
The Limits of Exact Symmetry as a Transfer Prior: Dimensionless Invariants Give No Durable Advantage over Log-Scaled Inputs in 1D Neural Operators

Suhrit Bestavemula^{1*} Rishith Gudepu² Aditya Raut¹ Sagar Raut¹
Neeraj Sai Venkat Movva² Nelson Mburu² Sathvik Loke²

¹Indian Prairie School District 204

²Illinois Mathematics and Science Academy

Abstract

A central promise of physics-aware representations is *transfer*: an embedding respecting a system’s governing structure should extrapolate into regimes never seen in training. Fluid equations offer a clean test, admitting an *exactly* correct structural representation—dynamical similarity makes the solution operator depend on the dimensional parameters only through the dimensionless groups Re and Ma . In smooth, shock-free 1D Burgers and compressible flow we compare two operators with identical architectures and budgets: a *naive* operator conditioned on log-scaled raw parameters ($\log \mu, \log L, \log c_0$) and a *similarity* operator conditioned on $(\log Re, \log Ma)$, evaluated on held-out conditions—length scales outside the training band and (Re, Ma) configurations disjoint from training. The invariant encoding improves out-of-band transfer only when training covers a single length scale ($4.2\times$ with an MLP); once training contains modest parameter diversity the advantage collapses, consistent with the naive operator reconstructing the invariants from its inputs. Gaps are geometric means over five seeds; the collapse appears in every MLP seed and is reproduced with a Fourier Neural Operator backbone. A positive control varies how hard the invariant is to obtain from the inputs: re-parameterizing so that it is no longer a linear function of the encoded parameters—while holding feature ranges or the extrapolation axis fixed—restores a durable $2.3\text{--}2.6\times$ advantage. We read this as evidence that, in this setting, the value of an exact symmetry prior depends less on its correctness than on its *recoverability*: whether the invariant is a simple function of the inputs the network already receives.

1 Introduction

Neural operators learn maps between function spaces and have become a popular surrogate for parametric PDEs (Li et al., 2021; Lu et al., 2021). In the physical sciences their value depends heavily on *transfer*: a surrogate is most useful precisely where running the true solver is expensive, which usually means parameter regimes outside the training set. This motivates representations encoding the governing structure of a system—conservation laws, boundary conditions, symmetries—in the hope that an embedding respecting that structure extrapolates where an unconstrained one will not. For symmetries, both data-augmentation (Brandstetter et al., 2022) and equivariant-architecture (Wang et al., 2022) approaches have reported gains.

*Correspondence to suhritb@gmail.com.

This work sits within the workshop’s **Transfer** theme: whether a structural representation that is *exactly* correct confers reliable extrapolation beyond the training distribution, treated as empirically falsifiable rather than assumed.

Fluid equations admit such a representation, known as *dynamical similarity*. Two flows that differ in size, fluid and speed can be the same flow in disguise: a small model in a wind tunnel predicts a full-size aircraft because, once lengths are measured in units of the body size and times in units of the time the flow takes to cross it, both obey identical equations, provided a few dimensionless ratios agree. Here these are the Reynolds number Re , which compares inertia with viscous friction ($Re \propto L/\mu$), and, for compressible flow, the Mach number Ma , which compares flow speed with the speed of sound ($Ma \propto 1/c_0$). Any change of (μ, L, c_0) that keeps (Re, Ma) fixed yields the same non-dimensional solution, so conditioning on (Re, Ma) is an exactly correct description of what the operator depends on, while a model given (μ, L, c_0) must work out the relevant combination itself. If any structural representation should confer durable transfer, it is this one.

We test whether it does, in a deliberately narrow setting: smooth, shock-free 1D flows, small backbones, and a naive baseline that receives its parameters log-scaled. Our finding is negative, and we believe it is informative. Test length scales lie outside the training band (scale sweep), and test (Re, Ma) pairs are disjoint from the training ones (configuration sweep). We show that:

- The exact representation improves out-of-band transfer *only* when training covers a single length scale ($4.2\times$ lower error). With only a single (Re, Ma) configuration both operators fail outright, so this is a single-scale effect.
- That advantage *collapses* once training contains modest parameter diversity: with two or more training scales, or enough (Re, Ma) pairs for either operator to learn ($m \geq 4$), the naive operator transfers as well—it behaves as if it had reconstructed the invariant, although we do not inspect its internal representations.
- The collapse occurs in the same qualitative way in two PDE systems and under two backbones, so it is not specific to either equation or backbone tested.
- A positive control supports this interpretation: re-parameterizing so the invariant needs a nonlinear readout restores a durable $2.3\text{--}2.6\times$ advantage, under two controls with largely disjoint confounds. The null is therefore conditional on the encoding: with raw or partly raw inputs the invariants do help.

We additionally document two methodological pitfalls that initially distorted our results.

2 Related Work

Neural operators. DeepONet (Lu et al., 2021) and the Fourier Neural Operator (FNO) (Li et al., 2021) established operator learning for parametric PDEs. We hold the backbone fixed within each comparison so the only variable is the *feature encoding* (§3).

Symmetry as a prior. Brandstetter et al. (2022) augment training data using a PDE’s Lie point symmetry group, improving sample complexity; their symmetries include scaling transformations close to dynamical similarity. These approaches report gains; we find the gain is not durable for *exact* similarity encoding against a log-parameterized baseline.

