae Kumar Bhartari, Bowen Li, and Paris Perdikaris. Gradient alignment in physics-informed neural networks: A second-order optimization perspective, 2025. URL https://arxiv.org/abs/2502.00604.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable CNNs: Learning rotationally equivariant features in volumetric data. *arXiv:1807.02547*, 2018. URL https://arxiv.org/abs/1807.02547.

YuQing Xie and Tess Smidt. A tale of two symmetries: Exploring the loss landscape of equivariant models, 2025. URL https://arxiv.org/abs/2506.02269.

Zhanhong Ye, Xiang Huang, Leheng Chen, Hongsheng Liu, Zidong Wang, and Bin Dong. PDEformer: Towards a foundation model for one-dimensional partial differential equations. In *ICLR 2024 Workshop on AI4DifferentialEquations In Science*, 2024. URL https://openreview.net/forum?id=GLDMCwdhTK.

Richard Zhang. Making convolutional networks shift-invariant again. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7324–7334. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhang19a.html.

Bo Zhao, Robert M. Gower, Robin Walters, and Rose Yu. Improving convergence and generalization using parameter symmetries, 2024. URL https://arxiv.org/abs/2305.13404.

Software and Data

The code used in this work is available at https://github.com/lanl/PDEHats. We trained our models on the openly available dataset PDEGym Herde et al. (2024) using Lux.jl Pal (2023b,a), with Zygote.jl as our auto-differentiation backend Innes (2018). Plots in this manuscript were generated using Makie.jl Danisch and Krumbiegel (2021).

Appendix A. Data Selection.

PDEGym data is particularly valuable for studying the progression from a linear wave regime with discontinuities to fully developed turbulence, a crossover that poses computational and analytical challenges due to the presence of sharp wave-fronts and emergent nonlinear interactions. While the CE flows exhibit comparable large-scale structures, they also display qualitatively distinct behaviors. In particular, CE-RPUI initial configurations give rise to complex finger-like instabilities in the flow field that are absent or less pronounced in both CE-RP and CE-CRP. In total, we used $6,500$ trajectories for each of the 3 classes of initial conditions; for each trajectory, we used the first 16 time steps, for total of $104,000$ training pairs per initial condition class, requiring greater than 150 GB memory.

Appendix B. Model Selection.

We focus on UNet and ViT backbones because they represent the dominant scalable image-to-image paradigms currently used in neural PDE emulation and PDE foundation modeling. Convolutional backbones derived from UNets remain standard architectures for grid-based surrogate modeling and large-scale benchmarking efforts (Takamoto et al., 2024; Gupta and Brandstetter, 2022). Likewise, transformer-based backbones now underpin many recent PDE foundation models and large-scale operator learners, including Poseidon and PDEformer (Herde et al., 2024; Ye et al., 2024), in addition to several contemporary multi-physics (Rautela et al., 2026) and shifting window transformer (Liu et al., 2021) architectures. Fourier neural operators provide an important complementary class of operator-learning models (Li et al., 2021; Subramanian et al., 2023); however, because their representations are constructed from truncated Fourier modes, the resulting plane-wave basis is less naturally suited to the spatially localized discontinuities and sharp shock fronts characteristic of compressible flow evolution (Ohana et al., 2025). Consistent with this expectation, we found FNO variants to be noncompetitive with both UNet and ViT backbones on the flows

considered in this work, which is consistent with recent results by the PDEGym authors (Mousavi et al., 2025). DeepONets provide a foundational operator-learning formalism (Lu et al., 2021), but are not the prevailing architecture for contemporary large-scale PDE solution operator emulation; this assertion is consistent with the present lack of competitive DeepONet models on PDEGym.

Appendix C. Loss Sensitivity as Gradient Overlap

The influence diagnostic admits a simple interpretation as a first-order sensitivity of the loss function. Let C_p denote the cost associated with a perturbing example p, and let the induced parameter displacement be $\delta\theta_p = -\eta^{-1}\nabla_\theta C_p$, where η is the Gauss-Newton metric. The corresponding first-order change in the cost of a response example r is $C_r(\theta + \delta\theta_p) - C_r(\theta) \approx -\left\langle \nabla_\theta C_r, \eta^{-1}\nabla_\theta C_p \right\rangle$. Thus, gradient alignment measures the first-order sensitivity of example r to an update generated by example p. Positive alignment indicates that the perturbing update acts coherently on the response example, whereas weak or negative alignment indicates that the corresponding learning signals are decoupled or conflicting.

