

Generalization vs. Memorization in Autoregressive Deep Learning: Or, Examining Temporal Decay of Gradient Coherence

Anonymous Authors¹

Abstract

Foundation models trained as autoregressive PDE emulators hold significant promise for accelerating scientific discovery through their capacity to both extrapolate beyond training regimes and efficiently adapt to downstream tasks despite a paucity of examples for fine-tuning. However, reliably achieving genuine generalization—a necessary capability for producing novel scientific insights and robustly performing during deployment—remains a critical challenge. Establishing whether these requirements are met demands evaluation metrics capable of clearly distinguishing genuine generalization from mere memorization. We apply the influence function formalism to systematically characterize how autoregressive PDE emulators assimilate and propagate information derived from diverse physical scenarios, revealing fundamental limitations of standard models and training routines in addition to providing actionable insights regarding the design of improved surrogates.

1. Introduction

Machine learning surrogate models has emerged as a powerful technique for efficiently approximating the solutions of computationally intensive partial differential equations (PDEs). These emulation methods range from purely data-driven approaches, trained on high-fidelity simulation data, to physics-informed neural networks, which integrate PDE structures directly into the training loss, enforcing physical laws as soft constraints (Raissi et al., 2019). Such models hold promise for achieving significant computational acceleration in applications such as fluid dynamics (Takamoto et al., 2024; Lippe et al., 2023; Gupta & Brandstetter, 2022; Ohana et al., 2025; Herde et al., 2024), climate modeling

(Bodnar et al., 2024), and materials science (Batatia et al., 2024), enabling rapid research and development across diverse scientific and engineering disciplines. Despite these advances, reliable generalization and robustness remains a critical challenge (Krishnapriyan et al., 2021). Before surrogate models can be safely deployed in operational environments demanding generalization beyond their training data, it is essential to develop methods capable of quantifying risk profiles and ensuring trustworthiness of predictions.

Distinguishing between memorization of training examples and genuine generalization is critical to evaluating model robustness; diagnostic tools such as influence functions (Koh & Liang, 2020; Bae et al., 2022), leverage scores, and gradient alignment analyses offer promising avenues for characterizing this balance, revealing whether models rely appropriately on generalized understanding or disproportionately on memorized patterns (Fort et al., 2020; Chatterjee, 2020; Chatterjee & Zielinski, 2022; Zielinski et al., 2020).

Autoregressive models, useful for their promising extrapolation capabilities, accumulate errors during inference, in part due to the inevitable distribution shift that originates with the usage of model outputs as inputs to drive the predicted evolution arbitrarily far into the future (Lee, 2023; Brandstetter et al., 2023). While such inference-time error accumulation is significant, we reveal a decisive learning limitation that also contributes to the difficulty of achieving stable long-term rollouts: gradient signals fail to propagate coherently across time, which implies that ordinary training lacks a mechanism for generalizing from supervision of one-step predictions to multi-frame dynamical evolution governed by a shared update structure. The length of time that a model performs (and is confident) prior to excessive prediction error defines a “trust horizon” for forward prediction that is contingent on its encoding of the true data-generating mechanisms—the physics—rather than merely exploiting empirical correlations among proximal points in feature space to perform local statistical interpolation.

Traditional benchmarks, such as point-wise mean squared error evaluations on limited validation datasets, often fail to adequately capture surrogate model reliability, especially when faced with variations in initial or boundary conditions, mesh resolutions, or varying physical parameter regimes

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

(Setinek et al., 2025). Physics-informed metrics, including conservation-law violation assessments, PDE residual norms, analytical-limit checks, and numerical stability evaluations, have been proposed to better reflect model robustness (Karniadakis et al., 2021), yet even these enriched criteria are not guaranteed to fully quantify the true worst-case prediction errors. Indeed, empirical accuracy metrics based on finitely many examples can dramatically underestimate the true worst-case error, especially when the data is noisy, sparse, or incompletely understood (Vapnik, 1998).

In scientific machine learning, limited availability of high-fidelity simulation data often results in narrow training distributions, making it challenging to develop robust emulators. On queries poorly represented by the training set, data-driven predictive models risk producing non-physical artifacts, such as violations of conservation laws, causality, or symmetry. While transfer learning and multi-fidelity methods have emerged to alleviate data scarcity, ensuring physically consistent generalization remains a significant challenge (Herde et al., 2024). Towards addressing this gap, current research increasingly emphasizes the development of PDE foundation models designed to achieve robust and unified generalization across diverse physical scenarios (Sun et al., 2025; Ye et al., 2024; Herde et al., 2024; Subramanian et al., 2023). Contemporary PDE emulators employ a variety of architectures (Li et al., 2021; Lu et al., 2021; Gregory et al., 2024; Shankar et al., 2023); however, most large-scale deployments rely on UNet (Ronneberger et al., 2015) or Transformer backbones (Vaswani et al., 2023; Liu et al., 2021; Dosovitskiy et al., 2021), and there remains no consensus on which model variant is most capable at scale. One must balance ease of optimization with the incorporation of physics priors, but quantitative tools for comparing loss-landscape properties across these architectures remain under-explored.

Insight into surrogate model behavior beyond static accuracy metrics can be gained through analysis of the model gradients. Combining test example error evaluation with gradient examination allows for interpolation of prediction errors across the underlying data manifold; for instance, PINNs can be certified with continuous-domain error bounds (Eiras et al., 2024). By quantifying gradient overlap among different training examples, it is possible to identify potential conflicts or synergies present during learning and inherent to fully trained models. Precisely how gradients derived from individual training samples propagate through model parameters is formalized through the use of influence functions (Hampel, 1974; Cook & Weisberg, 1982). Influence functions were originally developed in robust statistics (Huber & Ronchetti, 2009) to quantify how small perturbations of a data point in the training set affect model parameter estimations (Koh & Liang, 2020; Bae et al., 2022). Diagonal elements of the influence function measure each train-

ing example’s self-leverage; high-leverage points thereby identifying data that exerts disproportionate impact during training.

For PDE surrogates, the influence framework can also pinpoint examples providing gradient signals that exacerbate violations of physical constraints (Naujoks et al., 2024). Furthermore, influence functions reveal spatio-temporal correlations inherent in PDE emulator learning (Wang et al., 2025), distinguishing between memorization and generalization in cases where the underlying solution operator lacks explicit space-time dependence, in addition to exposing gradient misalignments across distinct initial conditions and inputs that are well separated in feature space. When applied to PDE foundation models, these techniques systematically characterize model stability, generalization capability, and uncertainty under structured domain shifts and multi-physics scenarios, in addition to uncovering subtle failure modes typically missed by conventional evaluation metrics, thereby enabling targeted refinements of model, architecture, and training routines that yield more robust, physically-consistent, data-driven models (Ren et al., 2019; Zhang & Pfister, 2021).

2. Related Work

Influence functions are powerful tools for understanding model behavior and data importance (Koh & Liang, 2020; Bae et al., 2022). Robust and interpretable criteria for detecting anomalous inputs follow from techniques that analyze the alignment of gradients (Wang et al., 2025) by quantifying directional consistency with in-distribution data (Huang et al., 2021), employ orthogonal projection (Behpour et al., 2023) to isolate anomalous components, and outlier gradient analysis (Chhabra et al., 2025).

Fort et al. (Fort et al., 2020) define stiffness in terms of the dot-product between the loss-gradients of two inputs. A positive stiffness then means that a stochastic gradient descent (SGD) step benefiting one example simultaneously lowers the loss of the other, evidence that the network assimilated shared, transferable features. Two summary statistics: sign-stiffness and cosine-stiffness, emphasize inter-class and intra-class correlations, respectively. Plotting stiffness against input-space distance yields a dynamic correlation length—the distance where average stiffness first crosses zero—which shrinks over epochs, revealing how the learned function becomes progressively more localized as specialization sets in.

The Coherent Gradients Hypothesis (Chatterjee, 2020) proposed that per-example gradients tend to align for similar inputs, so SGD steps amplify directions supported by many examples while suppressing idiosyncratic ones, steering the network toward functions that generalize rather than

memorize. Extensions of the Coherent Gradients Hypothesis (Zielinski et al., 2020) posit that SGD updates aligned across multiple training examples (strong directions) underpin generalization, whereas idiosyncratic updates (weak directions) promote memorization. They introduce optimizers that suppress weak directions without computing per-example gradients, dramatically reducing the train-test gap—even in the presence of heavy label noise—and thereby offer the first large-scale confirmation of the hypothesis. Complementing this view, He and Su (He & Su, 2020) establish the notion of local elasticity: in some neural networks, a parameter update perturbs predictions only within a narrow neighbourhood around the training point.

PINNfluence (Młodożeniec et al., 2025) interrogates a trained physics-informed neural network under perturbations to the PDE parameters and reweighting of collocation points. They distill raw pointwise influences into physically meaningful diagnostics such as the directional indicator, which measures the fraction of influence that propagates downstream with the fluid flow.

3. Our Contributions

We make three key advances toward principled analysis and validation of PDE emulators:

1. **Time-Aware Analysis of Off-Diagonal Influence Function Elements:** A systematic study off-diagonal influence function elements for PDE surrogate models, capable of quantifying training-sample leverage across physical time [see Figure 1]. This diagnostic sets standards for identifying the learning of persistent nontrivial correlations that extend across temporal horizons, thereby identifying when the emulator network has internalized fundamental, time-invariant PDE structures.
2. **Gradient-Coherence Diagnostics Across Initial Condition Classes:** We determine the degree of alignment of gradients computed across different classes of PDE solutions for two standard architectures, a UNet and a ViT. Strong alignment signals the learning of robust, transferable physics, whereas weak alignment suggests that the neural network embeds these classes on separated regions of the input manifold, with limited gradient coherence, despite the fact that the data represents solutions to the same underlying PDE.
3. **Dynamic Correlation Length and Curvature Diagnostics:** We show that autoregressive PDE emulators generically exhibit a limited dynamic correlation length (Fort et al., 2020), directly observable through the rapid decay of influence with increasing feature-space distance. Such feature-space localization provides an ex-

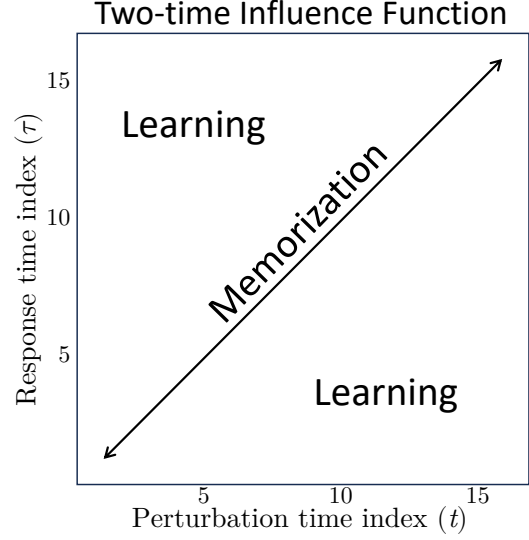


Figure 1. Conceptual schematic separating learning from memorization using the two-time influence diagnostic. A learning regime corresponds to broad support away from the diagonal, meaning that training information at one time affects predictions at many other times. A memorization regime corresponds to diagonal dominance, meaning that updates are effectively time-indexed and do not encode a time-consistent solution operator.

plicit, training-time explanation for why such models fail to reuse dynamical structure or internalize shared physical laws beyond narrow neighborhoods of the data manifold. Complementing this evidence, spectral analyses of the neural tangent kernel metric show that low test error is typically achieved in a highly anisotropic regime: while most directions remain flat, a small number of dominant eigenmodes exhibit large curvature, corresponding to sharp, high-sensitivity directions rather than globally robust solutions. This spectral imbalance clarifies why apparent interpolation success does not imply robustness, and why learned dynamics fail to transfer coherently across time or conditions despite favorable one-step performance (Karakida et al., 2019; Anonymous, 2025).

