

A Graph-Based Inspection and Intervention Tool for Assessing Mechanistic Learning in PINNs

Adwait Patkhedkar, Alifaraz Lakhani, Prathmesh Mohite, Abhijeet Salunke

Sardar Patel Institute of Technology, Mumbai, India

Corresponding author :adwait.patkhedkar23@spit.ac.in

Abstract

We ask whether physically meaningful correspondences discovered inside a trained scientific model remain meaningful outside the conditions under which they were discovered. We introduce GIIT (Graph-based Inspection and Intervention Tool), which represents governing physics as a computational physics dependency graph, maps graph nodes to internal network components via sensitivity- and trend-based discovery, and tests the resulting mapping under targeted intervention. Evaluating on temporal extrapolation out-of-distribution (OOD) regimes of increasing severity, we find that temporal extrapolation is associated with layer-wise correspondence drift and reduced functional correspondence. Specifically, while the model achieves low physics residual in-distribution (1.915×10^{-4} on full ID and 1.01×10^{-4} on an ID sub-window $t \in [0.5, 0.8]$), the functional mapping changes even before leaving the training domain, and the shift continues as the evaluation window extends beyond the training domain: residual error increases from 2.81×10^{-2} on the boundary-crossing window ($t \in [0.5, 1.5]$) to 3.30×10^{-1} on severe OOD ($t \in [1, 2]$), accompanied by a systematic leftward shift of internal layer mappings, where the average winner layer index drops from 5.71 (ID) to 3.00 (ID sub-window) down to 2.29 (severe OOD). Only 2 of 7 physical nodes maintain stable layer assignments across the boundary-crossing window, and only 1 of 7 under severe extrapolation, revealing potential internal functional correspondence changes before severe degradation in conventional output metrics. Additional results for linear oscillatory systems are further detailed in the appendix.

Keywords: Physics-Informed Neural Networks, Mechanistic Interpretability, Scientific Machine Learning, Representation Analysis, Intervention-Based Interpretability, Out-of-Distribution Generalization, Physical Reasoning

1 Introduction

Physics-Informed Neural Networks (PINNs) [Raissi et al., 2019] embed differential-equation residuals directly into the training loss, achieving strong performance in low-data, high-simulation-cost regimes. However, a near-zero physics residual and low output error are a form of *numerical* constraint satisfaction, not a guarantee of *mechanistic* correctness: a trained PINN’s hidden layers may encode a representation with little correspondence to the causal structure of the system it was fit to [Wang et al., 2024, Krishnapriyan et al., 2021].

General frameworks for understanding model behavior [Murdoch et al., 2019] and XAI approaches tailored to scientific settings [Roscher et al., 2020] both emphasize that explanations should align with domain knowledge, but the dominant tools remain correlational: saliency and SHAP-style attribution, and counterfactual explanation methods [Chou et al., 2022], operate on model *inputs* rather than internal components. Causal approaches such as Causal Dependence Plots [Loftus et al.,

2024] and DAG-based SHAP [Ter-Minassian et al., 2023] use causal structure to sharpen input-level explanations, but likewise stop at the input-output boundary. Work that does look inside the model tends to go the other direction (discovering causal structure *from* data or representations [Renero et al., 2025]), whereas we start from known physical relationships and test whether a trained model is consistent with them. Operator-learning methods such as DeepONet [Lu et al., 2021] extend PINN-style training to function-space mappings but do not test internal correspondence either. Our framing draws on the role of causal structure in scientific explanation [Buijsman, 2023] and on causal abstraction as a theoretical basis for mechanistic interpretability [Geiger et al., 2025], applying these ideas to PINNs via an explicit mapping between physical variables and internal tensors, followed by targeted intervention. Recent work has used PINNs to quantify the sensitivity of physical solutions to governing parameters through automatic differentiation. Existing PINN sensitivity analysis primarily studies physical-output sensitivity to physical parameters [Hanna et al., 2024]; GIIT studies sensitivity and intervention at the level of internal neural representations.

This motivates our central question: *do physically meaningful correspondences discovered inside a model remain meaningful when the model is evaluated outside the conditions under which the correspondence was found?* Specifically, we evaluate whether internal functional correspondences remain consistent when transitioning from in-distribution interpolation ($t \in [0, 1]$) through progressively more severe temporal extrapolation regimes.