Appendix D. Determination of η^{-1}

Influence-function computations in deep learning are commonly performed either through iterative matrix-free solvers, such as conjugate-gradient methods (Orban and Arioli, 2017; Montoison and Orban, 2023), or through scalable approximations to the curvature operator (George, 2021; TransferLab, 2024). While such approaches are often necessary in high-dimensional parameter spaces, iterative methods require repeated Hessian-vector products and convergence tolerances, whereas low-rank or diagonal approximations generically do not provide controlled error estimates.

In the present setting, however, the influence observable depends only on metric-weighted overlaps between example gradients. Let $G^n_\alpha = \partial_\alpha C^n$ denote the gradient associated with mini-batch example n. Performing a QR decomposition on G yields an orthonormal basis Q that projects onto the gradient subspace. Hence, the Gauss-Newton metric can be efficiently inverted. Since the dimension of the gradient-spanned subspace is bounded by the mini-batch size rather than the total number of trainable parameters, the resulting inversion is both exact and computationally efficient, eliminating the need for uncontrolled curvature approximations or iterative parameter-space solves.

For CE and NS data, hat-matrix elements were determined for 18 different mini-batches, each of which contained 3 trajectories corresponding to distinct initial conditions over 16 time-steps, for each seed of each model architecture trained. In practice, we additionally include an ℓ_2-type regularizer corresponding to the weight-decay term used during AdamW optimization (Kingma and Ba, 2017), $\eta_{\mu\nu} \to \eta_{\mu\nu} + \lambda\delta_{\mu\nu}$, with λ weakly lifting the zero modes of η. Our results are not sensitive to choosing either 10^{-6} or 10^{-4} for λ strength.

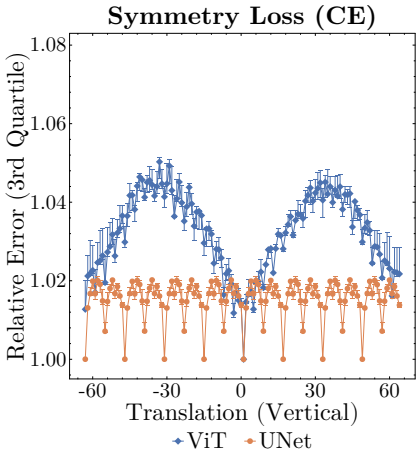
Appendix E. Limitations

Certain physically relevant symmetries remain outside the scope of the present study. Galilean boosts, in particular, impose a highly restrictive constraint: exact equivariance

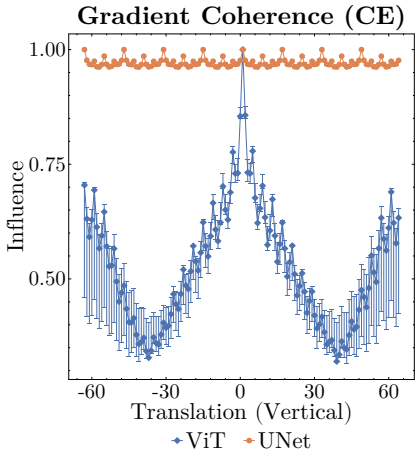
can be achieved only by affine transformations, while generic nonlinear architectures can at best satisfy this symmetry approximately (Wang et al., 2022). It would be interesting to extend our influence-based analysis to probe the response under small Galilean boosts. We did not examine the scaling symmetry of the continuum Navier-Stokes equations. In practical settings this symmetry is generically broken by spatial discretization, numerical regularization, and coarse graining of the data, rendering its faithful assessment ambiguous within the present framework. Finally, our analysis is restricted to two widely used architectural backbones, namely UNets and Vision Transformers. While these models represent dominant paradigms in contemporary large-scale PDE emulation, they do not exhaust the space of operator-learning methods. Consequently, the degree to which our measured gradient coherence properties depend on the chosen backbone rather than on the autoregressive training paradigm itself remains only partially resolved.

Finally, the reported influence profiles aggregate over multiple trajectory times and initial-condition classes, so their variability reflects not only estimator noise but also the intrinsic heterogeneity of the operator-learning problem: different flow regimes present distinct local loss geometries, and symmetry-compatible gradient transport need not be equally expressed across shocks, smooth advective structures, and vorticity-dominated states.

Appendix F. Supplementary Figures

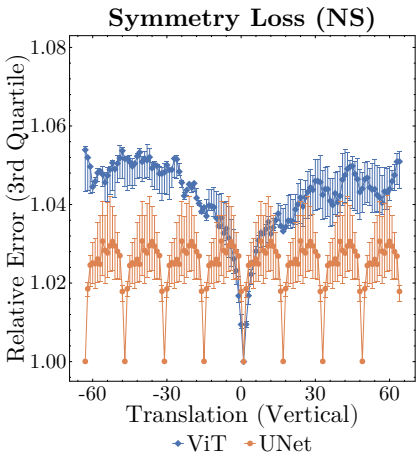


(*a*) Relative equivariance error as a function of vertical translation.