This paper is organized as follows. For the readers’ convenience, we first present our central results section 4, exposing pronounced lack of generalization capabilities in autoregressive PDE emulators. Technical details—those covering both the mathematical formulation of the influence-function framework in addition to our training procedures—are provided in section 5 and section 6, respectively.

4. Results

We examine how training information propagates across time and initial-condition classes in autoregressive PDE em-

ulators, using influence-based diagnostics evaluated on held-out test data. Across architectures, physical observables, and datasets, we find that gradient responses are strongly localized in both time and class, with off-diagonal influence rapidly decaying, indicating that these models primarily learn time- and class-indexed update rules rather than a globally consistent dynamical operator.

Test-data measurements of the two-time influence function [see Equation 6] for both a UNet and a ViT tasked with emulating fluid flow exhibit rapid temporal decay in the off-diagonal terms [see Figure 2], which indicates that surrogate training constructs localized vector fields suitable only for interpolation within small neighborhoods of the training data sub-manifold, rather than the universally consistent function that is desired based on expectations stemming from our knowledge of the underlying governing equations. If such models were truly learning the solution operator to a PDE that lacks explicit time dependence, gradients derived from examples at a given time would necessarily have a profound effect on the predictions at any other time, for we know that the true solution operator must take the same functional form at every point in phase space. This superfluous time awareness presents across the entire training trajectory, demonstrating that our models did not learn the underlying solution operator. Consistent with the two-time

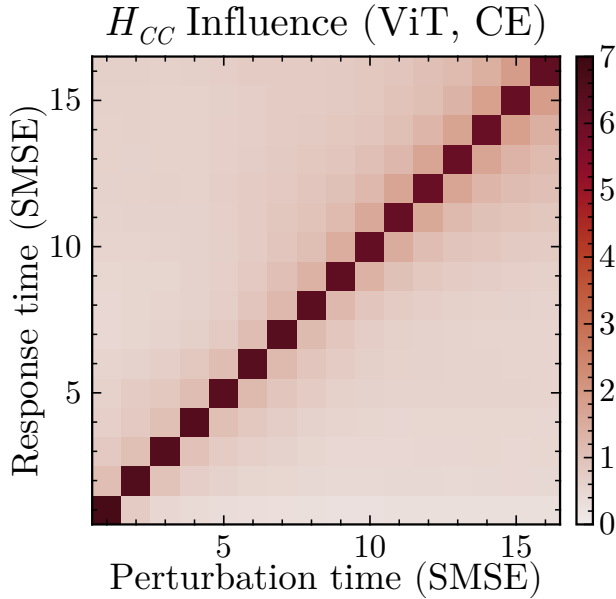


Figure 2. Heatmap of two-time influence for our ViTs trained on CE data, shown as a function of perturbation time (horizontal axis) and response time (vertical axis). Each pixel reports the intra-class averaged response induced by test example gradients at the perturbation time. A narrow diagonal ridge corresponds to time-local sensitivity consistent with interpolation, rather than generalization; substantial off-diagonal structure would indicate time-transferable learning. For the analogous plot using our UNets, see Figure 17. For NS data counterparts, see Figure 18 and Figure 19.

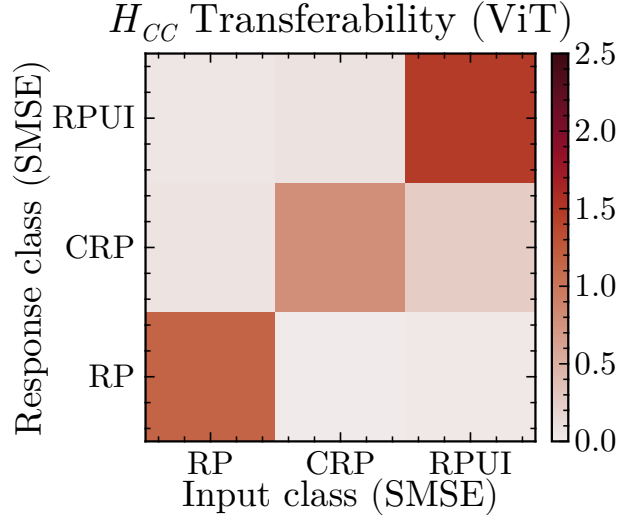


Figure 3. Class-to-class transferability matrix for our ViTs trained on the three-class CE split labeled RP, CRP, and RPUI. Each entry reports the time-averaged influence of test examples from the input class (horizontal axis) on examples from the response class (vertical axis). Diagonal dominance indicates class-locked gradient geometry; substantial off-diagonal values would imply reuse of dynamical features across classes. For the analogous plot using our UNets, see Figure 13. For NS data counterparts, see Figure 14 and Figure 15.

influence maps, the class-to-class transferability matrix in Figure 3 is strongly diagonal, indicating that gradient geometry is effectively class-locked: updates supported by one initial-condition family produce negligible response in the others. Furthermore, there is a near-total absence of inter-class influence [see Figure 4]. The degree of gradient alignment across examples also affords conclusions about the data manifold sparsity: while all inputs to the network are intimately related as unique solutions to a shared equation of motion under different initial conditions, both our ViTs and our UNets render inputs well separated in the sense that their gradients don’t meaningfully overlap unless their feature space distance small [see Figure 5], which implies limited generalization over dynamical structure away from nearby states. Such results challenge a fundamental assumption motivating the development of PDE foundation models, as it demonstrates that these models are prone to effectively treating different flow fields as distinct, isolated learning tasks. That this happens even when said classes of solutions arise merely from different initial conditions to the same physical process underscores the need for inductive biases to be explicitly incorporated during model development.

In addition to the overlap of cost function gradients, we also considered gradients derived from physics informed loss functions, such as global mass conservation [see Figure 6]

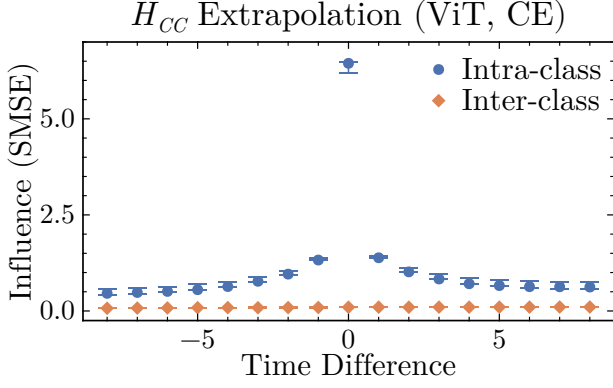


Figure 4. Time-lag summary of temporal transferability for our ViTs on CE data. The influence is averaged over all time-pairs of the same time difference and then split into intra-class pairs (gradient and response drawn from the same initial-condition class) and inter-class pairs (distinct initial-condition classes). Strong concentration near zero time difference indicates that gradient information fails to propagate coherently across time, while the lack of inter-class influence reveals an absence of physics consistent generalization. For the analogous plot using our UNets, see Figure 21. For NS data counterparts, see Figure 22 and Figure 23.

and global energy conservation. In all cases, we observed that the response function decayed off the time-diagonal and was dominated by intra-class matrix elements [see Figure 7]. Hence, we conclude that predictive models lacking explicit inductive biases are not internalizing a unified governing law, but merely allocating parameters tasked specifically with evolving states associated with a given time along a given class of trajectories.

Lastly, Figure 8 shows that the dominant NTK eigenmodes are large, revealing a stiff, highly anisotropic local response geometry; in particular, low test error coexists with sharp high-curvature modes rather than a uniformly flat, robust geometry.

5. Proximal Response Function

We develop the influence function as follows, taking inspiration from (Bae et al., 2022). Let θ be the current parameter values and consider the optimization step $\theta \leftarrow \theta + \delta\theta$, where the tangent-space displacement $\delta\theta$ minimizes the proximal objective

$$S(\delta\theta) = dC[\delta\theta] + \frac{1}{2} \|\hat{y}(\theta) - \hat{y}(\theta + \delta\theta)\|_{L_2}^2, \quad (1)$$

with \hat{y} a neural network. The stationarity condition $dS \stackrel{!}{=} 0$ is satisfied (to linear order) by

$$\delta\theta^\mu = -\eta^{\mu\nu} \partial_\nu C, \quad (2)$$

where $\eta_{\mu\nu} = J_\mu^n J_\nu^n$ is the neural tangent kernel metric, $J_\mu^n = \partial_\mu \hat{y}^n$ is the model Jacobian, and n indexes a given

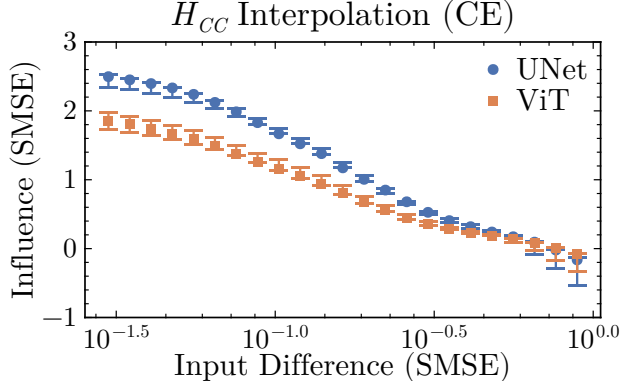


Figure 5. Curve fit of the influence as a function of feature-space separation between input states for CE data, comparing UNet and ViT; rangebars show uncertainty across seeds. A steep decay indicates short-range locality on the learned data manifold, implying that parameter updates affect only nearby states and generalization is limited. For NS data counterpart, see Figure 33.

mini-batch example. By convention, the components of η carry lowered indices, $\eta_{\mu\nu}$, while those of η^{-1} carry raised indices, $\eta^{\mu\nu}$, i.e. $\eta^{\mu\alpha} \eta_{\alpha\nu} = \delta^\mu_\nu$; η provides the canonical correspondence between covariant and contravariant components (Absil et al., 2008). Equation 1 balances the force term dC against the kinetic cost of the update distance in the η prescribed geometry. Convexity of η , together with mild regularity requirements on C , guarantees a unique stationary point of each proximal subproblem. Proximal gradient descent iterates these subproblems to accumulate a sequence of locally improving displacements that drives descent of cost function C . The inverse susceptibility tensor η^{-1} serves as a generalized stiffness operator by propagating gradient signals to parameter displacements (Fort et al., 2020).

