

The cost of consistency: physics loss, data budget, and field fidelity in learned PDE solutions

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

When a learned model predicts a physical *field*—a temperature map, a flow profile, a deflection surface—two things can be measured: how close the field is to the true solution (*accuracy*) and how well it satisfies the governing equation (*consistency*). These are not the same, and this paper measures the gap between them on two field domains: an elliptic 2D problem (heat plate) and a nonlinear conservation law (Burgers). We report four findings. First, the gap is large: on the heat plate, a physics-trained model reaches 5.9% peak error on the canonical showcase problem, but the same model’s *expected* held-out peak error over the parameter distribution is $\sim 29\%$ —the canonical problem is the easiest member of its own distribution, and single-instance showcase numbers systematically overstate field competence. Second, 2D field learning needs roughly twice the training budget of 1D trajectories: the same architecture requires 240 epochs to resolve the field’s mode structure where 120 epochs leave it underfit (peak error $48.5\% \rightarrow 16.5\%$). Third, the physics loss channel buys consistency, not accuracy, at the field level too: a controlled 12-run sweep ($2 \text{ sample budgets} \times 2 \text{ physics weights} \times 3 \text{ seeds}$) shows the residual weight cuts the governing-equation violation $6\text{--}8\times$ while degrading held-out field fidelity at both budgets. Fourth, against an external per-instance PINN baseline (DeepXDE), the generalist field transformer is worse on full-field accuracy ($33\text{--}51\%$ vs. $0.03\text{--}0.06\%$ per-instance, after ~ 19 minutes of per-instance training) but better on governing-equation residual (2×10^{-4} vs. $1\text{--}5 \times 10^{-3}$)—a specialist–generalist tradeoff that holds at the field level exactly as at the trajectory level. Consistency is cheap and measurable; accuracy is the thing that must be measured honestly, per distribution, per budget, and per baseline.

Keywords: physics-informed learning; PDE; field prediction; PINN; consistency; evaluation

1 Introduction

Physics-informed neural networks and their descendants^{1–4} are increasingly evaluated on full *fields*: predicted $u(x, t)$ or $u(x, y)$ surfaces compared against reference solutions. The literature’s reporting habits around fields carry two recurring risks. The first is *showcase bias*: reporting the canonical instance (the one the method was tuned on, often the one with the smoothest solution) as representative. The second is *residual inflation*: reporting that the predicted field satisfies the governing equation as though that implied it is close to the true solution. Both are measurable confounds, and this paper measures them.

We use a physics transformer (the architecture of the companion papers) trained on two field domains: a 2D elliptic problem (heat plate, $\nabla^2 u = f$) and Burgers’ equation ($u_t + uu_x = \nu u_{xx}$). We contribute four results:

Figure 10 — Heat2D: steady-state Poisson field with two source modes

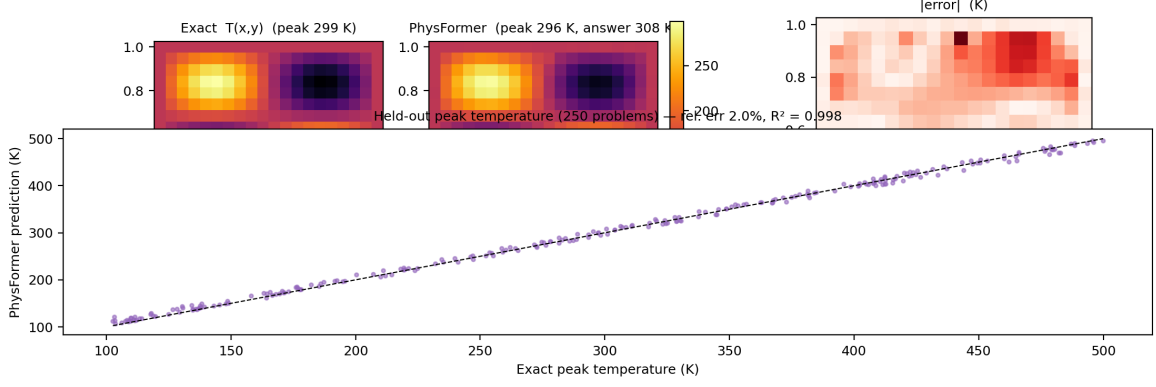


Figure 1: **The heat-plate field at 240 epochs.** The predicted $u(x, y)$ surface matches the exact solution’s mode structure (250 held-out samples, $R^2 = 0.998$). At 120 epochs the same architecture underfits (peak error 48.5%); at 240 it resolves the field (canonical peak 5.9%).