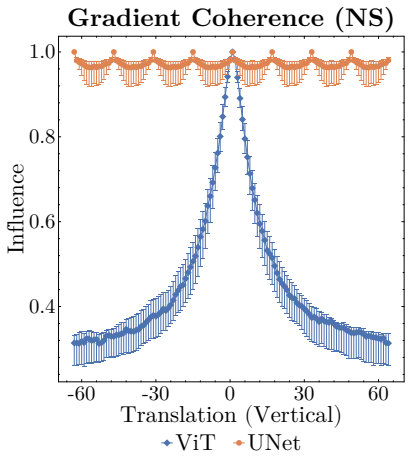


(*b*) Gradient alignment between an example and its vertically translated state.

Figure 3: Diagnostics under vertical translations on CE data. Markers and rangebars indicate the median and interquartile range over seeds. Nonuniform dependence on translation distance reflects vertical symmetry breaking and complements the horizontal response in Figure 1.

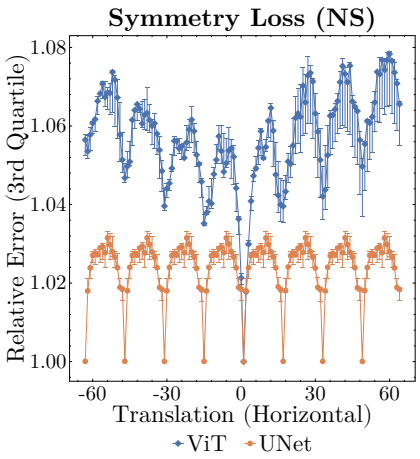


(*a*) Relative equivariance error as a function of vertical translation.

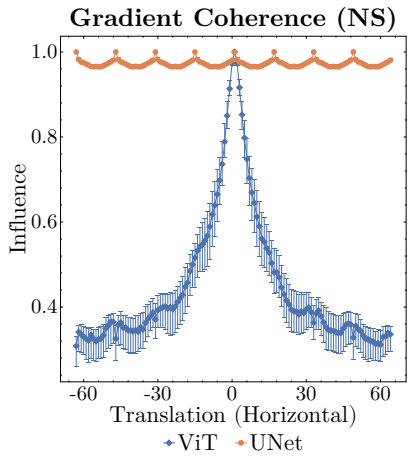


(*b*) Gradient alignment between an example and its vertically translated state.

Figure 4: Diagnostics under vertical translations on Navier-Stokes (NS) data. Markers and rangebars indicate the median and interquartile range over seeds. Nonuniform dependence on translation distance reflects vertical symmetry breaking; for horizontal translations, see Figure 5.

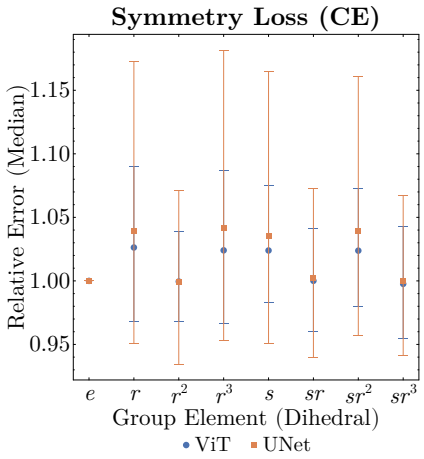


(*a*) Relative equivariance error as a function of horizontal translation.

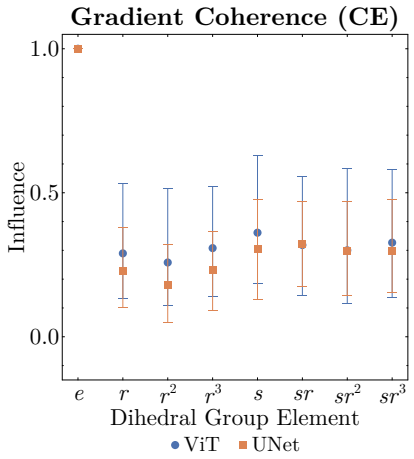


(*b*) Gradient alignment between an example and its horizontally translated state.

Figure 5: Diagnostics under horizontal translations on Navier-Stokes (NS) data. Markers and rangebars indicate the median and interquartile range over seeds. Nonuniform dependence on translation distance reflects horizontal symmetry breaking; for vertical translations, see Figure 4.