Classical influence functions can be expressed as the Lie derivative of a scalar; they’re capable of probing local gradient coherence, generalization capabilities, and adversarial sensitivity, in addition to enabling the identification of high-leverage examples. Consider a scalar observable Q and a vector field $V = -\eta^{-1}(dC)$ derived from the proximal objective. The Lie derivative of Q along V is

$$\begin{aligned} \mathcal{L}_V Q &= (\partial_\mu Q) \eta^{\mu\nu} (-\partial_\nu C) \\ &= - \left(\frac{\delta Q}{\delta \hat{y}^n}, \Pi^{nm} \frac{\delta C}{\delta \hat{y}^m} \right), \end{aligned} \quad (3)$$

where

$$\Pi^{nm} = J_\mu^n \eta^{\mu\nu} J_\nu^m, \quad (4)$$

and the inner product (\cdot, \cdot) is performed over feature indices. Hence, when Q is a loss function, Equation 3 reduces to the familiar form of an influence function in deep learning: the infinitesimal response of the loss, expressible as a metric-weighted gradient overlap. Likewise, when Q denotes a

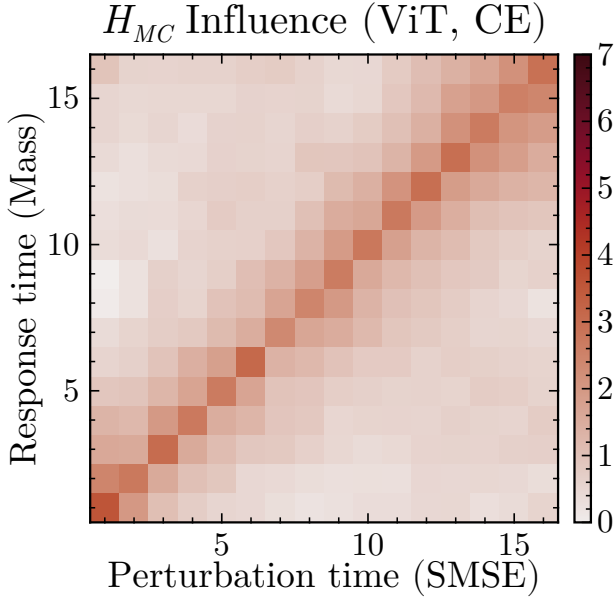


Figure 6. Two-time influence map for our ViTs on CE data when the response observable is the global mass-consistency signal. The horizontal axis indexes the time at which a test perturbation is applied, and the vertical axis indexes the time at which the mass-based response is evaluated. Off-diagonal support indicates that intra-class mass-related gradient information couples distant times. For the analogous plot using our UNets, see Figure 25. For plots concerning energy conservation, see Figure 28 and Figure 29.

model response and V encodes the perturbation to the gradient signal induced by a deformation of the input, Equation 3 reproduces the classical influence-function expression from robust statistics [see Appendix A].

Evidently, the Lie-derivative formulation of response is well defined at any point along the training trajectory, as it depends only on the instantaneous training-flow vector field and the induced local geometry; hence, influence-function analysis of neural networks does not require attainment of a stationary point to expand about. This perspective elevates influence from a static sensitivity relevant only near convergence to a dynamical linear response observable defined throughout optimization.

In the limit of vanishing regularization, $\lambda \rightarrow 0$, Π becomes idempotent, assuming the form of a classical hat matrix. We thus take the view that diagonal elements reflect statistical leverage, quantifying the self-influence of individual training examples, while off-diagonals measure cross-influence, i.e. influence between distinct examples. High leverage scores identify regions of parameter space with strong local curvature or limited redundancy, i.e., points with disproportionately large influence on the global response structure. Furthermore, the response matrix encodes the pairwise overlap of example gradients-effectively probing the local loss landscape by highlighting directions of correlated curva-

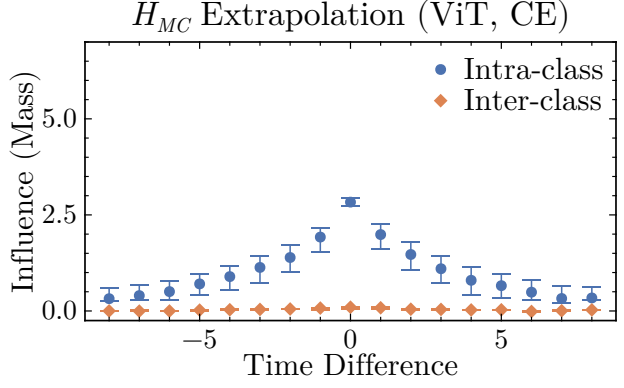


Figure 7. Time-lag transferability curve for our ViTs on CE data using the global mass-consistency observable. Intra-class averages measure how mass-based influence response align within an initial-condition class, while inter-class averages measure cross-class reuse of mass-related gradient directions. The decay pattern diagnoses whether conservation constraints induce transferable structure or remain class-locked. For the analogous plot using our UNets, see Figure 27. For plots concerning energy conservation, see Figure 30 and Figure 31.

ture and shared descent paths. Physical considerations that guide expectations for the structure of Π are evident on recognizing that we have so far suppressed feature indices in the model Jacobian. The response matrix Π^{nm} tracks how gradients derived from each output feature of prediction m influence each output feature of prediction n , offering an investigative level of detail across spacetime, channel, and class dimensions that remains unexplored. We emphasize that the proximal penalty in Equation 1 sets the geometry of the update and clearly identifies Π as the primary object governing to what extent an infinitesimal perturbation in the cost function propagates to an observable, such as the test error or physical consistency of predictions. To avoid materializing Π , which has $(128 \times 128 \times 4 \times 48)^2$ elements, we consider macroscopic observables: SMSE, in addition to global mass and energy conservation.

5.1. Observables

Our probe of generalization capabilities proceeds by quantifying the coherence of gradients derived from test data cost functions of physical and statistical significance. We introduce three generalized residuals

$$r_C = \frac{\delta C_{\text{SMSE}}}{\delta \hat{y}_\theta} = \frac{1}{4} \sum_c \frac{\hat{y}_\theta^c - y^c}{\text{RMS}(y^c)}, \quad (5a)$$

$$r_M = \frac{\delta C_{\text{Mass}}}{\delta \hat{y}_\theta} = \frac{M[\hat{y}_\theta] - M[x]}{M[x]}, \quad (5b)$$

$$r_E = \frac{\delta C_{\text{Energy}}}{\delta \hat{y}_\theta} = \frac{E[\hat{y}_\theta] - E[x]}{E[x]}, \quad (5c)$$

where M (E) computes the total mass (energy) of its argument, x is the input state that evolves to y , i.e. $y = U[x]$,

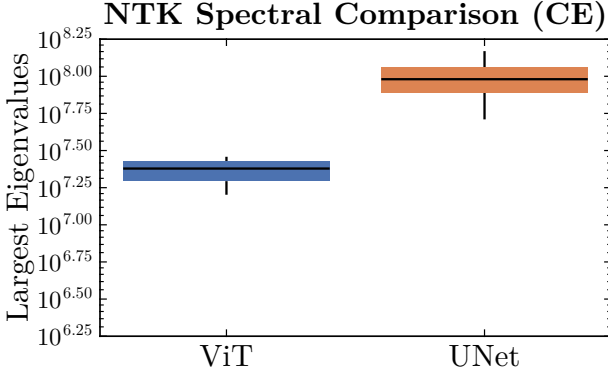


Figure 8. Spectral comparison of dominant neural tangent kernel eigenvalues for our UNets and our ViTs on CE data. The plotted distributions summarize the leading eigenvalue statistics across seeds and batches of the trained models. Larger dominant eigenvalues indicate a locally stiffer, sharper response geometry. For NS data counterpart, see Figure 35.

where U is defined in Equation 7, and \hat{y} is the neural network approximation to y . Recall that each training example is comprised of pairs $(x, y) = (s_t^n, s_{t+1}^n)$ of states s sharing a common initial configuration indexed by n , and related by the compressible Euler evolution operator.

Viewed as a coupling matrix over residuals, the diagonal blocks of Π recover the usual influence (e.g., how a perturbation in the SMSE affects SMSE itself), while the off-diagonal blocks encode cross-coupling, quantifying how a change in the SMSE residual at one time step or sample is converted into the conservation residual at another time step or sample, and vice versa.

It is useful to introduce following notation for the remaining external indices of the response matrix

$$H_{AB}(t, n | \tau, m) = (r_A^{nt}, \Pi_{t\tau}^{nm} r_B^{m\tau}), \quad (6)$$

where n, m index trajectories, defined by distinct initial conditions; indices t and τ specify the time step along said trajectories. H_{CC} gives the change in SMSE due to an SMSE perturbation, while H_{MC} and H_{EC} propagate the effect of gradients derived from SMSE into the physics informed and mass and energy conservation errors, respectively.

We report influence in a standardized form by normalizing with respect to the empirical variance of perturbations within each mini-batch, so that the baseline model-corresponding to unstructured stochastic variability-naturally sets the reference scale to unity. In this normalization, departures from one directly indicate influence beyond what is expected from random mini-batch fluctuations, providing a principled scale for interpreting both amplified self-responses and suppressed cross-responses (Héritier & Ronchetti, 1994; Lu et al., 1997). Matrix elements of Π were determined for six different mini-batches, each of which contained three

trajectories corresponding to distinct initial conditions, for each seed of each model architecture trained, across two datasets[see Appendix B].

6. Data and Training

We trained neural network surrogate models to approximate the evolution of two-dimensional compressible Euler flows, provided by the PDEgym dataset (Herde et al., 2024). Specifically, we used a dataset that contains three classes of initial conditions, namely, the four quadrant Riemann problem with (CE-RPUI) and without (CE-RP) uncertain interfaces, in addition to the curved Riemann problem (CE-CRP). This 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 three classes of initial conditions; for each trajectory, we used the first 16 time steps, for total of approximately 110,000 training pairs requiring greater than 150 GB memory.

