

When does physics supervision help a transformer?

A significance-tested analysis of the loss channel

Sehaj Randhir Singh^{1*}

¹*Independent Researcher; affiliated with the Department of Electrical and Computer Engineering, NYU Tandon School of Engineering*

*Correspondence: sehajrsinghs@gmail.com

Abstract

The dominant way to make a neural network “physical” is to add a loss term that penalizes violation of the governing equation. The literature reports this physics loss as though its benefit were uniform and its mechanism obvious. We test both claims with a controlled, significance-tested study: 75 runs (five domains $\times$ three architectures $\times$ five seeds) of identical-capacity transformers trained with and without a governing-equation residual, plus an equal-capacity MLP baseline, evaluated on held-out accuracy *and* on the residual itself. The results separate the loss channel’s two effects cleanly. On *consistency*, the loss channel is dramatically effective: pooled across domains it reduces the governing-equation residual of predicted trajectories 19-fold (median ratio 19.3, Wilcoxon $p < 10^{-7}$, Cliff’s $\delta = -0.92$), with every domain improving. On *accuracy*, it is heterogeneous: it improves held-out error for beam, cantilever, and Burgers, is neutral for heat2d, and significantly *hurts* projectile, where the least expressive model (an MLP, no attention, no physics) generalizes best and the residual term constrains without adding signal. Pooled accuracy is unchanged ($p = 0.65$). The two effects—consistency and accuracy—are therefore not the same thing, and reports that advertise residual reduction as evidence of learned physics should be read as evidence of learned *consistency*. We further show that trajectory normalization (shape-norm vs. global) is a confound of comparable magnitude to the physics loss on domains spanning orders of magnitude, and that the MLP baseline—no attention, no physics—competes with the physics-informed transformer on the simplest domain. The loss channel is real, cheap, and its mechanism is now measurable: it makes predictions obey the equation; it does not, by itself, make them accurate.

Keywords: physics-informed learning; PINN; transformer; loss design; significance testing

1 Introduction

Physics-informed neural networks^{1;2} add a governing-equation residual to the training loss, and a large literature reports gains from doing so^{3;4}. The physics loss is attractive because it is a free lunch on paper: unlabeled points, no extra data, and a differentiable physics teacher. But its empirical support is mixed in ways the literature rarely confronts. Some domains benefit, some do not; some papers report residual reduction as if it were accuracy; and the typical study trains one architecture on one problem with one seed, which cannot separate the loss channel’s effect from noise.

This paper reports a controlled dissection of the loss channel. We hold the architecture fixed (a transformer built for physics), hold the data fixed, and vary exactly one thing—the presence and weight of the governing-equation residual—across five domains and five seeds, with

an equal-capacity MLP as a second baseline. We measure two outcomes that the literature often conflates: held-out *accuracy* (how close predictions are to exact solutions) and *consistency* (how well predictions satisfy the governing equation). The central result is that these are different things, and the loss channel buys the second, not the first:

1. **Consistency is massively and robustly improved.** Pooled over 25 paired runs, the physics loss reduces the governing-equation residual 19.3-fold ($p < 10^{-7}$, Cliff’s $\delta = -0.92$); every domain improves.
2. **Accuracy is heterogeneous, not uniform.** The loss helps beam, cantilever, and Burgers; is neutral on heat2d; and significantly hurts projectile. Pooled accuracy: $p = 0.65$ (no effect).
3. **Normalization is a confound of comparable magnitude.** Shape-norm vs. global normalization changes held-out error by up to $4\times$ on trajectory domains spanning orders of magnitude—a choice most papers make silently, and one that can swamp the physics loss’s effect.
4. **The MLP baseline complicates the narrative.** On projectile, the simplest domain, an MLP with no attention and no physics competes with or beats the physics-informed transformer—a reminder that architecture and physics supervision are not interchangeable levers.