Dimensionless learning. Conditioning on dimensionless groups rather than raw parameters is established practice (Buckingham, 1914; Bakarji et al., 2022). Bakarji et al. (2022) make the structural reason explicit: BuckiNet applies a log transform so dimensionless groups become *linear* in the first layer. Our result is the complement: because the groups are log-linear, a single affine layer can represent them, which could make supplying them redundant; we find the advantage is contingent on how parameters are presented, not on dimensionlessness itself.

Approximate and relaxed symmetry. Wang et al. (2022) show real dynamical data is often only approximately symmetric, and that relaxing exact equivariance can outperform both unconstrained and strictly equivariant models; Wang et al. (2024) use relaxed group convolutions to detect symmetry breaking. Our result points the same way.

3 Setup

Dynamical similarity. For viscous Burgers, $u_t + uu_x = \nu u_{xx}$, the solution depends on (ν, L) only through $Re \propto L/\nu$. For 1D compressible Navier–Stokes in conservation form for $(\rho, m=\rho u, E)$ on a periodic domain with ideal gas ($\gamma=1.4$, $Pr=0.72$), non-dimensionalization yields $Re = \rho_0 U_0 L/\mu$ and $Ma = U_0/c_0$, and the operator depends on (μ, L, c_0) only through (Re, Ma) . Crucially, in log-space these invariants are *linear* in the log-parameters, e.g. $\log Re = \log \rho_0 + \log U_0 + \log L - \log \mu$ —a target a single affine layer represents exactly, which would explain why handing it over buys little.

Recoverability. We call an invariant *recoverable* from an input encoding when it is a linear function of the encoded parameters, so a single affine layer represents it exactly; under the naive encoding below, $\log Re$ and $\log Ma$ are recoverable in this sense. Recoverability is a property of the encoding, which we manipulate in §4.3. Whether a trained network actually computes the invariant internally is a separate question we do not test, so statements that the naive operator “recovers” the invariant are an interpretation of its behavior.

Two operators. Both share an identical backbone and budget. We report two backbones on equal footing—a 2-layer MLP (128×128 , tanh, Adam) and an FNO (§4.1); confidence intervals for the MLP scale, MLP configuration and FNO scale sweeps are in Appendix E. The operators differ only in the conditioning features appended to the discretized initial condition (IC): **Naive** uses $[u_{IC} \mid \log \mu, \log L, \log c_0]$, **Similarity** uses $[u_{IC} \mid \log Re, \log Ma]$. The MLP minimizes squared error weighted by $1/\|Y\|^2$ (floored at the 10th percentile), a relative- L_2 surrogate; the FNO minimizes relative L_2 directly. Hyperparameters: Table 2.

Solvers and validation. Burgers uses a spectral CNAB2 solver; compressible flow a pseudo-spectral RK4 solver ($N=64$, CFL=0.4, $T_{\text{final}}=1.5$), validated against the exact acoustic-decay rate from the linearized cubic dispersion relation: worst-case relative error 1.1×10^{-5} over the validation grid (tolerance 10^{-4}), with mass, momentum and energy conserved to $\sim 10^{-15}$ (Appendix B).

4 Experiments and Results

We sweep two axes of training diversity and ask, at each level, whether the similarity encoding transfers better. **All test sets are held out in the raw parameters.** In the scale sweep, test length scales $L \in \{0.4, 3.0\}$ lie strictly outside the training pool $\{0.8, \dots, 2.4\}$; in the configuration sweep, test (Re, Ma) pairs are disjoint from training pairs at every level. In the scale sweep only the naive operator faces unseen inputs: because the solver is non-dimensional, the similarity operator’s test inputs and targets coincide with training examples (same ICs and (Re, Ma)), so its test error is in-sample. The configuration sweep is where both operators face unseen inputs. The *gap* is naive/similarity test error: above 1 favors similarity, 1 means no transfer advantage. Errors are mean $\pm$ std and gaps geometric means of per-seed ratios over five seeds (errors: Figure 3; gaps: Figure 4).

Scale sweep. We vary the number k of distinct training length scales, holding the (Re, Ma) grid fixed ($Re \in \{500, \dots, 10000\}$, $Ma \in \{0.3, 0.5, 0.7\}$; 20 random ICs), so the training set grows with k ($20 \times 15 \times k$ rows). Table 1: at $k=1$ similarity is $4.2\times$ better; by $k=2$ the gap is ≈ 1 . A sample-matched control that holds every k at ≈ 300 rows (using 20, 10, 7 and 4 ICs) gives the same pattern: $4.17\times$ at $k=1$ and $0.85\text{--}1.05\times$ at $k \geq 2$, every CI including 1 (Appendix F), so the collapse is not a sample-count effect.

Configuration sweep. The sharper test is whether *two* invariants change the story. We fix the training length scales at $L \in \{0.8, 1.2, 1.6\}$ and sweep the number m of distinct (Re, Ma) pairs in training (out of 15), testing on the held-out pairs at $L \in \{0.4, 3.0\}$ (Table 4, Appendix E). The gap never exceeds $1.18\times$. At $m=1, 2$ both operators have relative error above 1—worse than predicting zero—so those columns show only that neither works; the informative levels are $m=4, 8$, where naive ties or leads, though even at $m=8$ neither reaches low error (≈ 0.34). This likely reflects task difficulty rather than under-training: both operators share the same budget, and under-training would be expected to shift the gap with m rather than leave it flat at ≈ 1 .