Since instantaneous flow states alone cannot distinguish viscous Navier-Stokes flows from their inviscid Euler counterparts, we do not combine compressible Euler data with Navier-Stokes data. Furthermore, rather than representing a fluid state using the velocities and pressure in addition to density, as was done by Poseidon (Herde et al., 2024), we used the momentum and energy fields; we expect that this setup will better facilitate the learning of all four conservation laws.

Each snapshot of the flow state at discrete time t is represented as a set of spatially discretized fields $\rho_{\text{mass}}, \rho_{\text{mom}}^i, \rho_{\text{energy}}$ on a uniform grid of size 128×128 , where ρ_{mass} denotes mass density, ρ_{mom}^i are the Cartesian components of momentum density, and ρ_{energy} is energy density. The model \hat{y}_θ is trained to emulate the compressible Euler evolution, i.e., $\hat{y}_\theta \approx U$, where the operator U enacts

$$s_{t+1} = U[s_t], \quad (7)$$

with s_t the collection of state variables at timestep t , via optimization of the weights θ . Specifically, we used the Adam optimizer (Kingma & Ba, 2017) with learning rate 5×10^{-4} and weight decay $\lambda = 10^{-4}$ to minimize a scaled mean squared error (SMSE) between predicted and true

states

$$C_{\text{SMSE}}(\theta) = \frac{1}{N} \sum_{n=1}^N \frac{1}{4} \sum_c \frac{\|\hat{y}_\theta^c(s_{t_n}) - s_{t_n+1}^c\|_{L_2}^2}{\text{RMS}(s_{t_n+1}^c)}, \quad (8)$$

on mini-batches containing $N = 48$ transitions $s_t \rightarrow s_{t+1}$, chosen randomly from the training set; here, the L_2 norm is computed over the spatial degrees of freedom and the channel index c runs through mass, both cartesian components of the momentum, and energy, respectively. $\text{RMS}(s^c)$ is computed channel-wise by taking the spatial root-mean-square of s^c . Thus, Equation 8 strikes a balance between ordinary and relative mean-squared-error; normalizing each channel’s squared error by the target fields’ characteristic amplitude favors examples containing pronounced features, but does not completely drown out gradients derived from relatively quiescent flows, thereby facilitating accurate capture of high-energy shocks and wavefronts without harsh under-emphasis of small-amplitude features. Moreover, the scaling of Equation 8 renders dimensionless the matrix elements of interest to this work.

In order to compare the results of our experiments across model architectures, we trained both a UNet (Ronneberger et al., 2015) and a vision transformer (ViT) (Vaswani et al., 2023; Liu et al., 2021; Dosovitskiy et al., 2021). Our UNet was based on BigGAN (Brock et al., 2019), with four down-sampling blocks and 24 channels after the initial embedding layer, for a total of about 13-million parameters. Our vision transformer was of layer-depth six, with 256 channels, for a total of about 5-million parameters, fewer than our UNet due to memory constraints on the 40 GB A100s that we used for training. To supplement our compressible Euler study, we repeat our analysis for velocity fields corresponding to solutions of the Navier-Stokes equations with NS-BB, NS-Gauss, and NS-Sines initial conditions, which presents a distinct feature space and flow morphology, involving smoother, viscosity-regularized transport with vorticity-dominated structure [see Figure 11].

The validation losses for each of our CE-tasks models is shown in Figure 9. Despite possessing fewer parameters, our ViT model consistently outperformed the UNet. Each of the two architectures was trained three times, sharing those three seeds that controlled initialization and dataset split. The training of each model was performed in distributed mode across two such A100s.

7. Conclusion

Our research reveals a critical shortcoming common to physics-agnostic PDE emulators: it is not an immediate consequence of large-scale multi-scenario training that the resulting trained model can satisfy those stringent expectations that follow from the governing equations one is trying to emulate. Physical expectations demand a shared feature

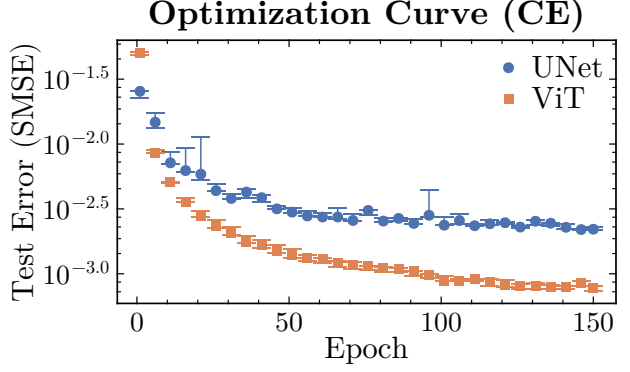


Figure 9. Optimization trajectories for the CE task, shown as test scaled mean squared error (SMSE) versus training epoch for our UNets and our ViTs. Markers denote the median model performance, while vertical range bars indicate the variance across model seeds at each epoch. Training is halted in the late-stage learning regime to test whether the induced local gradient geometry is sufficient to identify a physically consistent, generalizing solution. For the corresponding NS-data plot, see Figure 11.

basis across trajectories, yet our results reveal a failure of both UNets and ViTs to support a nontrivial off-diagonal response. This mismatch underscores the importance of enforcing principled physics-based constraints, either as weak regularizers during training, or baked in strictly through architectural design. By measuring gradient overlap between classes of initial conditions we reveal an absence of coherent gradients, which suggests limited learning of robust, transferable physics. This demonstrates that both ViT and UNet surrogates embed these solution classes on nearly disjoint manifolds, challenging the efficacy of current multi-scenario training pipelines.

We demonstrate that influence functions form a versatile diagnostic framework and demonstrate their effectiveness in revealing the degree of balance between memorization and generalization in autoregressive predictors. This analysis suggests that ordinary data-driven PDE emulators behave as statistical estimators, producing predictions primarily based on those training examples that lie within a neighborhood of the input query. While this localized learning mechanism provides resilience against noisy data, it also restricts generalization, and indicates that the learned data manifold geometry is composed of largely isolated regions.

In summary, we highlight a new, concrete, and targetable characteristics—time and class aware cross-influence—to guide researchers in designing algorithms capable of learning the underlying generative process and achieving reliable long-term rollouts.

8. Electronic Submission

Software and Data

We trained our models on the openly available dataset PDE-Gym (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 & Krumbiegel, 2021).

The code used in this work is publicly available at <https://anonymous.4open.science/r/PDEHats-B7BF/>. Additionally trained models and gradient data are available from the authors upon reasonable request.

Impact Statement

This work contributes a diagnostic framework for distinguishing mere memorization from genuine generalization in autoregressive surrogate models, with particular relevance to scientific and engineering applications where model failures can have downstream consequences. By exposing training-dynamics limitations that are not visible through standard accuracy metrics, our analysis helps identify when learned surrogates are likely to be brittle under long-horizon rollout or distribution shift, informing safer deployment in settings such as climate modeling, fluid dynamics, and materials simulation. The methods introduced here are diagnostic rather than prescriptive and do not directly enable new capabilities for misuse; instead, they promote transparency and reliability by clarifying when and why models fail to internalize shared physical structure. More broadly, our research encourages the development of learning algorithms and evaluation practices that prioritize robustness and interpretability, supporting the responsible use of machine learning in high-consequence scientific workflows.

References

- Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3.
- Anonymous. Measuring model robustness via fisher information: Spectral bounds, theoretical guarantees, and practical algorithms. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=BLhQv7iF3q>. under review.
- Bae, J., Ng, N., Lo, A., Ghassemi, M., and Grosse, R. If influence functions are the answer, then what is the question?, 2022. URL <https://arxiv.org/abs/2209.05364>.
- Batatia, I., Benner, P., Chiang, Y., Elena, A. M., Kovács, D. P., Riebesell, J., Advincula, X. R., Asta, M., Avaylon, M., Baldwin, W. J., Berger, F., Bernstein, N., Bhowmik, A., Blau, S. M., Cărare, V., Darby, J. P., De, S., Pia, F. D., Deringer, V. L., Elijošius, R., El-Machachi, Z., Falcioni, F., Fako, E., Ferrari, A. C., Genreith-Schriever, A., George, J., Goodall, R. E. A., Grey, C. P., Grigorev, P., Han, S., Handley, W., Heenen, H. H., Hermansson, K., Holm, C., Jaafar, J., Hofmann, S., Jakob, K. S., Jung, H., Kapil, V., Kaplan, A. D., Karimitari, N., Kermode, J. R., Kroupa, N., Kullgren, J., Kuner, M. C., Kuryla, D., Liepuoniute, G., Margraf, J. T., Magdău, I.-B., Michaelides, A., Moore, J. H., Naik, A. A., Niblett, S. P., Norwood, S. W., O’Neill, N., Ortner, C., Persson, K. A., Reuter, K., Rosen, A. S., Schaaf, L. L., Schran, C., Shi, B. X., Sivonxay, E., Stenczel, T. K., Svahn, V., Sutton, C., Swinburne, T. D., Tilly, J., van der Oord, C., Varga-Umbrich, E., Vegge, T., Vondrák, M., Wang, Y., Witt, W. C., Zills, F., and Csányi, G. A foundation model for atomistic materials chemistry, 2024. URL <https://arxiv.org/abs/2401.00096>.
- Behpour, S., Doan, T., Li, X., He, W., Gou, L., and Ren, L. Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients, 2023. URL <https://arxiv.org/abs/2308.00310>.
- 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. A foundation model for the earth system, 2024. URL <https://arxiv.org/abs/2405.13063>.
- Brandstetter, J., Worrall, D., and Welling, M. Message passing neural pde solvers, 2023. URL <https://arxiv.org/abs/2202.03376>.
- Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis, 2019. URL <https://arxiv.org/abs/1809.11096>.
- Chatterjee, S. Coherent gradients: An approach to understanding generalization in gradient descent-based optimization, 2020. URL <https://arxiv.org/abs/2002.10657>.
- Chatterjee, S. and Zielinski, P. On the generalization mystery in deep learning, 2022. URL <https://arxiv.org/abs/2203.10036>.
- Chhabra, A., Li, B., Chen, J., Mohapatra, P., and Liu, H. Outlier gradient analysis: Efficiently identifying detrimental training samples for deep learning models. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=v77ZMzbsBA>.