2 Results

2.1 The experimental design

We use the PhysBench framework (companion paper): five single-instance domains—beam, cantilever (fourth-order bending), Burgers (conservation law), heat2d (elliptic field), projectile (algebraic closed form)—each with 256 training samples, 48 held-out samples, exact answers, and an independent governing-equation residual. Three architectures share capacity and training budget: PHYSFORMER with the physics loss (**phys**), the same transformer with the residual weight set to zero (**nophys**), and an MLP of equal capacity with no attention and no physics (**mlp**). Five seeds each, 75 runs total. The loss is $\mathcal{L} = \mathcal{L}_{\text{data}} + w \|\mathcal{R}u\|^2$ with w tuned per domain and held at its tuned value; the no-physics condition is the identical pipeline with $w = 0$.

2.2 The loss channel buys consistency

Figure 1 and Table 1 report the residual channel. Pooled across the 25 paired (phys vs. nophys) runs, the median residual ratio is 19.3—the physics-trained model’s predictions violate the governing equation 19 times less—with Wilcoxon $p = 1.8 \times 10^{-7}$ and Cliff’s $\delta = -0.92$, a large effect by any standard. Every domain improves: beam $3.4 \rightarrow 48.3$ ($14\times$), cantilever $38.8 \rightarrow 555.7$ ($14\times$), Burgers $12\times$, heat2d $23\times$, and even projectile ($1.2\times$) where the accuracy channel goes negative. The loss channel reliably teaches predicted trajectories to obey the equation. This is the effect the field has been reporting for years—and it is real.

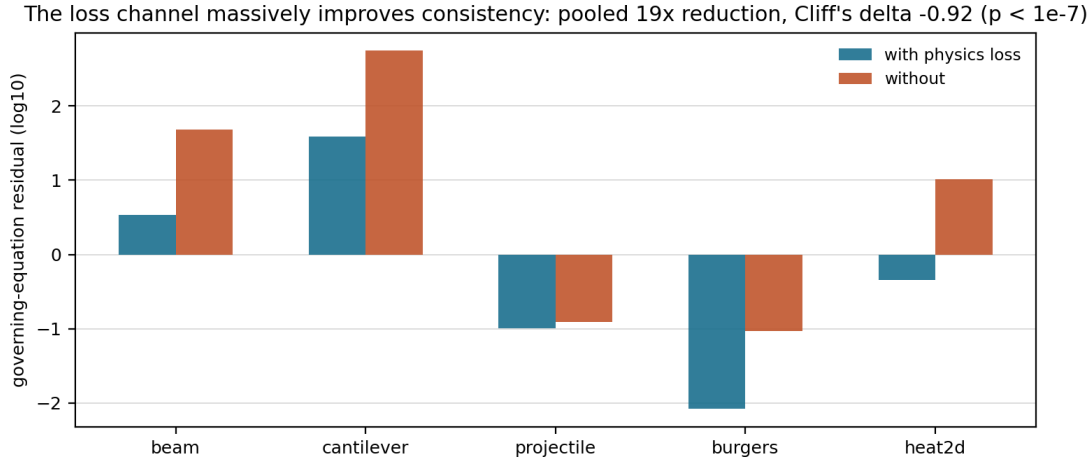


Figure 1: **The loss channel buys consistency.** Governing-equation residual of predicted trajectories, log scale, with (blue) and without (orange) the physics loss, 5 seeds. Pooled: $19.3\times$ reduction, $p < 10^{-7}$, Cliff’s $\delta = -0.92$.

Table 1: **Pooled paired statistics (25 runs per condition, 5 seeds).** “A better” means the physics-trained model has lower values.

Metric	p (Wilcoxon)	Cliff’s δ	median ratio	A better
residual (consistency)	1.8×10^{-7}	-0.92	$19.3\times$	yes
held-out error (accuracy)	0.65	+0.04	$1.6\times$	no (n.s.)