Contributions. We make three key contributions:

- **The GIIT Pipeline:** We introduce a tool that automatically links physical concepts (like mass or damping) to specific internal layers of a neural network using sensitivity analysis.
- **Intervention Testing:** We provide a way to test these links. By applying controlled multiplicative perturbations to specific layers, we test whether altering the identified components changes the corresponding physical observables.
- **Out-of-Distribution (OOD) Analysis:** We test whether these intervention-validated functional correspondences hold up under distribution shift. By tracking mapping discovery across evaluation regimes, we observe a systematic shift toward earlier-layer winners as the evaluation window moves from full ID through late-ID, boundary-crossing, and severe OOD regimes.

1.1 The GIIT Framework

Figure 1 summarizes the pipeline: a physical specification is compiled into a physics graph; a trained neural model is parsed for candidate tensors; a variable/parameter mapping is discovered between graph nodes and tensors; targeted interventions are applied to mapped tensors; and the resulting change in physical observables is compared against the analytically expected response.

2 Method

2.1 Physics Graph and Executable Ruleset

We represent the governing physical mechanism as a directed graph $G = (V, E)$, where V is a set of physical variables (parameters, e.g. mass, damping, viscosity; intermediate quantities, e.g. gradients; and outputs) and E encodes known directional dependencies (e.g. mass $\rightarrow$ acceleration). Some nodes are direct model outputs; others are derived via a small *executable ruleset* on top of automatic differentiation, e.g. $x \rightarrow x_t, x_{tt}$, or computed observables such as amplitude $x \rightarrow A = \max(x) - \min(x)$. The graph can be supplied directly as a JSON schema by a domain expert, or generated from source code with optional LLM assistance. The graph specifies the physical variables and their dependencies; the ruleset specifies *how* derived observables are computed from model output.

2.2 Neural-Physical Mapping

GIIT supports gradient-based sensitivity for output-derived variables and intervention-based sensitivity for parameters; in the experiments reported here, we use controlled perturbation sensitivity for the mapping analysis.

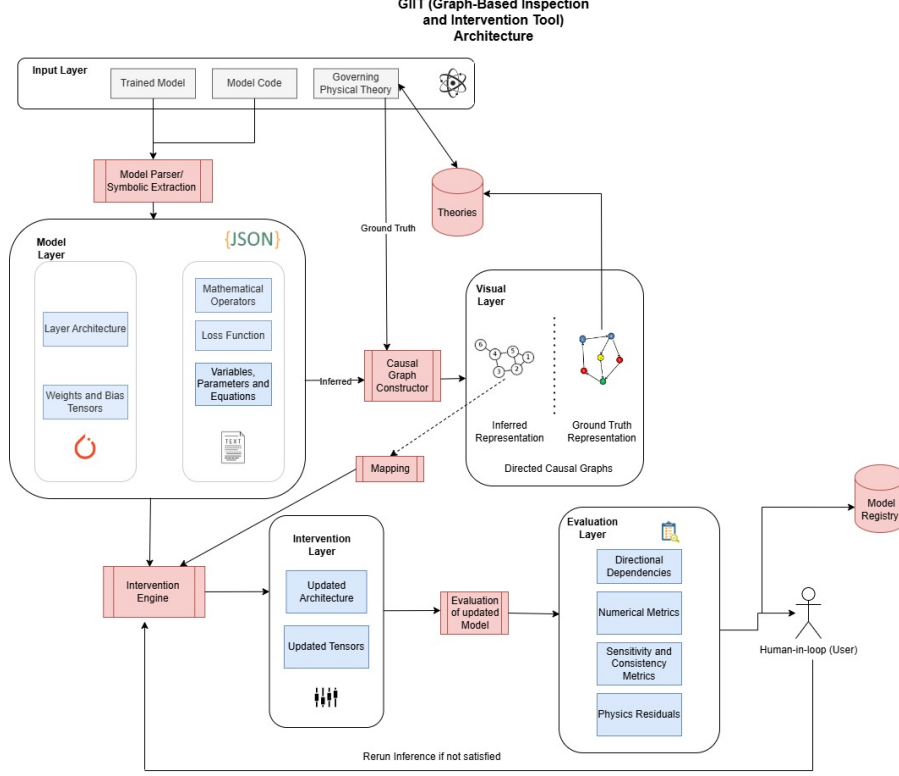


Figure 1: GIIT pipeline: Physical specification $\rightarrow$ Physics graph $\rightarrow$ Neural model $\rightarrow$ Variable/parameter mapping $\rightarrow$ Intervention $\rightarrow$ Physical observables $\rightarrow$ ID $\rightarrow$ OOD transfer test.