- Cook, R. D. and Weisberg, S. *Residuals and Influence in Regression*. Chapman and Hall, New York, 1982. ISBN 0-412-24280-0.
- Danisch, S. and Krumbiegel, J. 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>.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL <https://arxiv.org/abs/2010.11929>.
- Eiras, F., Bibi, A., Bunel, R., Dvijotham, K. D., Torr, P., and Kumar, M. P. Efficient error certification for physics-informed neural networks, 2024. URL <https://arxiv.org/abs/2305.10157>.
- Fort, S., Nowak, P. K., Jastrzebski, S., and Narayanan, S. Stiffness: A new perspective on generalization in neural networks, 2020. URL <https://arxiv.org/abs/1901.09491>.
- George, T. NNGeometry: Easy and Fast Fisher Information Matrices and Neural Tangent Kernels in PyTorch, February 2021. URL <https://doi.org/10.5281/zenodo.4532597>.
- Gregory, W. G., Hogg, D. W., Blum-Smith, B., Arias, M. T., Wong, K. W. K., and Villar, S. Equivariant geometric convolutions for emulation of dynamical systems, 2024. URL <https://arxiv.org/abs/2305.12585>.
- Gupta, J. K. and Brandstetter, J. Towards multi-spatiotemporal-scale generalized pde modeling, 2022. URL <https://arxiv.org/abs/2209.15616>.
- Hampel, F. R. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: 10.1080/01621459.1974.10482962.
- He, H. and Su, W. J. The local elasticity of neural networks, 2020. URL <https://arxiv.org/abs/1910.06943>.
- Herde, M., Raonić, B., Rohner, T., Käppeli, R., Molinaro, R., de Bézenac, E., and Mishra, S. Poseidon: Efficient foundation models for pdes, 2024. URL <https://arxiv.org/abs/2405.19101>.
- Héritier, S. and Ronchetti, E. 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>.
- Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild, 2021. URL <https://arxiv.org/abs/2110.00218>.
- Huber, P. J. and Ronchetti, E. M. *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>.
- Innes, M. Don’t unroll adjoint: Differentiating ssa-form programs. *CoRR*, abs/1810.07951, 2018. URL <http://arxiv.org/abs/1810.07951>.
- Karakida, R., Akaho, S., and ichi Amari, S. Universal statistics of fisher information in deep neural networks: Mean field approach, 2019. URL <https://arxiv.org/abs/1806.01316>.
- Karniadakis, G., Bilonis, I., and Perdikaris, P. Physics-informed machine learning. *Nature Reviews Physics*, 3: 422–440, 2021. doi: 10.1038/s42254-021-00314-5.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017. URL <https://arxiv.org/abs/1412.6980>.
- Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions, 2020. URL <https://arxiv.org/abs/1703.04730>.
- Krishnapriyan, A. S., Gholami, A., Zhe, S., Kirby, R. M., and Mahoney, M. W. Characterizing possible failure modes in physics-informed neural networks, 2021. URL <https://arxiv.org/abs/2109.01050>.
- Lee, Y. Autoregressive renaissance in neural pde solvers. In *ICLR Blogposts 2023*, 2023. URL <https://iclr-blogposts.github.io/2023/blog/2023/autoregressive-neural-pde-solver/>.
- Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations, 2021. URL <https://arxiv.org/abs/2010.08895>.
- Lippe, P., Veeling, B. S., Perdikaris, P., Turner, R. E., and Brandstetter, J. Pde-refiner: Achieving accurate long rollouts with neural pde solvers, 2023. URL <https://arxiv.org/abs/2308.05732>.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision

- transformer using shifted windows, 2021. URL <https://arxiv.org/abs/2103.14030>.
- Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.
- Lu, J., Ko, D., and Chang, T. 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, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G. E. Learning nonlinear operators via deepoNet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, March 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00302-5. URL <http://dx.doi.org/10.1038/s42256-021-00302-5>.
- Mlodozieniec, B., Eschenhagen, R., Bae, J., Immer, A., Krueger, D., and Turner, R. Influence functions for scalable data attribution in diffusion models, 2025. URL <https://arxiv.org/abs/2410.13850>.
- Montoisson, A. and Orban, D. Krylov.jl: A Julia basket of hand-picked Krylov methods. *Journal of Open Source Software*, 8(89):5187, 2023. doi: 10.21105/joss.05187.
- Naujoks, J. R., Krasowski, A., Weckbecker, M., Wiegand, T., Lapuschkin, S., Samek, W., and Klausen, R. P. Pinnfluence: Influence functions for physics-informed neural networks, 2024. URL <https://arxiv.org/abs/2409.08958>.
- 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. The well: a large-scale collection of diverse physics simulations for machine learning, 2025. URL <https://arxiv.org/abs/2412.00568>.
- Orban, D. and Arioli, M. *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>.
- Pal, A. On Efficient Training & Inference of Neural Differential Equations, 2023a.
- Pal, A. Lux: Explicit Parameterization of Deep Neural Networks in Julia, April 2023b. URL <https://doi.org/10.5281/zenodo.7808903>.
- Raissi, M., Perdikaris, P., and Karniadakis, G. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: <https://doi.org/10.1016/j.jcp.2018.10.045>.
- Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning, 2019. URL <https://arxiv.org/abs/1803.09050>.
- Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation, 2015. URL <https://arxiv.org/abs/1505.04597>.
- Setinek, P., Galletti, G., Gross, T., Schnürer, D., Brandstetter, J., and Zellinger, W. Simshift: A benchmark for adapting neural surrogates to distribution shifts, 2025. URL <https://arxiv.org/abs/2506.12007>.
- Shankar, V., Barwey, S., Kolter, Z., Maulik, R., and Viswanathan, V. Importance of equivariant and invariant symmetries for fluid flow modeling, 2023. URL <https://arxiv.org/abs/2307.05486>.
- Subramanian, S., Harrington, P., Keutzer, K., Bhimji, W., Morozov, D., Mahoney, M., and Gholami, A. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior, 2023. URL <https://arxiv.org/abs/2306.00258>.
- Sun, J., Liu, Y., Zhang, Z., and Schaeffer, H. Towards a foundation model for partial differential equations: Multi-operator learning and extrapolation, 2025. URL <https://arxiv.org/abs/2404.12355>.
- Takamoto, M., Praditia, T., Leiteritz, R., MacKinlay, D., Alesiani, F., Pflüger, D., and Niepert, M. Pdebench: An extensive benchmark for scientific machine learning, 2024. URL <https://arxiv.org/abs/2210.07182>.
- TransferLab. pyDVL, April 2024. URL <https://github.com/aai-institute/pyDVL>.
- Vapnik, V. N. *Statistical Learning Theory*. Wiley, New York, 1998. ISBN 978-0471030034.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023. URL <https://arxiv.org/abs/1706.03762>.

Wang, S., Bhartari, A. K., Li, B., and Perdikaris, P. Gradient alignment in physics-informed neural networks: A second-order optimization perspective, 2025. URL <https://arxiv.org/abs/2502.00604>.

Ye, Z., Huang, X., Chen, L., Liu, H., Wang, Z., and Dong, B. 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>.

Zhang, Z. and Pfister, T. Learning fast sample re-weighting without reward data, 2021. URL <https://arxiv.org/abs/2109.03216>.

Zielinski, P., Krishnan, S., and Chatterjee, S. Weak and strong gradient directions: Explaining memorization, generalization, and hardness of examples at scale, 2020. URL <https://arxiv.org/abs/2003.07422>.

A. Hat Matrix

To obtain the related expression familiar from classical influence function theory, let $Q = \hat{y}$ and

$$\delta C = \frac{\delta C}{\delta y} \delta y, \quad (9)$$

where δy is a target feature variation. Then

$$\delta \hat{y}^n = \mathcal{L}_V \hat{y}^n = -\Pi^{nl} \frac{\delta^2 C}{\delta \hat{y}^l \delta y^m} \delta y^m; \quad (10)$$

when C is the mean squared error cost, $\delta \hat{y} / \delta y = \Pi$. Equation 9 allows for investigating the influence of both physics-informed and numerical-routine aware data modifications, which we save for future work.

B. Determination of η^{-1}

In practice, we include an ℓ_2 -type regularizer corresponding to the weight-decay term used in AdamW driven training (Kingma & Ba, 2017; Loshchilov & Hutter, 2019). Although this penalty is not intrinsic to the model geometry—being defined with respect the ambient Euclidean coordinates—it remains a useful extrinsic regularizer when viewing the parameter manifold as embedded in a product of real coordinate spaces. Concretely, we consider the regularized metric

$$\eta_{\mu\nu} \rightarrow \eta_{\mu\nu} + \lambda \delta_{\nu\mu}, \quad (11)$$

with weight decay λ providing mass to the zero modes of η , thereby weakly lifting its flat directions.

We apply an iterative matrix-free solver, specifically the CRAIG method (Orban & Arioli, 2017) provided by the Krylov.jl package (Montoisson & Orban, 2023), which is formally equivalent to conjugate gradient descent, to efficiently approximate the required sensitivities. A direct inversion to determine η is not computationally feasible because of the large number of trainable parameters in our models. While this approach does not leverage commonly used scalable approximations (George, 2021; TransferLab, 2024), such approximations do not provide error control. When evaluating our UNet models, we determine the action of η with a relative error tolerance of 1.5×10^{-2} . We reach similar absolute error for our ViT models on using a relative tolerance of 5×10^{-2} ; our ViTs have both fewer parameters and smaller dominant NTK eigenvalues than our UNets [see Figure 8 and Figure 35].

C. Compressible Euler

The compressible Euler equations in two-spatial dimensions can be expressed in terms of four continuity equations, each of which is of the form

$$\partial_t \rho_c + \nabla \cdot \mathbf{J}_c = 0, \quad (12)$$

where ρ_c is a conserved density, \mathbf{J}_c is the associated conserved current, and c designates mass, momentum, and energy. Equation 12 follows directly from symmetry arguments: invariance under time translation yields energy conservation, spatial translation invariances implies momentum conservation, and an underlying global phase symmetry provides mass conservation. When Equation 12 is defined with periodic boundary conditions, the volume integrals of the conserved densities remain exact invariants for all time. Thus, both the local continuity relations and their associated global constraints must be respected: the domain-integrated mass, the two Cartesian components of momentum, and the total energy may not drift at any time during the rollout. Any surrogate or reduced-order model that aspires to physical fidelity must therefore honour these integral invariants, in addition to satisfying the differential conservation laws Equation 12.

C.1. Integral Invariants

For definiteness, we show that mass, momentum, and energy are conserved in this system. To this end, recall that

$$J_{\text{mass}}^j = \rho_{\text{mass}} v^j \quad (13a)$$

$$J_{\text{mom}}^{ij} = \rho_{\text{mass}} v^i v^j + p \delta^{ij} \quad (13b)$$

$$J_{\text{energy}}^j = (\rho_{\text{energy}} + p) v^j \quad (13c)$$

where p is the pressure and δ_{ij} is the Kronecker delta. Having introduced pressure as a fifth dynamical variable, a constitutive relation is needed in order to arrive at a closed system of equations, which is achieved on writing the energy density as

$$\rho_{\text{energy}} = \rho_{\text{mass}} e + \frac{1}{2} \rho_{\text{mass}} |\mathbf{v}|^2, \quad (14)$$

where the specific internal energy e is related to the pressure, and using the ideal-gas law

$$p = (\gamma - 1) \rho_{\text{mass}} e, \quad (15)$$

with $\gamma = 1.4$ the adiabatic index of a diatomic gas.

