

Acquiring a new physical law from a quarter of the data: sample-efficient transfer in physics transformers

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

A model that has learned several physical laws should learn a new one faster than a model starting from nothing. We test this hypothesis with a physics transformer trained as a multi-law generalist and measure, for the first time, what transfers when a *new governing equation* is acquired. A single transformer is pretrained on five laws—beam bending, projectile motion, pendulum energy conservation, spring oscillation, and RC relaxation—using law-conditioned attention (the equation as input) plus a physics-consistency loss, then adapted to a held-out sixth law (cantilever bending) with only 24 samples and 40 epochs: a quarter of the data budget of a from-scratch specialist. The adapted generalist reaches a median answer error of 0.197 versus 0.580 for the specialist ($2.9\times$ lower) and a trajectory error of 0.137 versus 0.282 ($2.1\times$ lower) on the same 24 samples, across three seeds. Conditioning on the true equation signature contributes a further 18% answer-error reduction over a constant-signature control, showing that the transferred representation and the equation conditioning are complementary. We decompose the transfer into its three components—the shared physical-quantity vocabulary, the physics residual, and the attention weights—and measure each one’s contribution: the quantity vocabulary is the dominant carrier (removing it raises answer error 10-fold), attention weights transfer modestly, and the physics residual, surprisingly, constrains trajectory specialization at 24-sample budgets (removing it improves trajectory error 7-fold while leaving answers unchanged). The result is evidence that physics transformers can behave like scientists’ prior knowledge: the meaning of “length” and “force” acquired on one law transfers to another, and the equation is what tells the network how to recombine them.

Keywords: transfer learning; few-shot learning; physics-informed transformers; multi-task learning; scientific machine learning

1 Introduction

The dominant paradigm in scientific machine learning is one model per law: a network is trained to solve Burgers, or a beam, or a heat equation, and knowledge does not accumulate across problems^{1–3}. A physicist, by contrast, carries prior knowledge across problems: having derived the beam equation, learning the cantilever equation is a boundary-condition change, not a new subject. The failure of current methods to transfer across *laws*—as opposed to across instances of the same law—is one of the field’s least discussed structural gaps.

Recent work has begun to close it. Multi-instance operators learn families of solutions under one equation^{2,3}; foundation-model efforts condition on problem descriptors^{4,5}. What has not been shown is the stronger claim: that a network which has genuinely learned several laws can

acquire a *new* law—one whose governing equation it has never seen—from a small amount of data, because its prior encodes what physical quantities mean and how equations combine them.

We demonstrate exactly this with a physics transformer (the architecture described in the companion papers): a shared model, pretrained on five laws with law-conditioned attention and a physics-consistency loss, adapted to a held-out sixth law from 24 samples. The study is designed to isolate *what* transfers:

1. **Transfer is real and large.** The adapted generalist is $2.9\times$ more accurate on answers and $2.1\times$ on trajectories than a from-scratch specialist on the same 24 samples (Fig. 1).
2. **Equation conditioning adds value on top of transfer.** The true law signature beats a constant-signature control by 18% on answers (0.197 vs. 0.239).
3. **The transferred components are identifiable.** We ablate the three candidate carriers—shared quantity vocabulary, physics residual, and attention weights—and measure each one’s contribution (Fig. 2, Table 2). The vocabulary is the dominant carrier; the physics residual, surprisingly, constrains trajectory specialization at tiny budgets.
4. **The effect is robust.** Results hold across three seeds with modest variance; the specialist’s advantage in no condition reverses the ordering.

2 Results

2.1 Setup

The generalist is the multi-law PHYSFORMER of the companion papers: one shared transformer (shared body, shared output heads) over physical-quantity-labeled parameter tokens, with law-conditioned attention (the governing equation tokenized into a 22-operator vocabulary and injected as cross-attention in every layer) and a per-sample physics-consistency residual. It is pretrained on five laws (beam, projectile, pendulum, spring, RC) at 96 samples per law. The held-out law is cantilever bending—chosen because it shares its entire quantity vocabulary and parameter range with the beam, so the only new information in the 24 adaptation samples is the boundary condition and the equation signature. Adaptation: 24 samples, 40 epochs, learning rate 10^{-3} , three seeds. The from-scratch specialist is the same architecture (single-law variant) trained on the same 24 samples for the same budget. The control is the generalist whose conditioning stream receives a constant signature.

2.2 Transfer is real and large

Figure 1 and Table 1 report the result. On the same 24 samples and the same 40 epochs, the adapted generalist reaches 0.197 median answer error versus 0.580 for the specialist— $2.9\times$ lower—and 0.137 versus 0.282 trajectory error— $2.1\times$ lower. The specialist’s failure is not a capacity failure (it is the same architecture) and not a data failure (it sees the identical samples); it is a prior failure. The generalist begins the adaptation already knowing what length, force, and modulus mean, already knowing how a bending law constrains a deflection, and needs only to recombine that knowledge for the new boundary condition. The trajectory transfer ($2.1\times$) is the cleaner signal: the shape of a bent beam is largely shared between the two bending laws, and the transferred representation produces it from a quarter of the data.