Sensitivity-based gradient mapping. For a node corresponding to a quantity computable from model output via a known formula F_n , GIIT can compute $q = \text{mean}(F_n(y))$ for $y = M(x)$ and evaluate gradient sensitivity across candidate layers ℓ :

$$S_{\text{grad}}(W_\ell, q) = \text{mean} \left| \frac{\partial q}{\partial W_\ell} \right|. \quad (1)$$

Trend-based perturbation mapping. Because governing parameters (e.g., mass, damping, or viscosity) are not direct network outputs, GIIT evaluates them through downstream behavioral observables computed by the ruleset (such as amplitude or frequency), assessing both response magnitude and directional consistency. Specifically, we apply a controlled multiplicative perturbation $W'_\ell = (1 + \epsilon)W_\ell$ (where the injected perturbation $\epsilon = \sigma|\mathcal{N}(0, 1)| \geq 0$ is strictly positive to preserve unambiguous directional trend evaluation), run a full forward pass, and measure the relative change in the target observable,

$$S_{\text{pert}} = \frac{|q(W'_\ell) - q(W_\ell)|}{|q(W_\ell)| + \delta}. \quad (2)$$

Candidate layers are ranked primarily by response magnitude; when a parameter has a known physical direction, we additionally evaluate whether the observed change has the expected sign. Tensor interventions are used to identify intervention-validated functional correspondences. Separately, analytical derivatives of the physics residual with respect to physical parameters ($\partial \mathcal{L}_R / \partial p$, where $\mathcal{L}_R = \text{mean}(R^2)$) quantify residual-loss sensitivity. Perturbing a tensor demonstrates that altering the identified component changes the target observable, rather than formal proof that the tensor uniquely or exclusively encodes the physical concept.

2.3 Transfer Protocol

To test whether discovered functional mappings transfer beyond discovery conditions, we define the full in-distribution training domain $\mathcal{D}_{\text{ID}} = \{t \in [0, 1]\}$, an in-distribution sub-window $\mathcal{D}_{\text{sub}} = \{t \in$

$[0.5, 0.8]$ (a subset of ID isolating late-time dynamics before the boundary), and two progressively severe out-of-distribution extrapolation domains: $\mathcal{D}_{\text{OOD}}^{(1)} = \{t \in [0.5, 1.5]\}$ (boundary-crossing window) and $\mathcal{D}_{\text{OOD}}^{(2)} = \{t \in [1, 2]\}$ (severe extrapolation). The protocol is:

1. Discover the baseline node-to-layer mapping distribution $\mathcal{M}_{\text{ID}} : V \rightarrow \mathcal{L}$ across multiple perturbation replicates on one trained model using full ID data.
2. Independently evaluate mapping distributions across the ID sub-window $\mathcal{D}_{\text{sub}}$ and each extrapolation domain $\mathcal{D}_{\text{OOD}}^{(i)}$ using the identical multi-replicate perturbation protocol.
3. Quantify layer-wise correspondence drift by evaluating modal layer stability (whether the top-mapped component is preserved) and tracking systematic layer index shifts across regimes, alongside predictive error (L_2) and physics residual adherence ($\text{mean}(R^2)$).

This tests whether functional correspondences remain stable under distribution shift, rather than relying only on output-level performance.

3 Experimental Setup

We evaluate GIIT across the four regimes defined in Section 2.3, using the 1D viscous Burgers’ equation benchmark ($u_t + uu_x = \nu u_{xx}$) adapted from PINNsTorch [Bafghi and Raissi, 2023] (`burgers_continuous_forward`), with spatial domain $x \in [-1, 1]$. We evaluate predictive accuracy (L_2 error), physics adherence (mean residual norm), and internal functional correspondence using targeted interventions and parameter-gradient residual sensitivity. For mapping discovery, we use perturbation sensitivity rather than gradient sensitivity because gradient sensitivity can also reflect bias parameters. L_2 error and residuals are computed on a uniform 100×100 space-time grid against a high-resolution finite-difference reference. For the Burgers’ experiments, mapping analysis uses 100 perturbation replicates on a single trained model ($\sigma = 0.1$, replicate index $\in [0, 99]$); the damped harmonic oscillator benchmark uses 21 perturbation replicates as detailed in Appendix B.

4 Results

4.1 Baseline Model Behavior

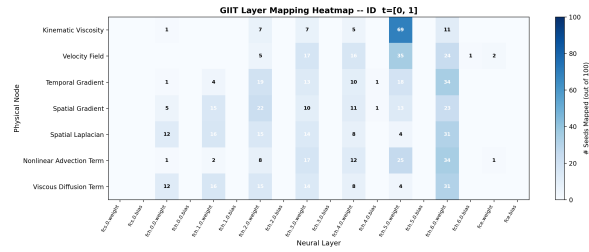
In-distribution, the PINN achieves tight physics adherence and low prediction error (Table 1). Baseline perturbation mapping identifies the deepest layers as dominant: viscosity (ν) and velocity (u) primarily map to `fch.5.0.weight`, while gradient and curvature terms map to `fch.6.0.weight`. However, viscosity shows an important failure: although consistently localized to `fch.5.0.weight`, its intervention response usually has the opposite of the expected physical direction. Thus, high sensitivity alone does not establish a physically valid correspondence.