While the continuity equations control the pointwise evolution of the densities, global conservation guarantees that the total amount of each conserved quantity is invariant under the flow map. Integrating the continuity equation for any density ρ_c over the periodic domain Ω and applying integration by parts (or, equivalently, the divergence theorem) yields

$$\frac{d}{dt} \int_{\Omega} \rho_c d\mathbf{A} + \underbrace{\int_{\partial\Omega} \mathbf{J}_c \cdot \mathbf{n} dS}_{\text{vanishes for periodic } \Omega} = 0. \quad (16)$$

Therefore,

$$Q_c(t) = \int_{\Omega} \rho_c d\mathbf{A} \quad (17)$$

is an integral of motion. This statement precludes secular drift of mass, momentum, or energy in long simulations, and serves as a primary evaluation metric for surrogate models. Note that the Navier-Stokes equations can also be expressed in the form Equation 12 and therefore also admit mass, momentum, and energy as conserved variables. However, in order to sensibly train a neural network on both Compressible Euler and Navier-Stokes, one should attach to the model input an indication of whether or not J_{mom} contains a viscous term.

D. Supplementary Figures

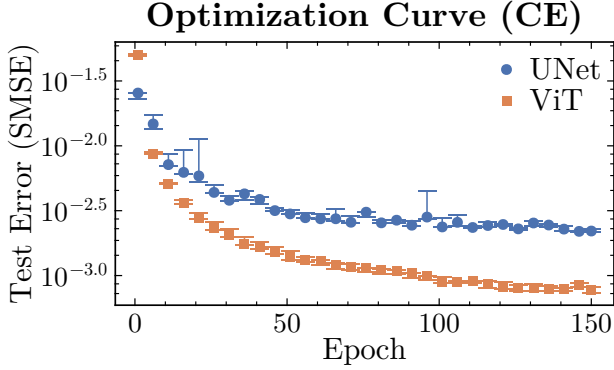


Figure 10. Optimization trajectories for the CE task, shown as test scaled mean squared error (SMSE) versus training epoch for our UNets and our ViTs. Markers denote the median performance, while vertical range bars indicate the variance across model seeds at each epoch. Training is halted in the late-stage learning regime to test whether the induced local gradient geometry is sufficient to identify a physically consistent, generalizing solution.

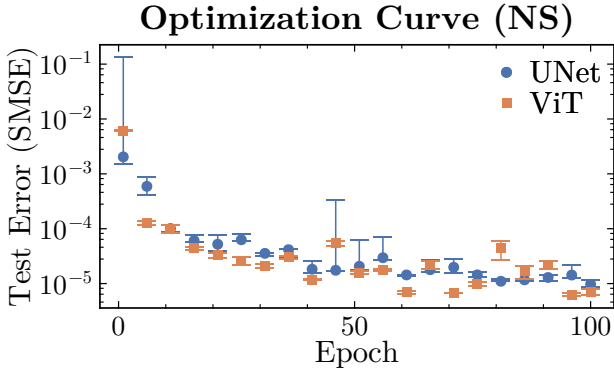


Figure 11. Same as Figure 10, except for NS data. In contrast to the CE problem, which involves four fields and has greater variability in time, one-step NS emulation is learned to comparable proficiency by both models.

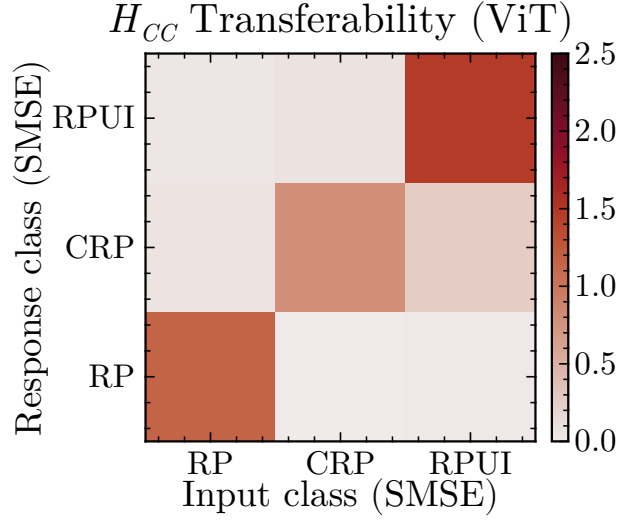


Figure 12. Class-to-class transferability matrix for our ViTs trained on the three-class CE split labeled RP, CRP, and RPUI. Each entry reports the time-averaged response of a given class (vertical axis) produced by test example gradients from the input class (horizontal axis). Diagonal dominance indicates class-locked gradient geometry; substantial off-diagonal values would imply reuse of dynamical features across classes.

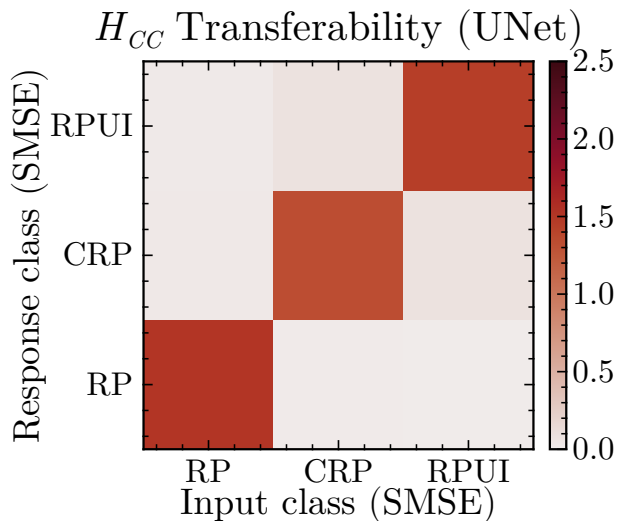


Figure 13. Class-to-class transferability matrix for our UNets trained on the CE split, with the same interpretation as Figure 12.

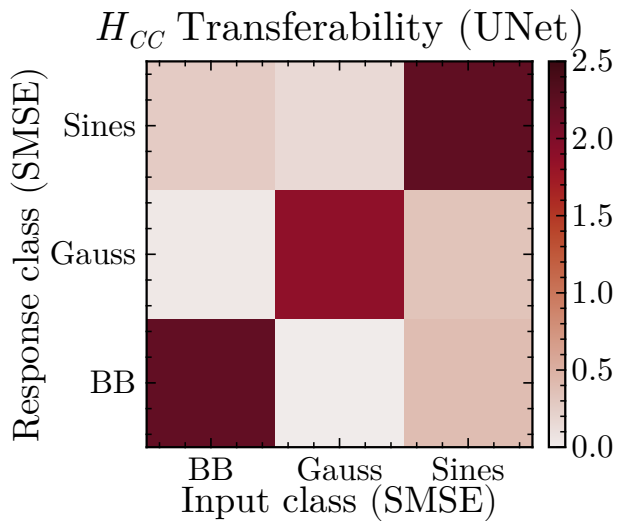


Figure 15. Same as Figure 14 except for out UNets.

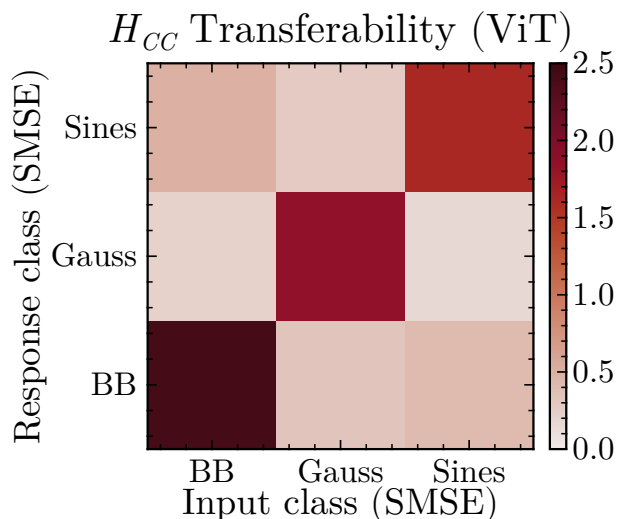


Figure 14. Same as Figure 12 except for NS data. Class-to-class transferability matrix for our UNets under the alternate three-class NS split labeled BB, Gauss, and Sines. Each cell is a time-averaged influence from one class to another, summarizing whether the learned representation supports feature sharing between qualitatively distinct initial-condition families. Relatively weak off-diagonal structure indicates that training organizes the data into largely disjoint gradient sectors.

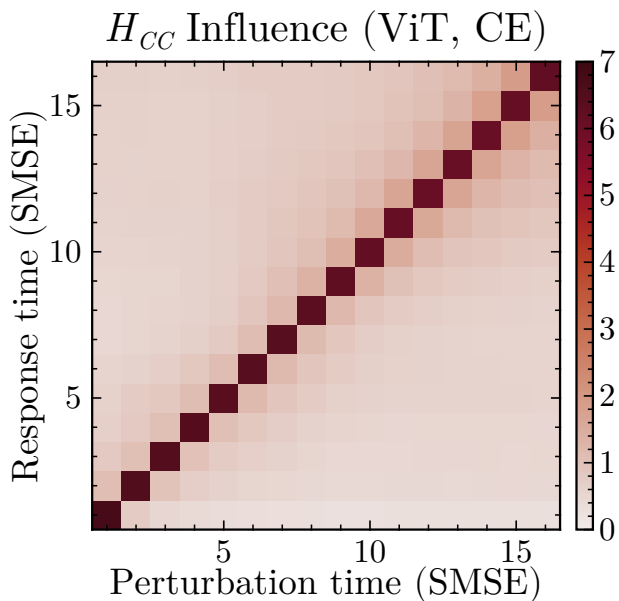


Figure 16. Heatmap of two-time influence for our ViTs trained on CE data, shown as a function of perturbation time (horizontal axis) and response time (vertical axis). Each pixel reports the intra-class averaged response induced by test example gradients at the perturbation time and the resulting change in test loss at the response time. A narrow diagonal ridge corresponds to time-local sensitivity consistent with interpolation, rather than generalization; substantial off-diagonal structure would indicate time-transferable learning.