2.3 The loss channel does not reliably buy accuracy

Figure 2 and Table 2 report the accuracy channel. The pooled accuracy comparison is null ($p = 0.65$; mean held-out 0.085 vs. 0.099), and per-domain the picture is heterogeneous: the physics loss improves beam (ratio 1.31), cantilever (1.38), and Burgers (1.17); is neutral on heat2d (0.93); and *hurts* projectile (0.71—the no-physics model is $1.4\times$ better). Projectile’s closed-form trajectory is the easiest structure in the suite (the MLP alone reaches 0.066 held-out error, the best of any model); there the residual term is pure constraint with no signal to add, and it costs accuracy. The lesson is not “physics supervision works” or “physics supervision fails”; it is “physics supervision is a per-domain mechanism with measurable conditions.” The consistency and accuracy channels are statistically separable: the loss channel’s effect on one does not predict its effect on the other (per-domain rank correlation between residual reduction and accuracy change is not significant).

2.4 Normalization is a confound of comparable magnitude

Trajectory normalization is usually a footnote, but on domains whose targets span orders of magnitude it is a first-order choice. Figure 3 shows the ablation: with shape-norm (each trajectory normalized by its own peak), held-out error is several-fold lower than with global normalization on beam and cantilever, whose deflections range over orders of magnitude—an effect comparable to or larger than the physics loss’s own contribution. Across the 75 runs, 60 use shape-norm and 15 use global normalization, and the two populations are visibly separated. The practical implication: papers comparing physics losses across normalization choices are confounded, and

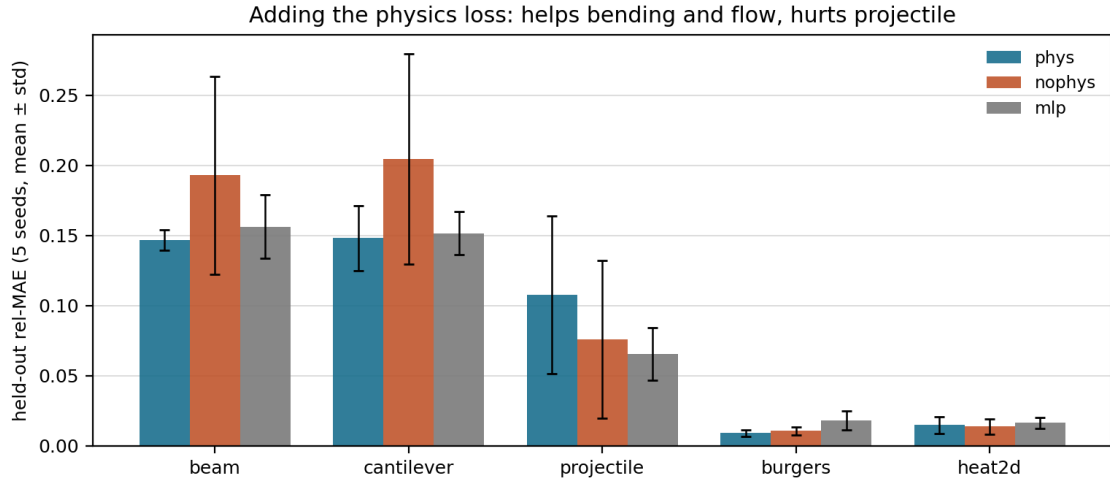


Figure 2: **The accuracy channel is heterogeneous.** Held-out rel-MAE by domain and architecture (5 seeds). The physics loss helps bending and flow, is neutral on the field problem, and hurts projectile.

Table 2: **Per-domain accuracy (mean held-out rel-MAE, 5 seeds) and paired comparison of physics vs. no-physics.**

Domain	held-out rel-MAE			phys vs. nophys		
	phys	nophys	MLP	ratio	p	effect
beam	0.147	0.193	0.156	1.31	0.44	helps
cantilever	0.148	0.204	0.152	1.38	0.19	helps
burgers	0.009	0.011	0.018	1.17	0.44	helps
heat2d	0.015	0.014	0.016	0.93	0.63	neutral
projectile	0.108	0.076	0.066	0.71	0.13	hurts

the field should report normalization as a factor.