Table 1: Viscous Burgers’ equation: ID (full and sub-window), boundary-crossing, and OOD extrapolation windows ($t \in [1, 2]$). A node mapping is defined as stable if its modal winning layer matches the full-ID baseline.

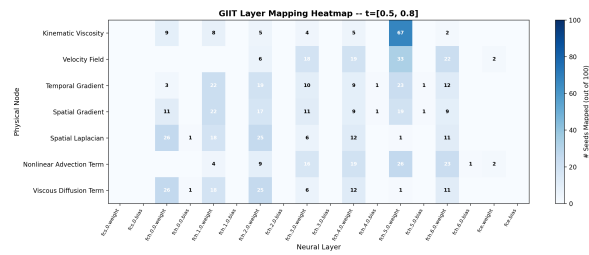
Regime	L_2 (%)	Residual $\text{mean}(R^2)$	$\partial \mathcal{L}_R / \partial \nu$	Stable Mappings
ID ($t \in [0, 1]$)	3.855	1.915×10^{-4}	−1.705	— (baseline)
ID Sub-window ($t \in [0.5, 0.8]$)	5.413	1.011×10^{-4}	−5.064	2/7 (28.6%)
Boundary-crossing ($t \in [0.5, 1.5]$)	5.883	2.811×10^{-2}	−73.576	2/7 (28.6%)
OOD ($t \in [1, 2]$)	31.752	3.301×10^{-1}	−154.115	1/7 (14.3%)

4.2 Correspondence Drift under Distribution Shift

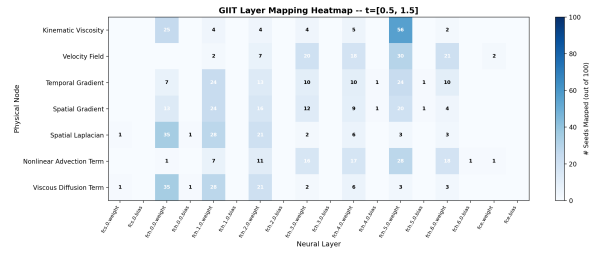
Mapping distributions shift toward earlier layers as evaluation moves away from the full-ID regime (Table 2, Figure 2). This reorganization is already evident within the training domain: curvature-related terms ($u_{xx}, \nu u_{xx}$) migrate to the earliest hidden layer (`fch.0.0.weight`) in the late-time ID sub-window, reducing stable mappings to 2/7. The drift continues through boundary-crossing into OOD extrapolation, where only velocity (u) retains its baseline modal assignment.



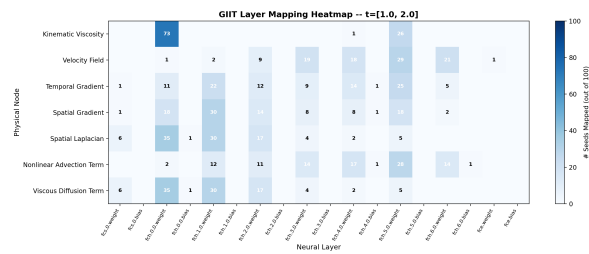
(a) In-Distribution ($t \in [0, 1]$)



(b) ID Sub-window ($t \in [0.5, 0.8]$)



(c) Boundary-crossing window ($t \in [0.5, 1.5]$)



(d) OOD Extrapolation ($t \in [1, 2]$)

Figure 2: GIIT layer mapping distributions across 100 perturbation replicates ($\sigma = 0.1$) for Burgers’ equation under ID and OOD regimes. Cell values show winner counts per layer, highlighting systematic leftward migration ($5.71 \rightarrow 2.29$).

Table 2: Neural Component Mapping Distributions across 100 Perturbation Replicates on One Trained Model ($\sigma = 0.1$, replicates $\in [0, 99]$) for ID (full and sub-window) and extrapolation regimes. The average winner layer index (bottom row) is the mean modal hidden layer index ($k \in [0, 6]$ for hkw) across all 7 physical nodes. Abbreviations: $hkw = fch.k.O.weight$.