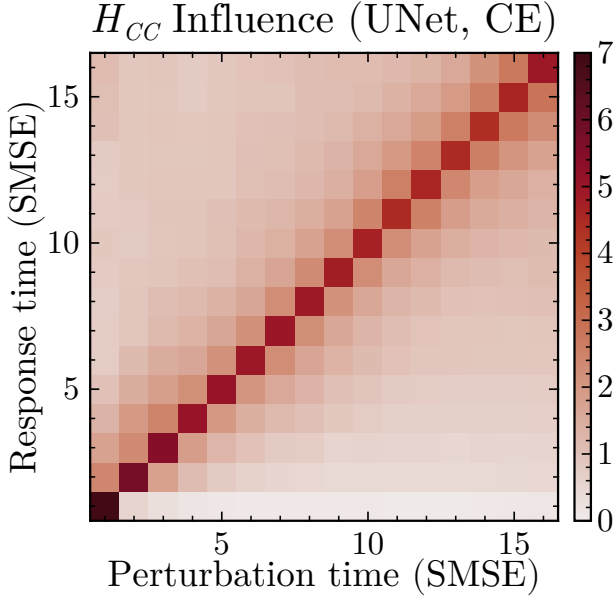


Figure 17. Same as Figure 16, except for our UNets. Desired off-diagonal coherence would indicate genuine temporal generalization.

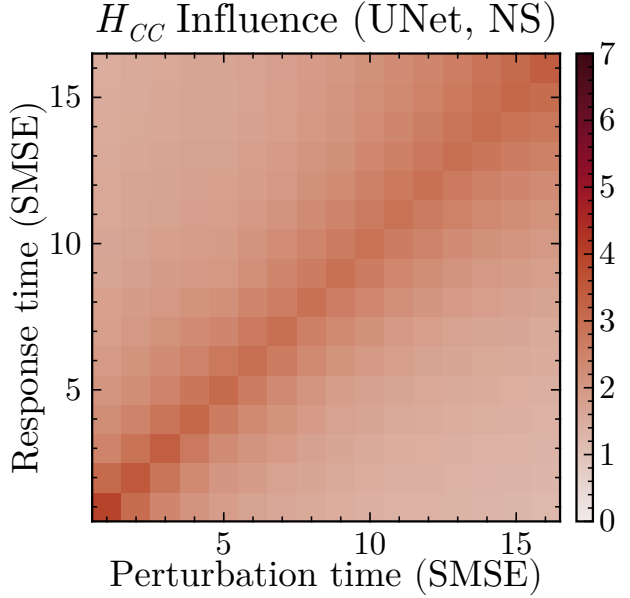


Figure 19. Same as Figure 18, except for our UNets. The enhanced temporal coherence relative to compressible Euler is consistent with viscosity-regularized dynamics.

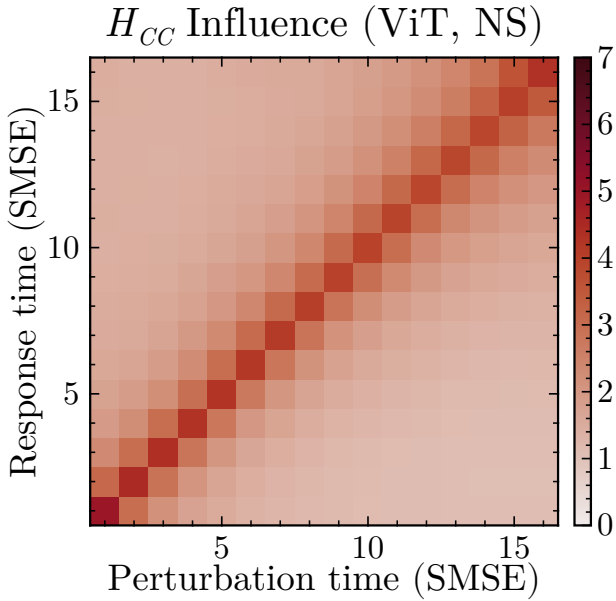


Figure 18. Same as Figure 16, except for NS data. The structure measures how test-time gradient information at one stage of a rollout affects loss at another stage.

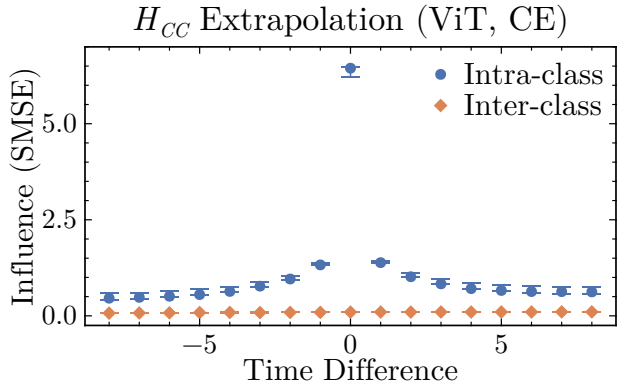


Figure 20. Time-lag summary of temporal transferability for our ViTs on CE data. The influence is averaged over all time-pairs of the same time difference and then split into intra-class pairs (gradient and response drawn from the same initial-condition class) and inter-class pairs (distinct initial-condition classes). Strong concentration near zero time difference indicates that gradient information fails to propagate coherently across time, while the lack off inter-class influence reveals an absence of physics consistent generalization.

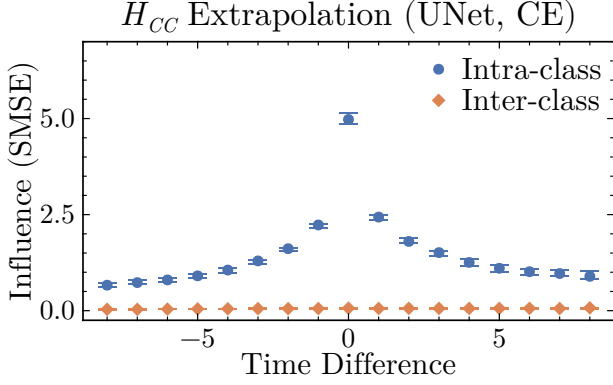


Figure 21. Same as Figure 20, except for our UNets. The gap between curves demonstrates that temporal coherence is predominantly a within-class phenomenon; the model does not learn to share dynamical structure across classes.

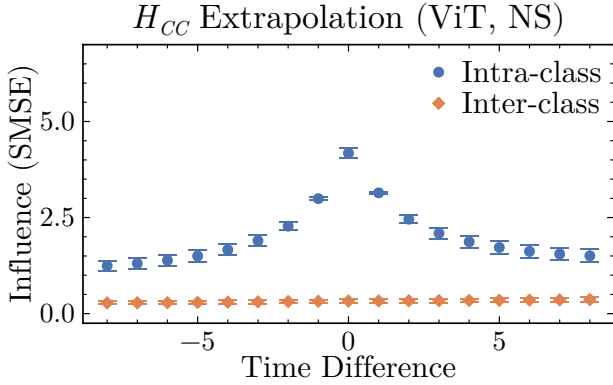


Figure 22. Same as Figure 20, except for NS data. The intra-class and inter-class averages separate temporal coherence within an initial-condition family from cross-family reuse. Persistent suppression of the inter-class curve indicates that the model behaves as a collection of class-conditioned local estimators rather than a shared dynamical emulator.

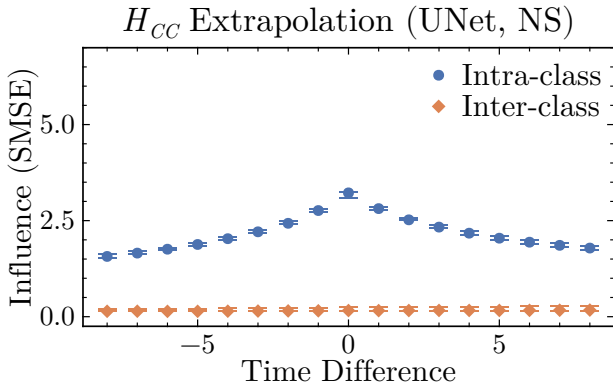


Figure 23. Same as Figure 22, except for our UNets.

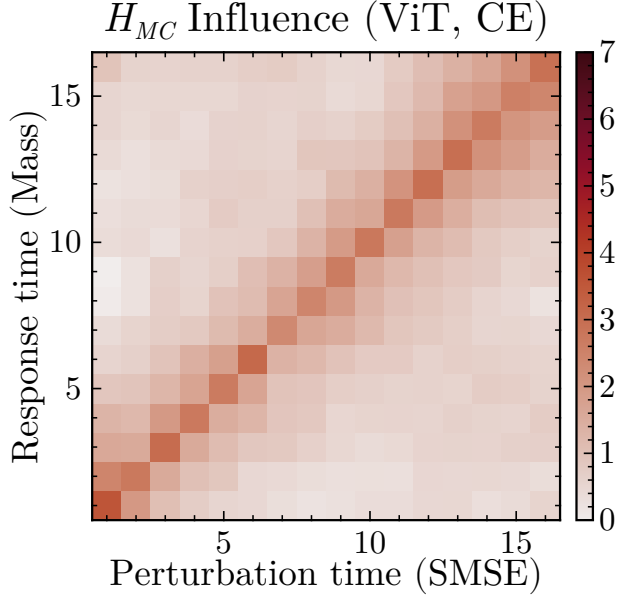


Figure 24. Two-time influence map for our ViTs on CE data when the response observable is the global mass-consistency signal. The horizontal axis indexes the time at which a test perturbation is applied, and the vertical axis indexes the time at which the mass-based response is evaluated. Off-diagonal support indicates that intra-class mass-related gradient information couples distant times.

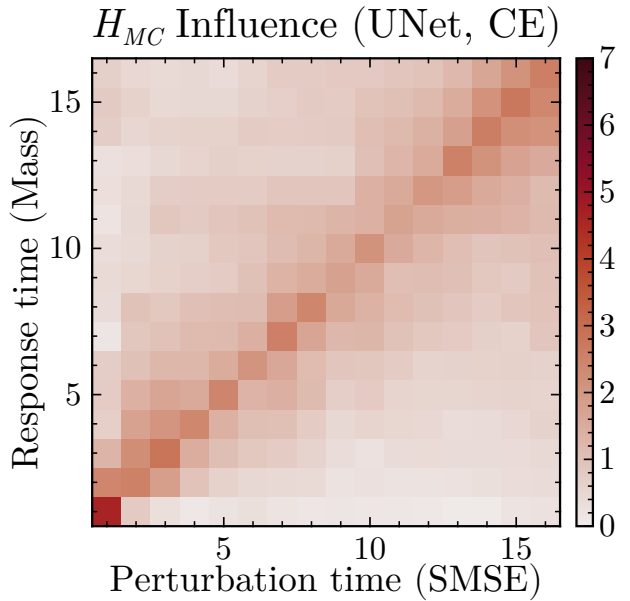


Figure 25. Same as Figure 6, except for our UNets.