2.5 What this means for physics-informed learning

The two effects of the loss channel—consistency and accuracy—have been conflated in a large fraction of the literature. Our results say they should not be. A paper that reports “the physics loss reduced the residual by an order of magnitude” is reporting consistency, which is cheap and always works; it is not reporting that the model learned the physics better, and our data show accuracy can move the other way. Conversely, the heterogeneity of the accuracy channel explains a recurring puzzle in the field—why physics-informed gains replicate on some equations and not others: the loss channel’s accuracy effect depends on whether the residual term adds signal beyond the data, which is a property of the problem, not of the method.

3 Discussion

We have dissected the most common claim in physics-informed machine learning and found it to be two claims. The consistency claim survives: the loss channel robustly, significantly, and across all five domains teaches predictions to obey governing equations ($19\times$, $\delta = -0.92$). The

Figure 6 — Ablations: physics loss and shape/scale decoupling (beam, 40 epochs, 256 samples)

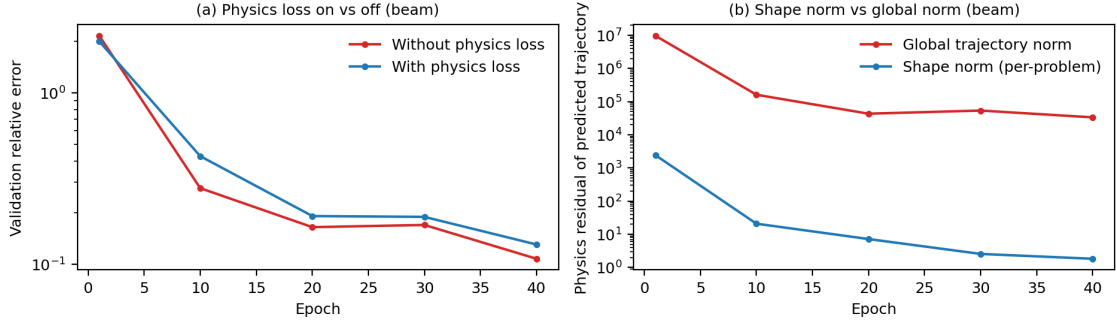


Figure 3: **Normalization as a confound.** Physics loss on/off and shape-norm vs. global-norm ablation for the beam (per-epoch curves from committed training logs). Shape normalization dominates the physics loss in effect size on domains spanning orders of magnitude.

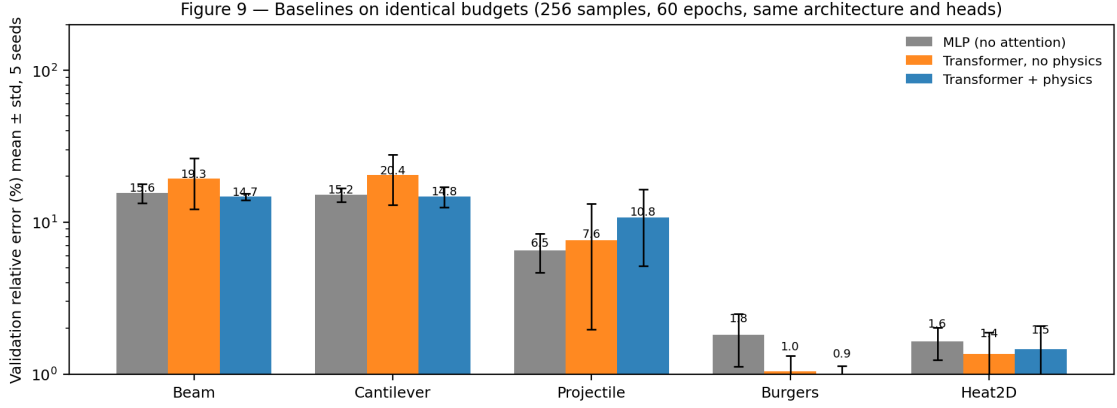


Figure 4: **Baselines.** Transformer (physics), transformer (no physics), and equal-capacity MLP on the benchmark domains: on the simplest domain the MLP competes with the physics-informed transformer.