Physical Node	ID ($t \in [0, 1]$)		ID Sub ($[0.5, 0.8]$)		Bound.-cross ($[0.5, 1.5]$)		OOD ($[1, 2]$)	
	Top	Cons.	Top	Cons.	Top	Cons.	Top	Cons.
Kinematic Viscosity (ν)	h5w	69/100	h5w	67/100	h5w	56/100	h0w	73/100
Velocity Field (u)	h5w	35/100	h5w	33/100	h5w	30/100	h5w	29/100
Temporal Gradient (u_t)	h6w	34/100	h5w	23/100	h1w	24/100	h5w	25/100
Spatial Gradient (u_x)	h6w	23/100	h1w	22/100	h1w	24/100	h1w	30/100
Spatial Laplacian (u_{xx})	h6w	31/100	h0w	26/100	h0w	35/100	h0w	35/100
Nonlinear Advection (uu_x)	h6w	34/100	h5w	26/100	h5w	28/100	h5w	28/100
Viscous Diffusion (νu_{xx})	h6w	31/100	h0w	26/100	h0w	35/100	h0w	35/100
Avg. layer index	5.71		3.00		2.43		2.29	

4.3 Output Degradation vs. Internal Correspondence

Output-level performance and internal representation stability diverge across regimes (Table 1). The boundary-crossing regime reveals substantial internal reorganization despite relatively stable predictive error ($L_2 \approx 5.88\%$). In the OOD regime ($t \in [1, 2]$), predictive accuracy deteriorates sharply, accompanied by a corresponding increase in physics residual. Residual-loss sensitivity also increases under distribution shift, remaining negative as increasing ν locally reduces the physics residual. However, it does not capture the layer-wise correspondence shifts detected by GIIT.

5 Discussion and Limitations

The observed leftward migration suggests that the functional role of network layers is context-dependent. One possibility is that deeper representations become less sensitive during extrapolation, causing interventions to select earlier layers; activation-level analysis is needed to distinguish this from a broader redistribution of sensitivity. Thus, internal representation stability cannot be inferred from low output loss alone.

The identified tensors should be interpreted as intervention-validated functional correspondences, not unique causal encodings or isolated modular circuits. The viscosity result further shows that high response magnitude does not guarantee the expected physical direction. Finally, our evidence is limited to two low-dimensional dynamical systems, so generalization to larger or higher-dimensional SciML models remains untested.

6 Conclusion

GIIT provides a model-agnostic framework for testing whether internal neural representations retain functional correspondence with known physical structure. Our experiments show that these correspondences can change substantially under temporal distribution shift, including within the nominal training domain. This suggests that output-level accuracy and physics residuals alone may be insufficient for diagnosing changes in internal functional organization. Future work should test whether these effects persist across independently trained models and higher-dimensional scientific systems.

References

- Reza Akbarian Bafghi and Maziar Raissi. PINNs-torch: Enhancing speed and usability of physics-informed neural networks with pytorch. In *The Symbiosis of Deep Learning and Differential Equations III*, 2023. URL <https://openreview.net/forum?id=nl1ZzdHpab>.
- Stefan Buijsman. Causal scientific explanations from machine learning. *Synthese*, 202(6), 2023.
- Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022. doi: 10.1016/j.inffus.2021.11.003.
- Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL <https://jmlr.org/papers/v26/23-0058.html>.
- John M. Hanna, José V. Aguado, Sebastien Comas-Cardona, Ramzi Askri, and Domenico Borzacchiello. Sensitivity analysis using physics-informed neural networks. *Engineering Applications of Artificial Intelligence*, 135:108764, 2024. doi: 10.1016/j.engappai.2024.108764.
- Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 26548–26560, 2021.
- Joshua R. Loftus, Lucius E. J. Bynum, and Sakina Hansen. Causal dependence plots. In *Advances in Neural Information Processing Systems*, volume 37, pages 112656–112683, 2024. URL <https://arxiv.org/abs/2303.04209>.
- Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021. doi: 10.1038/s42256-021-00302-5.
- W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019. doi: 10.1073/pnas.1900654116.
- Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. 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. doi: 10.1016/j.jcp.2018.10.045.
- Jesus Renero, Roberto Maestre, and Idoia Ochoa. Rex: causal discovery based on machine learning and explainability techniques. *Pattern Recognition*, page 112491, 2025.
- Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020. doi: 10.1109/ACCESS.2020.2976199.
- Lucile Ter-Minassian, Oscar Clivio, Karla Diazordaz, Robin J. Evans, and Christopher C. Holmes. PWSHAP: A path-wise explanation model for targeted variables. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34054–34089. PMLR, 2023.
- Tao Wang, Bo Zhao, Sicun Gao, and Rose Yu. Understanding the difficulty of solving cauchy problems with pinns. In *6th Annual Learning for Dynamics & Control Conference*, pages 453–465. PMLR, 2024.