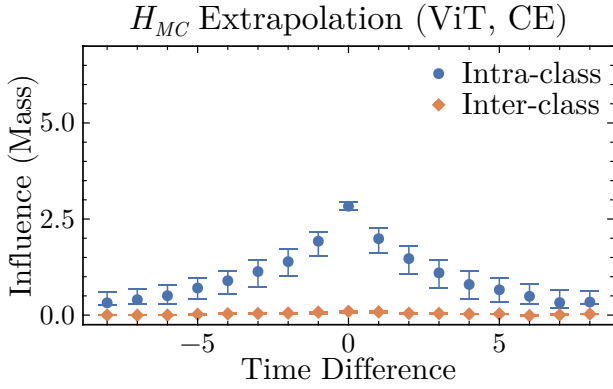


Figure 26. Time-lag transferability curve for our ViTs on CE data using the global mass-consistency observable. Intra-class averages measure how mass-based influence response align within an initial-condition class, while inter-class averages measure cross-class reuse of mass-related gradient directions. The decay pattern diagnoses whether conservation constraints induce transferable structure or remain class-locked.

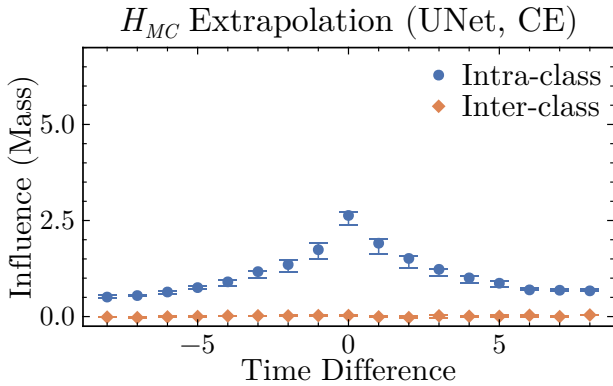


Figure 27. Same as Figure 7, except for our UNets. The intra-class and inter-class separation indicates that mass-informed gradients decouple different classes, and shared physics learning does not transfer out of class.

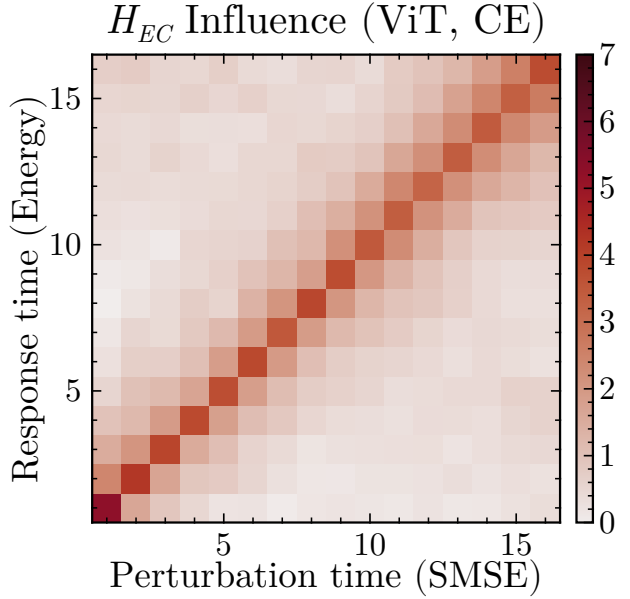


Figure 28. Two-time influence map for our ViTs on CE data when the response observable is the global energy-consistency signal. The horizontal axis indexes the time at which a test perturbation is applied, and the vertical axis indexes the time at which the energy-based response is evaluated. Off-diagonal support indicates that intra-class energy-related gradient information couples distant times.

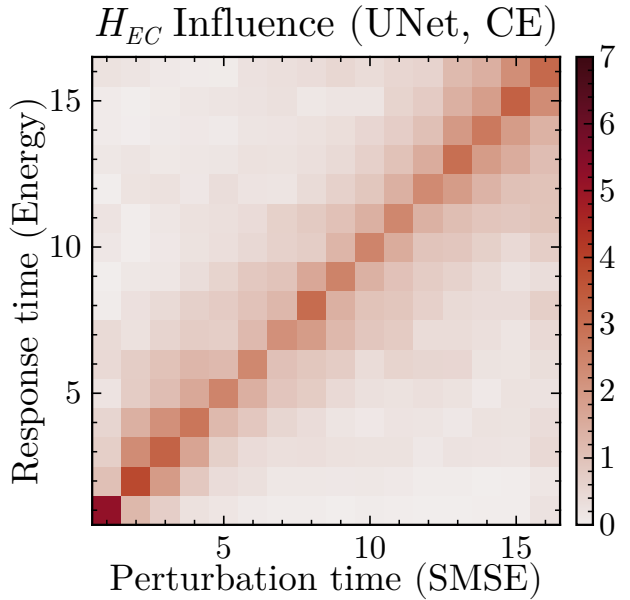


Figure 29. Same as Figure 28, except for our UNets.

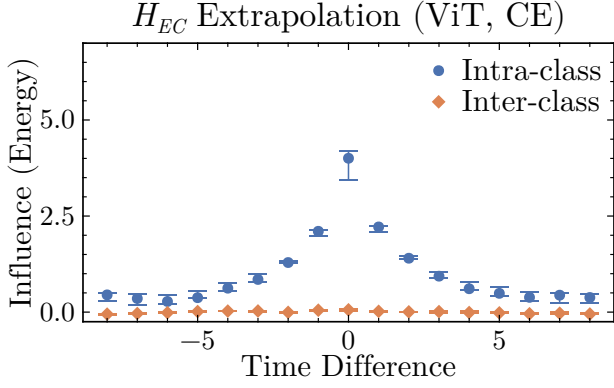


Figure 30. Time-lag transferability curve for our ViTs on CE data using the global energy-consistency observable. Intra-class averages measure how energy-based influence response align within an initial-condition class, while inter-class averages measure cross-class reuse of energy-related gradient directions. The decay pattern diagnoses whether conservation constraints induce transferable structure or remain class-locked.

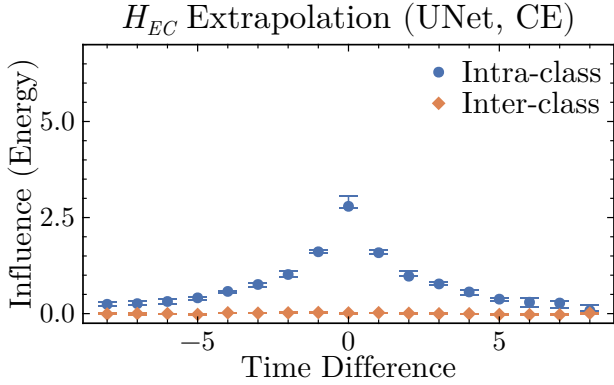


Figure 31. Same as Figure 30, except for our UNets. The intra-class and inter-class separation indices energy-informed gradients decouple different classes, and shared physics learning does not transfer out of class.

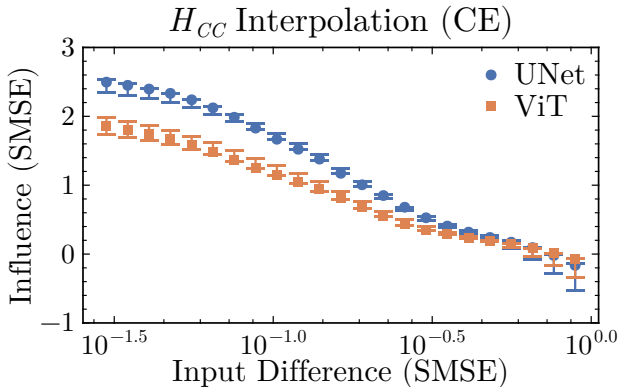


Figure 32. Curve fit of the influence as a function of feature-space separation between input states for CE data, comparing UNet and ViT, rangebars show uncertainty across seeds. Decay to zero indicates short-range locality on the learned data manifold, implying that parameter updates affect only nearby states and generalization is limited.

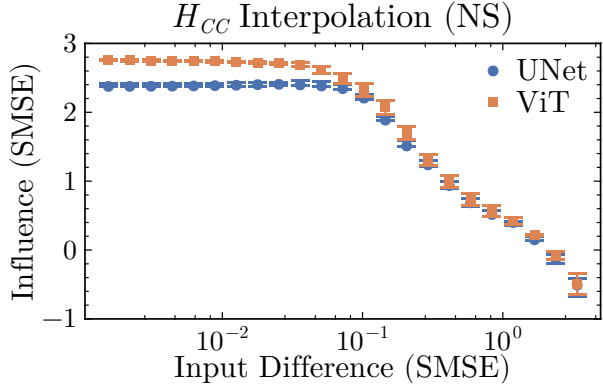


Figure 33. Same as Figure 5, except for NS data.

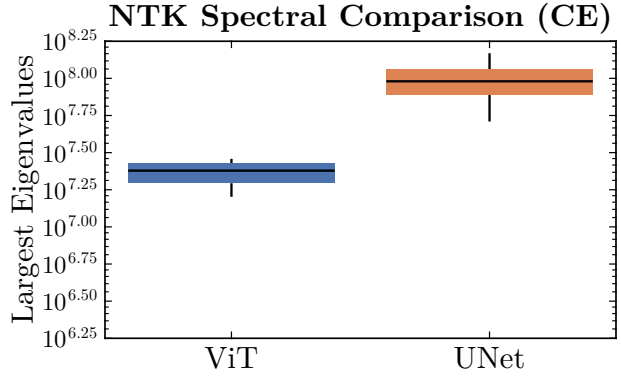


Figure 34. Spectral comparison of dominant neural tangent kernel eigenvalues for our UNets and our ViTs on CE data. The plotted distributions summarize the leading eigenvalue statistics across seeds and batches of the trained models. Larger dominant eigenvalues indicate a locally stiffer, sharper response geometry.

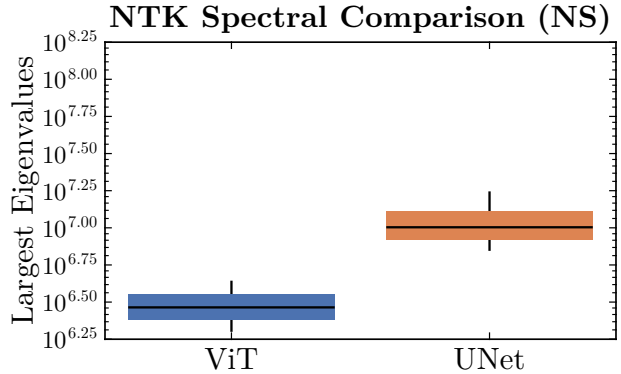


Figure 35. Same as Figure 8, except for NS data.