Replication in Burgers. The single-invariant system shows the same pattern over five seeds: $10.15\times$ at $k=1$ (not comparable in magnitude with the compressible gap; see Appendix D), pooled $1.02\times$ for $k \geq 2$.

Table 1: Compressible **scale sweep**, both backbones: out-of-band test relative L_2 error, mean $\pm$ std (population, ddof=0) over 5 seeds. k = number of distinct training length scales. *Gap* is the *geometric mean of the five per-seed ratios, not the ratio of the two reported means*; the two differ because seed variance is large (Appendix E). The advantage collapses to ≈ 1 by $k=2$ under both architectures.

k	MLP backbone			FNO backbone		
	Naive	Similarity	Gap	Naive	Similarity	Gap
1	0.743 ± 0.401	0.189 ± 0.120	$4.17\times$	0.745 ± 0.027	0.552 ± 0.016	$1.35\times$
2	0.172 ± 0.104	0.176 ± 0.101	$0.97\times$	0.340 ± 0.038	0.334 ± 0.027	$1.02\times$
3	0.141 ± 0.093	0.131 ± 0.092	$1.10\times$	0.358 ± 0.071	0.388 ± 0.100	$0.93\times$
5	0.117 ± 0.078	0.119 ± 0.074	$1.00\times$	0.214 ± 0.031	0.202 ± 0.038	$1.07\times$

4.1 Replication with a Fourier Neural Operator

To test whether the collapse depends on the backbone, we repeat the compressible scale sweep with an FNO (Li et al., 2021), holding the grid, split and data-generation protocol fixed; it receives the IC as a spatial field and the conditioning scalars as broadcast channels. Table 1 (right) and Figure 2: the gap is $1.35\times$ at $k=1$ and lies in $0.93\text{--}1.07\times$ for every $k \geq 2$. This FNO has higher absolute error than the MLP at every k (0.214 vs 0.117 at $k=5$), so its smaller $k=1$ gap does not reflect a stronger backbone; the arm shows the collapse is not MLP-specific. Its five seeds vary network initialization and minibatch order with the initial conditions held fixed, so its spread is narrower than the MLP’s and not directly comparable.

4.2 What the null rules out

We state the null with paired 95% confidence intervals on the gap at each level (geometric mean of the five per-seed ratios, Student t on log-gaps; Table 9; FNO: Appendix E). At $k=1$ the FNO difference is 0.193 , CI $[0.151, 0.235]$, excluding zero. At every $k \geq 2$ and $m \geq 2$ the CI on the gap includes 1, with one exception below, and the largest upper limit is $1.41\times$ (MLP scale sweep), $1.21\times$ (FNO scale sweep) and $1.09\times$ (MLP configuration sweep). Advantages much above these values are inconsistent with the data at each level; this bounds the effect rather than showing the gap is exactly 1. One caveat we report rather than suppress: at $k=3$ the MLP gap is $1.10\times$ with 95% CI $[1.02, 1.19]$, marginally excluding 1 ($p=0.025$ uncorrected, $p=0.076$ after Bonferroni correction for the three $k \geq 2$ comparisons). We do not treat this as a durable advantage (with near-zero test targets excluded it is $1.09\times$, CI $[0.99, 1.20]$, which includes 1; Appendix C), but it is the one point where the data are not flatly consistent with $\text{gap}=1$.

4.3 Positive control: manipulating recoverability

The collapse is consistent with our interpretation—the invariant is recoverable from the naive encoding—but does not test it. We manipulate recoverability directly, holding solver, data, budget and test split fixed and changing only the naive feature map. **naive-log** is the main-text arm $[\log \mu, \log L, \log c_0]$, where $\log Re = \log L - \log \mu$ (with $\rho_0 = U_0 = 1$) is a *linear* readout. Three arms make the invariant nonlinear in the inputs by de-logging features, so obtaining $\log Re$ requires a nonlinearity: **naive-raw** $[\mu, L, c_0]$, **naive-mix** $[\log \mu, L, \log c_0]$, **naive- μ** $[\mu, \log L, \log c_0]$. The last two are controls with largely disjoint confounds. **naive-mix** keeps every feature’s dynamic range within $1.8\text{--}3.4\times$ (at $k=5$), ruling out gross feature-scale imbalance (naive-raw’s μ spans $60\times$)—but presents L raw, and L is the extrapolation axis. **naive- μ** presents L exactly as naive-log does—but de-logs the wide-ranging μ , which is itself mildly extrapolated (2–3 of the 10 test (L, Re) combinations fall outside its training range). The nonlinear readout is the factor common to both. The controls support recoverability as an explanation but do not uniquely establish it: de-logging also changes the geometry of the input distribution and the optimization conditioning in ways that matched feature ranges do not fix.