A Technical Appendix: Viscous Burgers' PINN Architecture and Trajectory Analysis

The viscous Burgers' PINN follows a fully connected architecture (FCN) matching `burgers_temporal_experiment.py`:

- **Input Layer (fcs):** Linear($2 \rightarrow 20$) $\rightarrow \tanh$, taking spatial and temporal coordinates $(x, t) \in [-1, 1] \times [0, 1]$.
- **Hidden Blocks (fch):** 7 sequential hidden blocks indexed 0 through 6 (fch.0.0 to fch.6.0), each consisting of Linear($20 \rightarrow 20$) $\rightarrow \tanh$ (total $N_{\text{LAYERS}} = 8$ layer stages).
- **Output Layer (fce):** Linear($20 \rightarrow 1$) projecting to the scalar velocity field prediction $\hat{u}(t, x)$.
- **Parameters:** 3,021 total trainable parameters with kinematic viscosity set to $\nu = 0.01/\pi \approx 0.003183$.

Figure 3 compares the trained PINN base prediction against the high-resolution numerical reference solution in-distribution ($t \in [0, 1], x \in [-1, 1]$), alongside the pointwise absolute error surface displaying localized discrepancy near the steepening shock front ($x \approx 0$).

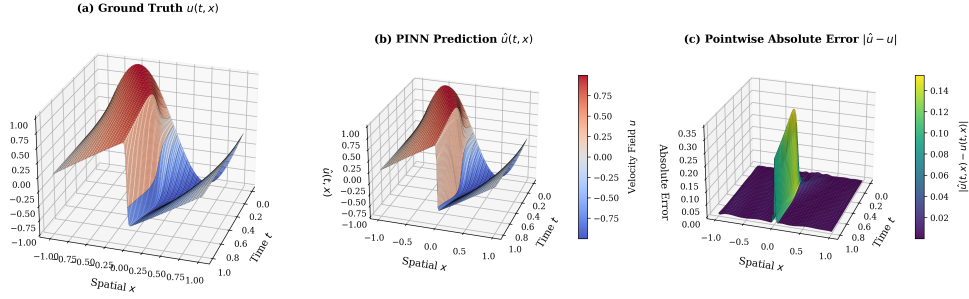


Figure 3: 3D comparison for the 1D viscous Burgers' velocity field across the in-distribution space-time domain $(t, x) \in [0, 1] \times [-1, 1]$: (a) high-resolution finite-difference reference solution $u(t, x)$, (b) PINN base model prediction $\hat{u}(t, x)$ ($L_2 \approx 3.86\%$), and (c) pointwise absolute error $|\hat{u}(t, x) - u(t, x)|$, illustrating that predictive errors are primarily concentrated along the steep internal shock layer at $x \approx 0$.

Under temporal extrapolation ($t \in [1, 2]$), model prediction diverges substantially from the governing physics (Figure 4), with relative L_2 error jumping to 31.75% and maximum pointwise absolute error exceeding 1.22. While the true solution exhibits gradual viscous dissipation, the unconstrained PINN exhibits severe unphysical distortion and trough sagging near $x \approx 0$.

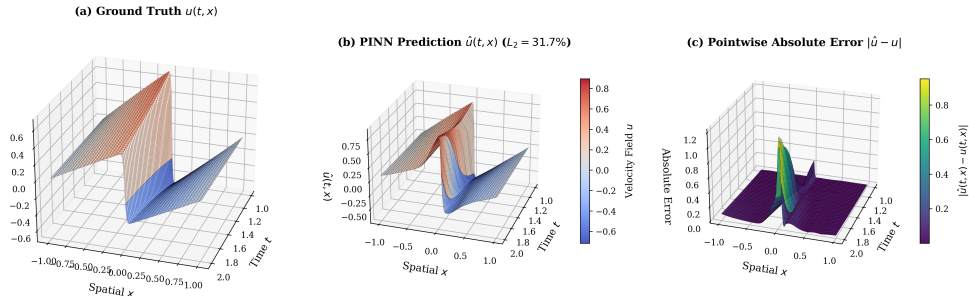


Figure 4: 3D comparison for the 1D viscous Burgers' velocity field under out-of-distribution temporal extrapolation ($t \in [1, 2], x \in [-1, 1]$): (a) high-resolution reference solution $u(t, x)$ continuing smooth viscous dissipation, (b) PINN model prediction $\hat{u}(t, x)$ showing severe unphysical deformation ($L_2 \approx 31.75\%$), and (c) pointwise absolute error $|\hat{u}(t, x) - u(t, x)|$, highlighting large localized error growth along the extrapolated shock front.