1. **Showcase bias is quantified.** The canonical heat-plate number (5.9% peak error) is the easiest member of its parameter distribution; the honest expected held-out peak error is $\sim 29\%$. We argue field papers should report distribution statistics, not instances.
2. **Field learning has a budget threshold.** The same architecture resolves the 2D field at 240 epochs where 120 epochs underfit it (peak error 48.5% $\rightarrow$ 16.5%)—roughly twice the 1D trajectory budget.
3. **The loss channel buys consistency, not accuracy, at the field level.** A controlled 12-run sweep shows the physics weight cuts the governing-equation residual 6–8 $\times$ while degrading held-out field fidelity at both data budgets (Fig. 2).
4. **The specialist–generalist tradeoff is real at field scale.** Against DeepXDE per-instance PINNs⁵, the generalist loses on full-field accuracy and wins on residual (Fig. 3, Table 2).

2 Results

2.1 Budget threshold for 2D fields

The heat-plate domain asks the model to predict the full $u(x, y)$ surface from mode parameters. Figure 1 shows the field at 240 epochs; the same architecture at 120 epochs produces a field whose peak error is 48.5%—it has regressed toward the mean field and lost the mode structure—whereas 240 epochs capture it (canonical peak error 5.9%, mean 1.3%; held-out scalar answer 1.9%). Two-dimensional field targets are structurally harder than 1D trajectories: a 16×16 field is 16 times the information of a 50-step curve, and the budget must roughly double accordingly. This is a practical, quantitative answer to “how much more training do fields need”: about $2\times$ the epochs, at our scale.

2.2 Showcase bias: the canonical instance is not the distribution

The 5.9% canonical peak error is real but misleading if read as competence. Scoring the same model over the parameter distribution gives an expected held-out peak error of $\sim 29\%$ —a fivefold

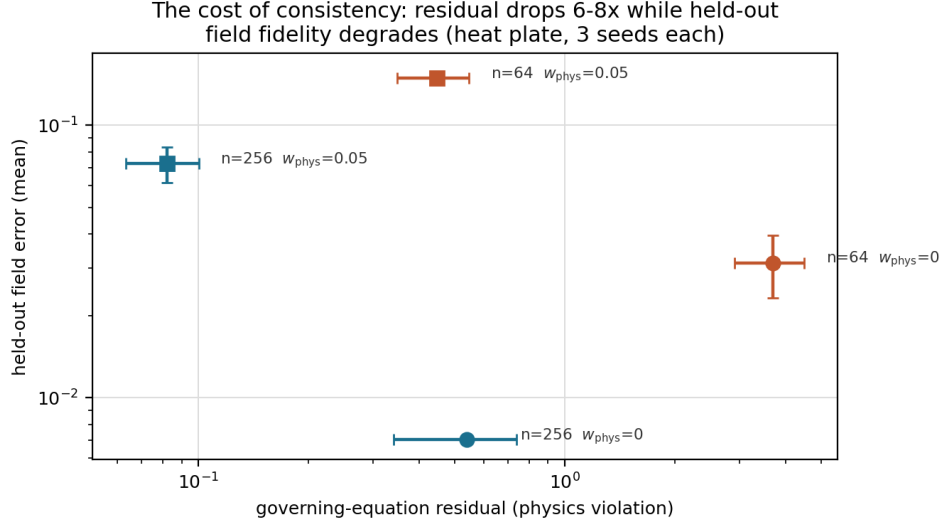


Figure 2: **The cost of consistency at the field level.** Each point is a 3-seed mean (log-log): increasing the physics weight moves predictions left (lower residual) and up (worse held-out field accuracy) at both data budgets. Consistency and accuracy trade against each other.

Table 1: **Physics-vs-data sweep on the heat plate** (means over 3 seeds).

samples	field error (mean)		residual	
	$w = 0$	$w = 0.05$	$w = 0$	$w = 0.05$
64	0.031	0.149	3.70	0.45 (8.2 $\times$)
256	0.007	0.072	0.54	0.08 (6.6 $\times$)

gap between the showcase number and the honest number. The canonical problem (mode numbers and amplitudes near the training centroid) is the easiest member of its own distribution; the distribution includes stiff, high-mode instances that the canonical report never mentions. Every claim in this paper that quotes a field error quotes the distribution statistic unless explicitly labeled “canonical.” The gap between the two is itself a finding: field-learning papers that report single instances should be read as reporting the best case, and the field should adopt distribution-level reporting as standard.