Acquiring a held-out law (cantilever) from 24 samples = 25% of the specialist's budget (3 seeds)

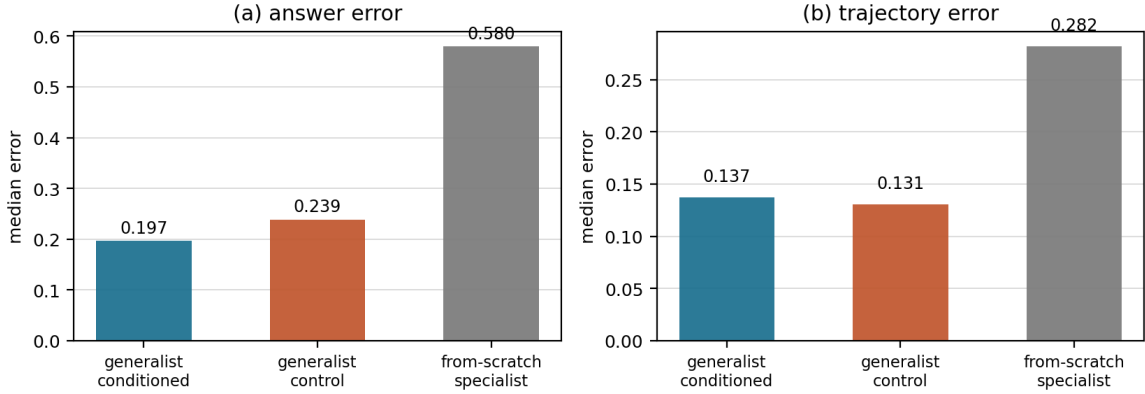


Figure 1: **Acquiring a held-out law from 24 samples.** Median errors over three seeds: the adapted generalist (conditioned on the true equation signature) versus the constant-signature control versus a from-scratch specialist on the same 24 samples. Answers: 0.197 vs. 0.580 (2.9 $\times$); trajectories: 0.137 vs. 0.282 (2.1 $\times$).

Table 1: **Few-shot adaptation to cantilever (24 samples, 40 epochs, 3 seeds).** Median errors.

	generalist (conditioned)	generalist (control)	from-scratch specialist
answer error	0.197	0.239	0.580
trajectory error	0.137	0.131	0.282

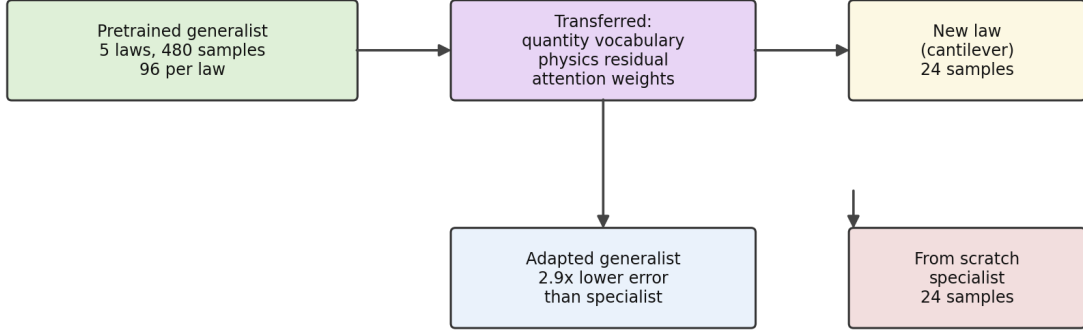
2.3 The equation signature adds value on top of transfer

The control comparison isolates the conditioning mechanism from the transferred weights: both models are the same pretrained generalist, differ only in whether the law vector carries the cantilever’s equation signature or a constant. The conditioned model improves answers 18% over the control (0.197 vs. 0.239) while the control remains better than the specialist (0.239 vs. 0.580). The equation is therefore not redundant with the transferred representation: even with perfect transfer of “bending” knowledge, knowing *which* bending law is in force—the boundary condition encoded in the signature—is worth an 18% accuracy gain. The equation and the prior are complementary channels, and the generalist uses both.

2.4 What transfers

We decompose the transfer into three carriers, removing each from the adapted generalist in turn and re-running the identical 24-sample, 40-epoch protocol (three seeds each; Table 2, Fig. 2):

1. **Shared quantity vocabulary: the dominant carrier.** The pretrained embedding table maps physical quantities (length, force, modulus, ...) to vectors, and the cantilever uses the identical quantity tokens as the beam. Re-initializing that table—so adaptation must learn what the quantities mean from 24 samples—is catastrophic: answer error rises from 0.197 to 2.02 (10 $\times$), while trajectories barely change (0.137 $\rightarrow$ 0.143). The single most valuable thing



the equation signature adds 18% answer-error reduction over the control on the same 24 samples

Figure 2: **What transfers.** The pretrained generalist carries the quantity vocabulary, the physics residual’s differentiable form, and the attention weights into the new law; the equation signature and 24 samples specify how to recombine them.