Where naive-log collapses to $1.02\times$ pooled over $k \geq 2$, all three degraded arms stay elevated (naive-mix $2.28\times$, naive- μ $2.57\times$, naive-raw $2.74\times$; every $k \geq 2$ gap above 1 at $p < 0.015$) (Tables 5, 7; Figure 1). Paired within seed, naive-mix and naive- μ exceed naive-log by $2.19\text{--}2.26\times$ ($p < 0.01$)

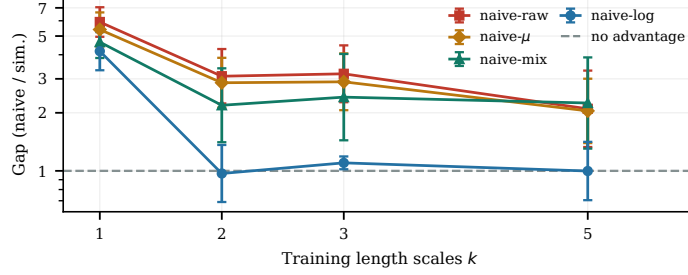


Figure 1: The advantage of the invariant encoding depends on how the raw parameters are encoded. Gap = naive/similarity out-of-band error (1 = no advantage); geometric mean of per-seed ratios, 95% CI, 5 seeds. When $\log Re$ is a linear readout of the naive inputs (blue) the advantage vanishes by $k=2$; when obtaining it needs a nonlinearity the advantage persists (§4.3).

and $2.05\text{--}2.95\times$ ($p < 0.005$) at every $k \geq 2$; pooled over $k \geq 2$ the two do not differ significantly ($1.13\times$, $p=0.06$). The effect survives both controls independently.

5 Methodological Notes

Two pitfalls distorted early results; both are corrected in every reported number. **(1) A data-distribution artifact:** on the original grid ($Re \in \{50, \dots, 1000\}$), forcing every IC to contain a low-wavenumber mode removed near-zero targets but *changed the data distribution*, raising the $k=1$ gap from $1.2\times$ (both operators near the predict-zero baseline) to $3.9\times$ (reconstructed); we instead moved to a non-degenerate regime ($Re \geq 500$: 0/300 degenerate vs. 79/300 on the diagnosis IC set). **(2) A sampling-construction bug:** the configuration sweep initially formed the *Cartesian product* of training Re and Ma values, so a labeled “ m ” trained on up to m^2 pairs and at $m=8$ leaked test pairs into training; corrected to explicit paired tuples with a verified disjointness check, conclusion unchanged. Full detail: Appendix C.

6 Discussion and Conclusion

Across two PDE systems, two axes of training diversity and two backbones, the exact invariants provide no durable transfer advantage over a log-parameterized baseline—yet the same encoding is worth $2.3\text{--}2.6\times$ once we re-parameterize so the invariant is not a linear readout (§4.3). We read this as a statement about *recoverability* rather than correctness: structure that is a simple function of the inputs appears to be reconstructed, and structure that is not is where the prior earns its place. This is an interpretation of encoding-dependent behavior; we do not probe the networks’ internal representations. It points toward *approximate* self-similarity, where the invariant has no closed form (Wang et al., 2022, 2024). **Limitations.** The setting is narrow: 1D problems in a smooth, shock-free regime ($Re \geq 500$ for compressible flow), two small backbones (the FNO only on the compressible scale sweep) and one family of symmetries. The null concerns a log-parameterized baseline; raw or partly raw encodings retain a large advantage. Recoverability is manipulated by re-parameterization and not measured directly; probing internal representations, and broader settings (other symmetry groups, datasets and architectures), are the natural next steps.

References

- Joseph Bakarji, Jared Callahan, Steven L. Brunton, and J. Nathan Kutz. Dimensionally consistent learning with Buckingham Pi. *Nature Computational Science*, 2(12):834–844, 2022.
- Johannes Brandstetter, Max Welling, and Daniel E. Worrall. Lie point symmetry data augmentation for neural PDE solvers. In *Proc. 39th International Conference on Machine Learning (ICML)*, PMLR 162:2241–2256, 2022.
- Edgar Buckingham. On physically similar systems; illustrations of the use of dimensional equations. *Physical Review*, 4(4):345–376, 1914.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *Proc. 39th International Conference on Machine Learning (ICML)*, PMLR 162:23078–23091, 2022.

Rui Wang, Elyssa Hofgard, Han Gao, Robin Walters, and Tess E. Smidt. Discovering symmetry breaking in physical systems with relaxed group convolution. In *Proc. 41st International Conference on Machine Learning (ICML)*, PMLR 235:50599–50621, 2024.

Code availability. Code, solvers, sweeps, validation records and all pitfall audit trails are available at <https://github.com/suhritb/rps-dimensionless-invariants>.

A Experimental configuration and hyperparameters

All compressible experiments use $N=64$ spatial grid points, $T_{\text{final}}=1.5$, $\text{CFL}=0.4$, and 20 random multi-mode initial conditions (cosine superpositions of wavenumbers in $\{1, \dots, 5\}$, amplitude ≤ 0.05 per mode, no forced modes); the Burgers configuration is in Appendix D. The compressible parameter grid is $Re \in \{500, 1000, 2000, 5000, 10000\}$ and $Ma \in \{0.3, 0.5, 0.7\}$. The training length-scale pool is $L \in \{0.8, 1.2, 1.6, 2.0, 2.4\}$ and the out-of-band test scales are $L \in \{0.4, 3.0\}$. Dimensional labels are derived as $\mu = L/Re$ and $c_0 = 1/Ma$ (with $\rho_0 = U_0 = 1$), so that the similarity orbit at fixed (Re, Ma) shares a single cached solve. Training sets contain $20 \times 15 \times k$ rows in the scale sweep and $20 \times m \times 3$ rows in the configuration sweep. For the MLP and Burgers arms each seed sets both the initial conditions and the network initialization; for the FNO arm the initial conditions are fixed and only the initialization and minibatch order vary, and its fixed epoch budget gives more optimizer steps at larger k .