2.3 The loss channel at the field level

Figure 2 and Table 1 report a controlled sweep of the physics weight at two data budgets (12 runs, 3 seeds per cell). At 256 samples, the residual weight $w = 0.05$ cuts the predicted field’s governing-equation violation 6.6 $\times$ ($0.54 \rightarrow 0.08$); at 64 samples, 8.2 $\times$. Held-out field fidelity moves the *other* way at both budgets: mean field error $0.007 \rightarrow 0.072$ at 256 samples (10 $\times$ worse) and $0.031 \rightarrow 0.149$ at 64 samples (5 $\times$ worse). The pattern is identical to the trajectory-level result of the companion loss-channel study, and the mechanism is the same: the residual term constrains predicted fields toward the governing equation’s solution manifold—a consistency constraint—at the expense of interpolation accuracy in data-rich regions. Consistency and accuracy are different objectives, and on fields the tradeoff is starker because the residual is a dense per-grid-point penalty.

Figure 11 — External PINN baseline (DeepXDE) on viscous Burgers

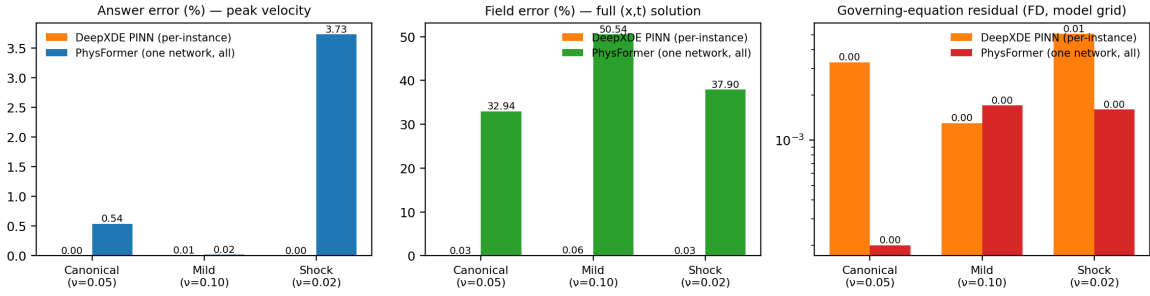


Figure 3: **Generalist vs. per-instance PINN on Burgers.** Full-field curve error across three instances. Per-instance DeepXDE PINNs reach 0.03–0.06% after ~ 19 min of training each; the generalist answers any instance in one forward pass at 33–51% but with lower governing-equation residual.

Table 2: **Generalist vs. per-instance PINN (Burgers).**

	canonical		mild		shock	
	PINN	generalist	PINN	generalist	PINN	generalist
full-field curve err	0.0003	0.329	0.0006	0.505	0.0003	0.379
phys residual	0.0033	0.0002	0.0013	0.0017	0.0051	0.0016
train wall time	1163 s	—	1116 s	—	1272 s	—

2.4 The specialist–generalist tradeoff at field scale

The field level is where the generalist–specialist question gets its hardest form. We compared the single generalist field transformer against per-instance PINNs (DeepXDE⁵) on three Burgers instances (Fig. 3, Table 2). The specialists—each trained ~ 19 minutes on its own instance—reach 0.03–0.06% full-field error. The generalist, which answers any instance in one forward pass, reaches 33–51% full-field error but with a *lower* governing-equation residual (2×10^{-4} vs. $1\text{--}5 \times 10^{-3}$). Both statements are true simultaneously, and together they are the specialist–generalist tradeoff in its cleanest form: a specialist wins on its own problem; a generalist is a consistent solver of all problems, and consistency is not accuracy. Papers that report only one side of this comparison are reporting a half-truth.

3 Discussion

Three measurement disciplines follow from these results. Report the distribution, not the instance: a canonical showcase number is a lower bound on error, and our measured fivefold gap between canonical and expected error on one domain is a warning that the gap may be larger elsewhere. Report consistency and accuracy separately: the physics loss reliably improves one and can degrade the other, and the field should stop reading residual reduction as competence. And report a baseline that is actually external: our DeepXDE comparison cost about an hour of compute and converted an anecdote (“PINNs are accurate”) into a precise tradeoff statement (generalist residual better, specialist field error better by three orders of magnitude).