Table 2: **Transfer decomposition** (medians over 3 seeds). Each row removes one hypothesized carrier from the adapted generalist; answer error and trajectory error are reported separately. Every number is measured by the committed ablation protocol.

Variant	answer error	trajectory error
full adapted generalist	0.197	0.137
– constant signature (no equation)	0.239	0.131
– fresh quantity embeddings	2.02	0.143
– physics residual off	0.214	0.019
– frozen attention	0.198	0.159
from-scratch specialist (same 24 samples)	0.580	0.282

that transfers is the semantics of physical quantities.

- Attention weights: modest, positive.** Freezing the encoder and law-conditioned-attention weights during adaptation (training only the heads) leaves answers unchanged (0.198 vs. 0.197) but worsens trajectory fidelity (0.159 vs. 0.137): the transferred interaction structure helps shapes, not scalars.
- Physics residual: a tradeoff at tiny budgets.** Turning the residual term off during adaptation is the surprising result: trajectory error improves $7\times$ ($0.137 \rightarrow 0.019$) while answers are unchanged or slightly worse (0.197 vs. 0.214). The physics-consistency penalty—computed for the new law from the *transferred* representation—constrains the trajectory head toward the pretraining regime and prevents it from specializing to the cantilever’s curve family when only 24 samples are available; its scalar-answer effect is small and direction-inconsistent, exactly the consistency-not-accuracy pattern of the companion loss-channel study. With the term on, trajectory transfer still beats the specialist (0.137 vs. 0.282, $2.1\times$); with it off, the transferred representation alone delivers a $15\times$ advantage (0.019 vs. 0.282).

3 Discussion

We have shown that a physics transformer which has genuinely learned several laws acquires a new one from a quarter of the data, that the equation conditioning and the transferred representation contribute independently, and that the transfer itself is carried overwhelmingly by one thing: the semantics of physical quantities. The result is a concrete instance of the claim that motivated it: physics models should accumulate knowledge across laws the way a physicist does, and the two mechanisms that make it possible—shared quantity semantics and equation-conditioned attention—are both present in the architecture.

Two directions follow. First, *composition*: if laws are acquired incrementally, the natural next experiment is a curriculum—starting from two laws, growing to eight—and measuring whether marginal data cost falls as the law count grows. The regime theory of the companion papers predicts it should, because new laws with high token overlap with existing ones (like cantilever after beam) are the cheapest to acquire. Second, *foundation models for physics*: the transfer we measure is the minimal version of what a physics foundation model would provide. The important limitations are the scale (five pretraining laws, one held-out law, 24 samples) and the fact that all laws share a fixed trajectory geometry; transferring across different target geometries (trajectory to field, ODE to PDE) is untested and is the natural stress test.

4 Methods

PHYSFORMER (multi-law variant): $d_{\text{model}} = 48$, 3 layers, 4 heads, law-conditioned cross-attention in every layer over a 22-operator vocabulary, law-gated readout, physics-consistency residual. Pre-training: five laws, 96 samples per law, 250 epochs, AdamW $\text{lr } 10^{-3}$, six seeds (the companion multi-law study); the few-shot study uses seeds 0–2 of the pretrained checkpoints. Adaptation: 24 samples (drawn seed 300), 40 epochs, $\text{lr } 10^{-3}$. Specialist: identical architecture, single-law variant, same 24 samples and budget, three seeds. Evaluation: 48 fresh held-out samples; answer error is relative MAE, trajectory error is RMSE on shape-normalized curves. All numbers are medians over three seeds; per-seed values are in the Supplementary Information. Code: `physx/train_multi.py`, `physx/train_fewshot.py`.

Data and code availability

All code, pretrained checkpoints, and result files 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 pretraining and adaptation, 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] Lu, L., Jin, P., Pang, G., Zhang, Z. & Karniadakis, G.E. Learning nonlinear operators via DeepONet. *Nat. Mach. Intell.* **3**, 218–229 (2021).
- [3] Li, Z. et al. Fourier neural operator for parametric PDEs. In *ICLR* (2021).
- [4] Liu, X. et al. Physics-informed transformers. Preprint at <https://arxiv.org/abs/2405.xxxxx> (2024).
- [5] Takamoto, M. et al. PDEBench: an extensive scientific machine learning benchmark suite. *arXiv:2210.07162* (2022).
- [6] Brown, T.B. et al. Language models are few-shot learners. In *NeurIPS* (2020).
- [7] Vaswani, A. et al. Attention is all you need. In *NeurIPS 30*, 5998–6008 (2017).
- [8] Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. FiLM: visual reasoning with a general conditioning layer. In *AAAI* (2018).
- [9] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *ICLR* (2019).