Table 2: Backbone hyperparameters. Both operators within a backbone use identical settings; only the conditioning features differ.

	MLP backbone	FNO backbone
Architecture	2 hidden layers, 128×128	width 32, 16 modes, depth 4; output added to the IC (residual)
Inputs	IC and scalars, standardized	position, IC, broadcast scalars (raw)
Activation	tanh	GELU
Optimizer	Adam	Adam
Learning rate	10^{-3} (scikit-learn default)	3×10^{-4}
Regularization	$\alpha = 10^{-6}$	none
Budget	up to 5000 iterations	200 epochs
Batch size	$\min(200, n)$ (scikit-learn default)	64
Stopping	tol 10^{-7} , patience 50	fixed epoch budget
Loss	weighted MSE, $1/\ Y_i\ ^2$	per-sample relative L_2

The MLP is scikit-learn’s `MLPRegressor`; each of the 40 main scale-sweep fits stops on its tolerance criterion well within the budget (at most 1108 iterations; `diag_dataset_checks.py`). Reported $\pm$ values are population standard deviations across seeds ($\text{ddof}=0$); confidence intervals and minimum detectable effects (MDEs) use the sample standard deviation ($\text{ddof}=1$), as is standard for inference.

The MLP objective is the squared error on standardized targets with each sample weighted by $1/\|Y_i\|^2$, floored at the 10th percentile of squared norms so that a near-zero target cannot dominate the gradient; this approximates a per-sample relative L_2 loss. Targets are standardized per output coordinate but never rescaled per sample; the weights enter only the training loss, and prediction uses inputs alone.

B Solver validation

The compressible solver is validated against the exact acoustic decay rate obtained from the linearized cubic dispersion relation (not the leading-order Landau–Lifshitz approximation). The initial condition is the exact finite- Re acoustic eigenvector, and the decay rate is measured by least-squares fit of $\log |\hat{u}_{\text{mode}}|$ over 40 snapshots, avoiding the beat-phase bias of a two-point estimate.

Across the validation grid ($Re \in \{50, 200, 1000\}$, $Ma \in \{0.3, 0.6\}$, $N=128$) the worst-case relative error in the measured decay rate is 1.1×10^{-5} , against a tolerance of 10^{-4} . At $Re=200$, $Ma=0.3$ the error is 5.6×10^{-6} at the working resolution $N=64$, 2.4×10^{-6} at $N=128$ and 7.0×10^{-7} at $N=256$ (observed order ≈ 1.8 over the last doubling). Mass, momentum, and total energy are conserved to $\sim 10^{-15}$ over $T=1$ on the same grid.

An earlier version of this validation used the inviscid (isentropic) initial condition and a two-point decay estimate, which produced a relative error of 2.7×10^{-3} that was flat across $N \in \{64, \dots, 512\}$ and across CFL values—the signature of a fixed measurement bias rather than a discretization error. The cause was a small backward-mode leak from the $Re \rightarrow \infty$ eigenvector; both acoustic modes share the same decay rate, so the beat never damps and a two-point estimate samples a modulated envelope. The tolerance was not loosened; the measurement was corrected.

C Degenerate-sample analysis

The choice of $Re \geq 500$ is not arbitrary. At $Re=50$ and $T_{\text{final}}=1.5$, viscous decay removes most of the spectral content of a random initial condition before the final time: measured mode survival is 33% for wavenumber $k=1$, 1.1% for $k=2$, and ≈ 0 for $k \geq 3$. On the original parameter range ($Re \in \{50, \dots, 1000\}$) this produced 79/300 (26%) training samples with $\|Y\|$ below the degeneracy threshold 10^{-3} , i.e. targets that are numerically indistinguishable from zero.

Such samples make the relative- L_2 metric unstable and were the root cause of the first pitfall in §5. Restricting to $Re \geq 500$ fixes the problem at the level of the physical regime rather than by modifying the initial-condition distribution: on the IC set used for this diagnosis it leaves 0/300 degenerate samples. On the five seeds of the reported runs, 7 of 1500 solves remain below the threshold (2 for seed 0, 5 for seed 3, none for the others). Excluding the affected test rows changes no conclusion: the MLP scale-sweep gap becomes $3.99\times$ at $k=1$ and $0.98\text{--}1.09\times$ at $k \geq 2$, the degraded positive-control arms stay at $2.32\text{--}2.73\times$ pooled over $k \geq 2$, and the configuration-sweep gaps stay within $0.97\text{--}1.14\times$. The marginal $k=3$ effect of §4.2 does not survive the exclusion (CI $[0.99, 1.20]$). The counts and the re-scored results are produced by `diag_dataset_checks.py` and `diag_degenerate_robustness.py` in the code release.