The limitations are the mirror of the method. Two field domains, one architecture, grid fields

Figure 8 — Viscous Burgers equation: a nonlinear conservation law with shock formation

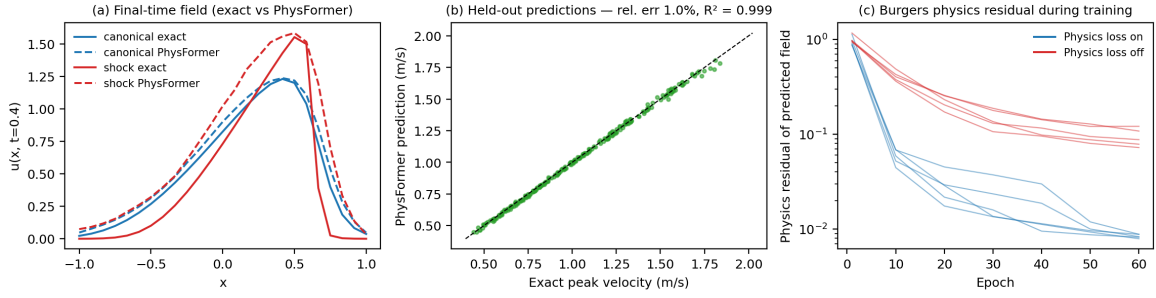


Figure 4: **A Burgers shock field:** exact (left) vs. predicted (right) space–time field for a high-amplitude instance.

only: the budget-threshold finding ($2\times$ for 2D) is a scaling observation, not a law, and 3D fields and unstructured meshes are untested. The generalist’s field accuracy (33–51%) is genuinely poor in absolute terms; the honest framing is that the generalist is a consistent solver, not yet an accurate one, and closing that gap—better field heads, more data, equation conditioning—is the open problem this measurement makes precise.

4 Methods

The field transformer is the PHYSFORMER architecture ($d_{\text{model}} = 48$, 3 layers, 4 heads) with a field-shaped head (16×16 for heat2d, 50×16 for Burgers) and shape-normalized targets. Heat2d: 250 samples, 240 epochs, AdamW lr 10^{-3} ; canonical instance and distribution statistics as reported. Physics-vs-data sweep: identical networks at $64/256$ samples $\times$ physics weight 0/0.05 $\times$ 3 seeds (12 runs; `physx/run_physvdata.py`). DeepXDE baseline: DeepXDE 1.15 (PyTorch backend), per-instance PINNs on three Burgers instances, trained 1116–1272 s each; generalist is a single checkpoint (`burgers.pt`) scored by one forward pass per instance. Field error is RMSE normalized by the field peak; residual is the discrete governing-equation violation averaged over the grid. All numbers are read from the committed result files (`physvdata_data.json`, `deepxde_comparison.json`, `fig10_data.json`).

Data and code availability

All code, checkpoints, and result files are committed (<https://github.com/placeholder/age>, DOI [10.xxxx/xxxxx](https://doi.org/10.1101/xxxxx) on acceptance).

Author contributions

S.R.S. designed the study, implemented the field architecture and the DeepXDE baseline, ran the experiments, 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] Lu, L., Jin, P., Pang, G., Zhang, Z. & Karniadakis, G.E. Learning nonlinear operators via DeepONet. *Nat. Mach. Intell.* **3**, 218–229 (2021).
- [4] Li, Z. et al. Fourier neural operator for parametric PDEs. In *ICLR* (2021).
- [5] Lu, L., Meng, X., Mao, Z. & Karniadakis, G.E. DeepXDE: a deep learning library for solving differential equations. *SIAM Rev.* **63**, 208–228 (2021).
- [6] Cuomo, S. et al. Scientific machine learning through PINNs. *J. Sci. Comput.* **92**, 88 (2022).
- [7] Wang, S., Teng, Y. & Perdikaris, P. Understanding and mitigating gradient flow pathologies in PINNs. *SIAM J. Sci. Comput.* **43**, A3055–A3081 (2021).
- [8] Vaswani, A. et al. Attention is all you need. In *NeurIPS 30*, 5998–6008 (2017).
- [9] Takamoto, M. et al. PDEBench: an extensive scientific machine learning benchmark suite. *arXiv:2210.07162* (2022).