B Technical Appendix: Damped Harmonic Oscillator Benchmark

B.1 Experimental Setup and Baseline Results

We evaluated GIIT on the 1D damped harmonic oscillator ($m\ddot{x} + c\dot{x} + kx = 0$ with $m = 1, c = 4, k = 400$), modeled with a fully connected PINN (Linear($1 \rightarrow 32$) $\rightarrow$ tanh $\rightarrow$ Linear($32 \rightarrow 32$) $\rightarrow$ tanh $\rightarrow$ Linear($32 \rightarrow 32$) $\rightarrow$ tanh $\rightarrow$ Linear($32 \rightarrow 1$), 2,209 parameters), trained with a combined data-fit and physics-residual loss. The ID regime covers $t \in [0, 1]$ (500 uniform collocation points), while the OOD transfer regime tests temporal extrapolation over $t \in [1, 2]$ (500 collocation points) referenced against the analytical damped trajectory. Note that the reported physics residual (mean(R^2)) measures the raw, unnormalized squared ODE operator norm $(m\ddot{x} + c\dot{x} + kx)^2$; because $k = 400$, nominal values of kx and $\ddot{x}$ are on the order of 10^1 to 10^2 , which naturally leads to an ID residual scale of $\sim 2 \times 10^2$ (contrasting with Burgers’ equation where normalized fields and $\nu \approx 0.003$ yield smaller numerical residual scales). Under OOD extrapolation, this residual explodes by nearly three orders of magnitude to 1.317×10^5 .

Table 3: Oscillator results across In-Distribution ($t \in [0, 1]$) and Out-Of-Distribution ($t \in [1, 2]$) regimes.

Regime	Pred. L_2 error	Physics residual	Layer correspondence
ID ($t \in [0, 1]$)	1.13%	2.007×10^2	Distinct parameter mappings identified
OOD ($t \in [1, 2]$)	2007.34%	1.317×10^5	Mass/damping collapse to shared tensor; spring shifts

In-distribution, the model accurately predicts displacement ($L_2 = 1.13\%$) with residual-loss sensitivities $\frac{\partial \mathcal{L}_R}{\partial m} = +1377.28$ and $\frac{\partial \mathcal{L}_R}{\partial c} = -3.84$. Across the 21 perturbation replicates on one trained model, mapping identifies hidden tensors associated with the physical parameters (modal winners `fch.1.0.weight` and `fch.0.0.weight`).

B.2 DHO Out-of-Distribution Transfer Analysis

Evaluating mapping distributions under the OOD regime ($t \in [1, 2]$) reveals significant divergence in residual-loss sensitivity and layer-wise functional correspondence. As shown in Table 3 and Figure 5, while the model achieves low predictive error in-distribution ($L_2 = 1.13\%$), its OOD predictive error explodes ($L_2 = 2007.34\%$), and the physics residual norm increases by nearly three orders of magnitude (1.317×10^5). This catastrophic divergence occurs because the true oscillator trajectory rapidly decays toward equilibrium ($x \rightarrow 0, \dot{x} \rightarrow 0$), whereas the neural network ceases oscillating near $t \approx 1.0$ and drifts toward an unphysical fixed point ($\hat{x} \approx -1.19, \hat{\dot{x}} \approx 0$). In phase space (Figure 5b), while the analytical ground truth forms a decaying spiral converging to the equilibrium origin $(0, 0)$, the PINN trajectory departs from the physical limit cycle at the $t = 1.0$ boundary and freezes into a spurious fixed point at $(-1.19, -0.19)$. Dividing the non-decaying prediction error by this vanishing true signal drives the relative L_2 error to 2007.34%, while the stiff restoring spring force ($k = 400$) causes the squared ODE residual $(m\ddot{x} + c\dot{x} + kx)^2$ to explode past 10^5 . In addition, the residual-loss sensitivities shift substantially under extrapolation: $\frac{\partial \mathcal{L}_R}{\partial m}$ shifts from +1377.28 to -1747.92 and $\frac{\partial \mathcal{L}_R}{\partial c}$ shifts from -3.84 to +579.62 (detailed in Table 4). Furthermore, re-running auto-mapping in the OOD regime demonstrates that mass and damping mappings collapse to a shared early-layer tensor (`fch.0.0.weight`), while the spring-constant mapping shifts to a neighboring early layer (`fch.1.0.weight`).