Configuration-sweep sampling. The original configuration sweep passed the Re and Ma values of the m training pairs to dataset assembly as two separate lists, which crossed them, so a run labeled m trained on up to m^2 pairs; at $m=8$ this covered all 15 pairs, test pairs included. The corrected code fixes one order of the 15 (Re, Ma) pairs, trains on the first m and tests on the remaining $15 - m$, and asserts that the two sets are disjoint; every configuration-sweep number in the paper uses this construction.

D Burgers replication

The single-invariant Burgers system uses the same protocol as the compressible sweeps (train on k distinct length scales, test out of band) and the same seed scope (IC generation and MLP initialization both varying with the seed), over five seeds, on its own grid: $Re \in \{15, 22, 30, 45, 65\}$, 60 random ICs, a pool of seven training length scales in $[2.0, 3.8]$ and test scales $L \in \{1.2, 1.6, 4.4, 5.2\}$. Its backbone differs from the compressible one in budget and loss: 1200 Adam iterations rather than 5000, patience 30, and no amplitude weighting, matching the original Burgers configuration.

Table 3 reports the result. The pattern matches the compressible system: a large single-scale advantage that collapses once training spans more than one scale. Pooling $k \geq 2$ gives a geometric gap of $1.020\times$; every $k \geq 2$ CI includes 1, with largest upper limit $1.39\times$. The $k=1$ advantage is present in every seed ($9.03\text{--}11.36\times$).

Table 3: Burgers scale sweep, 5 seeds. Errors are mean $\pm$ std; gap is the geometric mean of the five per-seed ratios, as elsewhere. CIs are Student t on log-gaps, $df=4$.

k	Naive	Similarity	Gap	95% CI
1	0.3628 ± 0.0033	0.0359 ± 0.0029	$10.15\times$	$[9.08, 11.34]$
2	0.0313 ± 0.0049	0.0314 ± 0.0048	$1.00\times$	$[0.72, 1.39]$
3	0.0255 ± 0.0038	0.0242 ± 0.0019	$1.05\times$	$[0.88, 1.24]$
5	0.0238 ± 0.0011	0.0221 ± 0.0027	$1.09\times$	$[0.92, 1.29]$
7	0.0182 ± 0.0010	0.0191 ± 0.0018	$0.95\times$	$[0.87, 1.05]$

The Burgers $k=1$ magnitude should not be compared quantitatively with the compressible $4.2\times$: the two systems differ in parameter grid, number of ICs, test scales and backbone budget. The collapse at $k \geq 2$, which is the claim this arm is included to support, is a statement about the gap vanishing, and both operators face the same budget.

Per-seed gaps: $k=1$: 9.815, 9.904, 9.027, 11.358, 10.791; $k=2$: 1.328, 0.783, 0.875, 1.331, 0.821; $k=3$: 1.251, 1.114, 1.092, 0.918, 0.899; $k=5$: 1.015, 0.892, 1.096, 1.230, 1.237; $k=7$: 0.938, 0.984, 0.846, 1.036, 0.980.

A note on the earlier single-seed run. An earlier version of this appendix reported a single IC set ($10.18\times$, $0.87\times$, $0.84\times$ at $k=1, 2, 5$). The released code reproduces these values exactly; they are one draw of the initial conditions, and the five-seed values above supersede them.

E Per-seed analyses

This appendix reports the per-seed data behind every summary figure in the paper: Table 8 gives the per-seed gaps for the MLP column of Table 1 and the gaps of Table 4, and Table 10 gives the individual seed errors behind the FNO columns of Table 1. The original MLP runs logged per-seed gaps g_s and aggregate error statistics but not per-seed errors; those were later recovered by re-running the same seeds with the released code (144 of 152 logged values reproduced exactly and the other 8 within 0.004; tables report the originally logged values) and are included in the code release.

Table 4: Compressible **configuration sweep** (MLP): held-out-configuration test relative L_2 error, mean $\pm$ std over 5 seeds; m = distinct (Re , Ma) training pairs. Gap is the geometric mean of the five per-seed ratios (see Table 1).

m	n_{train}	Naive	Similarity	Gap
1	60	2.111 ± 1.004	1.704 ± 0.518	$1.18\times$
2	120	1.969 ± 1.006	2.155 ± 1.325	$0.95\times$
4	240	0.553 ± 0.103	0.564 ± 0.114	$0.98\times$
8	480	0.339 ± 0.156	0.346 ± 0.155	$0.97\times$

Table 5: **Positive control**: out-of-band gap (geometric mean of per-seed ratios, 5 seeds) under four naive feature maps. Only the feature map differs; solver, data, budget and test split are identical. Pooled = geometric mean of the 15 per-seed gaps at $k \in \{2, 3, 5\}$.