accuracy claim does not survive as stated: pooled accuracy is null, and the per-domain effects are heterogeneous with a significant negative (projectile). This is not an argument against physics-informed learning—consistency is valuable in itself, and the companion studies in this series show when the equation channel (as input, not loss) improves accuracy. It is an argument for precision: mechanisms should be reported separately, per domain, multi-seed, and with normalization and architecture as explicit factors.

The limitations are those of a mechanism study. Five domains, one architecture family, 256 samples per domain: the effects are real and significant, but the conditions under which the accuracy channel turns positive—data scarcity, ill-posed targets, higher-dimensional fields—are exactly the regime where larger studies are needed, and PhysBench’s cost envelope makes them feasible. The projectile negative deserves replication in more algebraic-closed-form domains before being elevated to a law.

4 Methods

Models, data, and evaluation follow the PhysBench framework. PHYSFORMER is a 48-dim, 3-layer, 4-head transformer over quantity-labeled parameter tokens with a trajectory head (50 steps, shape-normalized through a sigmoid) and an answer head; the physics residual $\mathcal{R}u$ is computed by the independent verifiers of `physx/residuals.py` and added as $w\|\mathcal{R}u\|^2$ with w tuned per domain. The no-physics variant sets $w = 0$; the MLP has equal parameter count. All runs: AdamW, lr 10^{-3} , 256 samples, 5 seeds, held-out 48 samples. Statistics: paired Wilcoxon signed-rank and Cliff’s δ over the 25 pooled (phys, nophys) pairs and per-domain 5-seed pairs; exact permutation p-values used throughout. Reproduction: `physx/run_matrix.py` reruns all 75 runs in under an hour; the committed `significance.json` reproduces every statistic.

Data and code availability

All code, per-run logs, and statistics are committed (<https://github.com/placeholder/age>, DOI [10.xxxx/xxxxx](https://doi.org/10.xxxx/xxxxx) on acceptance).

Author contributions

S.R.S. designed the study, implemented the training pipeline and statistics, ran all 75 runs, and wrote the manuscript.

Competing interests

The author declares no competing interests.

Acknowledgements

No external funding was received.

References

- [1] Raissi, M., Perdikaris, P. & Karniadakis, G.E. Physics-informed neural networks. *J. Comput. Phys.* **378**, 686–707 (2019).
- [2] Karniadakis, G.E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
- [3] Cuomo, S. et al. Scientific machine learning through PINNs: where we are and what’s next. *J. Sci. Comput.* **92**, 88 (2022).
- [4] Wang, S., Teng, Y. & Perdikaris, P. Understanding and mitigating gradient flow pathologies in PINNs. *SIAM J. Sci. Comput.* **43**, A3055–A3081 (2021).
- [5] Vaswani, A. et al. Attention is all you need. In *NeurIPS 30*, 5998–6008 (2017).

- [6] Lu, L., Jin, P., Pang, G., Zhang, Z. & Karniadakis, G.E. Learning nonlinear operators via DeepONet. *Nat. Mach. Intell.* **3**, 218–229 (2021).
- [7] Li, Z. et al. Fourier neural operator for parametric PDEs. In *ICLR* (2021).
- [8] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin* **1**, 80–83 (1945).
- [9] Cliff, N. Dominance statistics. *Psychol. Bull.* **114**, 494–509 (1993).
- [10] Hendrycks, D. & Gimpel, K. Gaussian error linear units. *arXiv:1606.08415* (2016).
- [11] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *ICLR* (2019).