Table 4: Residual-loss Sensitivities ($\partial \mathcal{L}_R / \partial p$).

Physical Parameter	ID ($t \in [0, 1]$)	OOD ($t \in [1, 2]$)
Mass (m)	+1377.28	-1747.92
Damping Coeff (c)	-3.84	+579.62
Spring Constant (k)	+111.77	+458.12

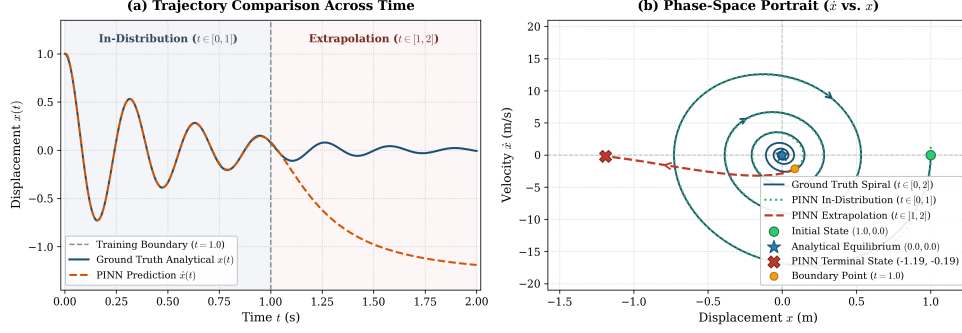


Figure 5: Damped harmonic oscillator dynamics in-distribution ($t \in [0, 1]$) and under extrapolation ($t \in [1, 2]$): **(a)** displacement $x(t)$ comparing ground truth against the PINN, which diverges to a flatline ($\hat{x} \approx -1.19$) after $t = 1.0$; **(b)** phase portrait ($\dot{x}$ vs. x) showing the true trajectory spiraling into the origin $(0, 0)$, whereas the PINN terminates at $(-1.19, -0.19)$.

B.3 Perturbation-Replicate Stability Analysis

To rigorously evaluate whether the discovered neural-physical mappings are artifacts of a specific random noise draw or represent stable topological properties of the trained PINN, we executed 21 independent perturbation replicates on one trained model (replicate $\in [0, 20]$ with noise std $\sigma = 0.1$).

Table 5 summarizes the modal (top) mapped neural tensor, the number of replicates where that tensor was selected as the highest-scoring component, and the distribution across candidate layers.

Table 5: Neural Component Mapping Distributions across 21 Perturbation Replicates (replicate $\in [0, 20]$, $\sigma = 0.1$) for both In-Distribution ($t \in [0, 1]$) and Out-Of-Distribution ($t \in [1, 2]$) evaluation.

Physical Variable	In-Distribution ($t \in [0, 1]$)		Out-Of-Distribution ($t \in [1, 2]$)	
	Top Component	Consistency	Top Component	Consistency
Mass (m)	fch.1.0.weight	10/21 (47.6%)	fch.0.0.weight	16/21 (76.2%)
Damping Coeff (c)	fch.1.0.weight	10/21 (47.6%)	fch.0.0.weight	16/21 (76.2%)
Spring Constant (k)	fch.0.0.weight	8/21 (38.1%)	fch.1.0.weight	12/21 (57.1%)
Displacement (x)	fch.1.0.weight	11/21 (52.4%)	fch.1.0.weight	16/21 (76.2%)
Velocity ($\dot{x}$)	fch.0.0.weight	17/21 (81.0%)	fch.0.0.weight	13/21 (61.9%)
Acceleration ($\ddot{x}$)	fch.0.0.weight	16/21 (76.2%)	fch.0.0.weight	10/21 (47.6%)

Key Observations from Multi-Replicate Analysis:

- **Derivative Localization Stability:** Physical derivative states exhibit high consistency in-distribution. Velocity ($\dot{x}$) and Acceleration ($\ddot{x}$) map to the first hidden block (fch.0.0.weight) in 81.0% and 76.2% of perturbation replicates, respectively.
- **OOD Correspondence Reorganization:** Under out-of-distribution temporal extrapolation ($t \in [1, 2]$), mass and damping mappings collapse to a shared early-layer tensor (fch.0.0.weight in 76.2% of replicates), while the spring-constant mapping shifts to a neighboring early layer (fch.1.0.weight in 57.1% of replicates), demonstrating that the network’s internal capacity to separate functional parameter correspondences degrades outside the training envelope.

B.4 Supplementary Material

Anonymized code, dataset configurations, and model weights are available at <https://github.com/adwaitdagitguy/GIIT>.