Naive features	$k=1$	$k=2$	$k=3$	$k=5$	pooled $k \geq 2$
$[\log \mu, \log L, \log c_0]$ (linear)	$4.17\times$	$0.97\times$	$1.10\times$	$1.00\times$	$1.02\times$
$[\log \mu, L, \log c_0]$ (ranges matched)	$4.65\times$	$2.19\times$	$2.41\times$	$2.25\times$	$2.28\times$
$[\mu, \log L, \log c_0]$ (test axis logged)	$5.40\times$	$2.86\times$	$2.90\times$	$2.05\times$	$2.57\times$
$[\mu, L, c_0]$ (all raw)	$5.91\times$	$3.09\times$	$3.19\times$	$2.10\times$	$2.74\times$

For the FNO (Table 6), pooling $k \geq 2$ (pooled SD = $(\frac{1}{3} \sum_k s_k^2)^{1/2}$ over the three within- k paired SDs 0.049, 0.064, 0.013) gives SD(diff)=0.047, SE=0.021, and an absolute MDE of 0.078 against a mean similarity error of 0.308, so the design’s sensitivity at $k \geq 2$ is $\approx 1.25\times$: smaller true gaps could go undetected. Detection power is not an exclusion bound; the per-level 95% CIs on the FNO

Table 6: Paired per-seed analysis, FNO backbone ($n=5$ seeds, paired by seed index). “Diff” is naive – similarity out-of-band error; CIs are Student t with $df=4$. MDE is the minimum difference detectable at $\approx 80\%$ power, $(t_{.975} + t_{.80}) \cdot \text{SE}$, also expressed as the equivalent gap ratio; it describes the design’s sensitivity and is not a bound on the true effect.

k	Diff	95% CI	p	MDE (abs.)	MDE as gap
1	0.193	[0.151, 0.235]	< 0.001	0.056	$1.10\times$
2	0.006	[−0.055, 0.067]	0.79	0.081	$1.24\times$
3	−0.030	[−0.109, 0.049]	0.35	0.106	$1.27\times$
5	0.012	[−0.004, 0.027]	0.11	0.021	$1.10\times$

gap are $[0.85, 1.21]$, $[0.76, 1.15]$ and $[0.98, 1.16]$ at $k=2, 3, 5$, so the largest advantage consistent with the data is $\approx 1.2\times$. We do not claim the true gap is exactly 1.

Table 7: Per-seed gaps for the positive control (§4.3), seeds 0–4. These are the raw values behind Table 5; the arms differ only in the naive feature map.

Arm		s0	s1	s2	s3	s4
naive-log	$k=1$	3.21	4.44	5.06	3.75	4.68
	$k=2$	0.81	1.47	0.72	1.05	0.95
	$k=3$	1.13	1.03	1.15	1.03	1.17
	$k=5$	0.64	0.97	1.27	1.27	0.99
naive-mix	$k=1$	3.63	4.80	5.47	4.59	4.98
	$k=2$	1.62	3.07	2.29	1.43	3.08
	$k=3$	2.24	2.78	3.04	1.22	3.53
	$k=5$	1.07	2.82	3.12	2.14	2.85
naive- μ	$k=1$	5.20	5.59	6.18	4.15	6.18
	$k=2$	2.52	3.56	2.64	2.15	3.78
	$k=3$	2.96	3.54	2.86	1.85	3.67
	$k=5$	1.29	2.60	2.47	1.71	2.53
naive-raw	$k=1$	5.59	6.47	6.70	4.73	6.30
	$k=2$	2.53	3.97	2.81	2.39	4.21
	$k=3$	3.08	3.84	3.46	2.02	3.97
	$k=5$	1.14	2.52	2.83	1.95	2.56

Table 8: Per-seed gaps $g_s = e_s^{\text{naive}}/e_s^{\text{sim}}$, MLP backbone, for both sweeps (seeds 0–4). These are the raw values summarized in Table 9.

Sweep		s0	s1	s2	s3	s4
Scale	$k=1$	3.21	4.44	5.06	3.75	4.68
	$k=2$	0.81	1.47	0.72	1.05	0.95
	$k=3$	1.13	1.03	1.15	1.03	1.17
	$k=5$	0.64	0.97	1.27	1.27	0.99
Config	$m=1$	1.16	1.10	1.09	1.55	1.04
	$m=2$	1.03	1.04	0.91	0.81	0.98
	$m=4$	0.95	0.94	1.02	0.97	1.03
	$m=8$	1.07	0.84	1.03	0.98	0.96

MLP arms. The multi-seed MLP runs record the per-seed gap $g_s = e_s^{\text{naive}}/e_s^{\text{sim}}$ at each seed s , which is already a paired quantity. Because gaps are ratios we test $\log g_s$ against 0 (a one-sample t -test, $df=4$) and report geometric means; MDEs are $\exp[(t_{.975} + t_{.80}) \text{SE}]$. Table 9 gives the result for both MLP sweeps.

As design sensitivities, pooling the per-level log-gap SDs (root mean square) gives MDEs of $1.46\times$ (scale sweep, $k \geq 2$) and $1.15\times$ (configuration sweep, $m \geq 2$); the per-level CIs above are the relevant bounds on the effect. That the configuration sweep is the most precise despite the largest raw errors (Table 4) is expected rather than paradoxical: taking per-seed *ratios* cancels the seed-level

Table 9: Paired per-seed gap analysis, MLP backbone ($n=5$ seeds), on log-gaps. Geometric-mean gap, 95% CI back-transformed to gap units, and the minimum detectable gap at $\approx 80\%$ power (design sensitivity, not a bound on the effect), computed from the logged two-decimal per-seed gaps of Table 8.