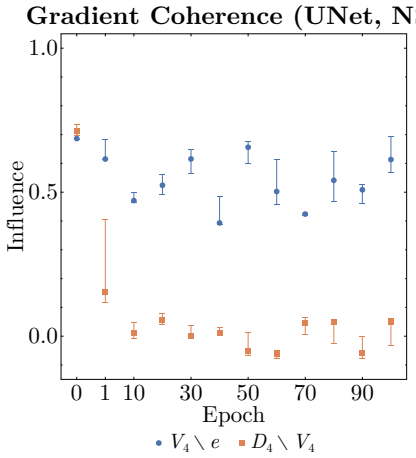


(a) Relative equivariance error for each D_4 group action.

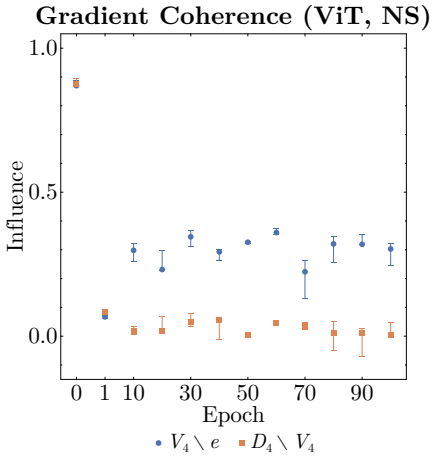


(b) Gradient alignment between an example and its D_4-transformed state.

Figure 6: Diagnostics under D_4 symmetry on CE data. Markers and rangebars indicate the median and interquartile range over seeds. This error-wise uniformity is consistent with the broadly uniform cross-influence profile, indicating that gradient updates couple across the sampled D_4 actions. The CE case therefore contrasts with the NS diagnostics in Figure 2, where equivariance error and influence vary strongly across group elements. Since our models are evaluated in the late-stage learning regime [see Figure 8], the observed off-identity gradient alignment supports continued symmetry-coherent gradient transfer across the D_4 orbit. Pearson correlations, computed over all measurements across batches, time steps, initial-condition classes, and D_4 elements, except the identity, between relative equivariance error and cross-influence normalized by self-influence on the identity element are -0.17 ± 0.01 for both UNet and ViT, where $\pm$ denotes the standard deviation about the mean over model seeds.

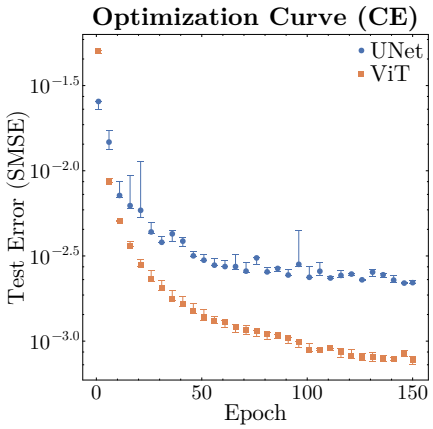


(*a*) Cross-influence into the learned and unlearned D_4 transformations across UNet training.

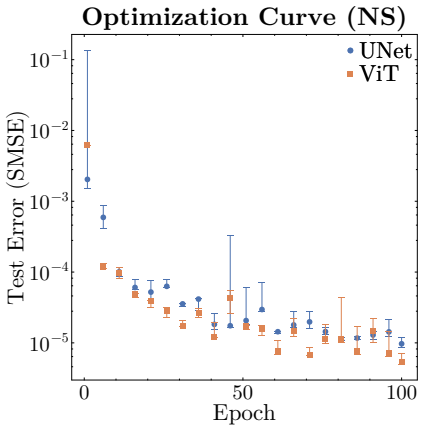


(*b*) Cross-influence into the learned and unlearned D_4 transformations across ViT training.

Figure 7: Evolution of cross-influence within the residual Klein four-subgroup $V_4 = \{e, r^2, sr, sr^3\}$ and its complement in D_4 for Navier–Stokes models. Influence is aggregated over the nonidentity learned transformations $V_4 \setminus e$ and the unlearned transformations $D_4 \setminus V_4$. The two classes exhibit comparable cross-influence at initialization, but separate rapidly during training: transformations in $D_4 \setminus V_4$ receive systematically less cross-influence than those in $V_4 \setminus e$ across subsequent epochs. This separation persists for both architectures. Markers and rangebars indicate the median and interquartile range over seeds.



(*a*) CE optimization trajectories.



(*b*) NS optimization trajectories.

Figure 8: Optimization trajectories for the CE and NS tasks. Test scaled mean squared error (SMSE) is shown as a function of training epoch for UNets and ViTs. Markers denote median performance across seeds, and rangebars indicate variability across seeds at each epoch. The curves establish that the subsequent symmetry diagnostics are evaluated after both model classes have reached the late-stage learning regime.