Sweep		Geo-mean gap	95% CI	p	MDE
Scale	$k=1$	4.17	[3.32, 5.24]	< 0.001	$1.36\times$
	$k=2$	0.97	[0.69, 1.36]	0.81	$1.58\times$
	$k=3$	1.10	[1.02, 1.19]	0.025	$1.11\times$
	$k=5$	1.00	[0.70, 1.41]	0.99	$1.59\times$
Config	$m=1$	1.18	[0.96, 1.43]	0.086	$1.30\times$
	$m=2$	0.95	[0.84, 1.08]	0.33	$1.19\times$
	$m=4$	0.98	[0.93, 1.03]	0.37	$1.07\times$
	$m=8$	0.97	[0.87, 1.09]	0.54	$1.17\times$

variation in task difficulty, which dominates the absolute errors but affects both operators equally. The MLP scale, MLP configuration and FNO scale sweeps are mutually consistent. The single exception is the scale sweep at $k=3$, where the CI marginally excludes 1; the effect is $1.10\times$, does not survive Bonferroni correction across the three $k \geq 2$ comparisons ($p=0.076$), and is not reproduced in the FNO arm ($0.93\times$ at $k=3$). We report it because suppressing a marginally significant result that runs against our conclusion would be selective reporting. The qualitative conclusion—advantage at $k=1$, collapse thereafter—holds in every MLP seed.

Table 10: Per-seed out-of-band test relative L_2 error, FNO backbone, compressible scale sweep (seeds 0–4).

k	Operator	s0	s1	s2	s3	s4
1	Naive	0.7458	0.7466	0.7155	0.7229	0.7942
1	Similarity	0.5291	0.5805	0.5475	0.5503	0.5536
2	Naive	0.3124	0.3584	0.3285	0.4039	0.2972
2	Similarity	0.3685	0.3600	0.2990	0.3298	0.3124
3	Naive	0.4848	0.3425	0.3670	0.3258	0.2687
3	Similarity	0.5701	0.3100	0.3280	0.4210	0.3112
5	Naive	0.2178	0.1959	0.1854	0.2711	0.1975
5	Similarity	0.1966	0.1947	0.1570	0.2715	0.1890

F Sample-matched control

In the scale sweep the training set grows with k ($20 \times 15 \times k$ rows), so the collapse at $k \geq 2$ could in principle reflect more data rather than more length scales. This control holds the training-set size at ≈ 300 rows—the $k=1$ size—by using fewer initial conditions as k grows: 20, 10, 7 and 4 ICs at $k=1, 2, 3, 5$ (300, 300, 315 and 300 rows). Everything else follows the main MLP sweep: the same seeds, the first n ICs of each seed’s sequence, the same (Re, Ma) grid, length-scale pool, test scales, objective and backbone; as in the main sweep, testing uses the training ICs at the held-out scales. At $k=1$ the control is identical to the main sweep and reproduces it seed for seed (Table 11).

Table 11: Sample-matched scale sweep (MLP, compressible, 5 seeds). Gap is the geometric mean of per-seed ratios; CIs are Student t on log-gaps, $df=4$, computed from the unrounded per-seed gaps.

k	ICs	Rows	Gap	95% CI	Per-seed gaps
1	20	300	$4.17\times$	[3.32, 5.24]	3.21, 4.44, 5.06, 3.75, 4.68
2	10	300	$1.05\times$	[0.62, 1.77]	0.73, 1.71, 1.58, 0.90, 0.71
3	7	315	$0.85\times$	[0.65, 1.13]	0.68, 0.75, 1.19, 0.76, 0.98
5	4	300	$0.99\times$	[0.62, 1.57]	0.77, 1.55, 1.15, 0.61, 1.15

The advantage collapses at $k \geq 2$ with the training-set size held fixed (pooled $0.96\times$), so the collapse is not a sample-count effect. The CIs are wider than in the main sweep because each run trains and tests on fewer ICs.

G Additional figures

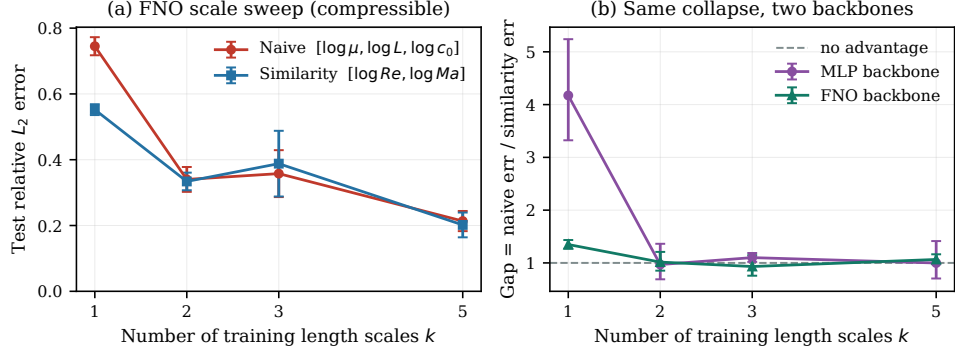


Figure 2: FNO replication of the compressible scale sweep, 5 seeds. (a) Both operators converge to the same out-of-band error as training diversity grows (mean $\pm$ std). (b) The gap collapse under the FNO backbone tracks the MLP backbone: an advantage at $k=1$ that vanishes by $k=2$ (geometric mean of per-seed ratios, 95% CI).

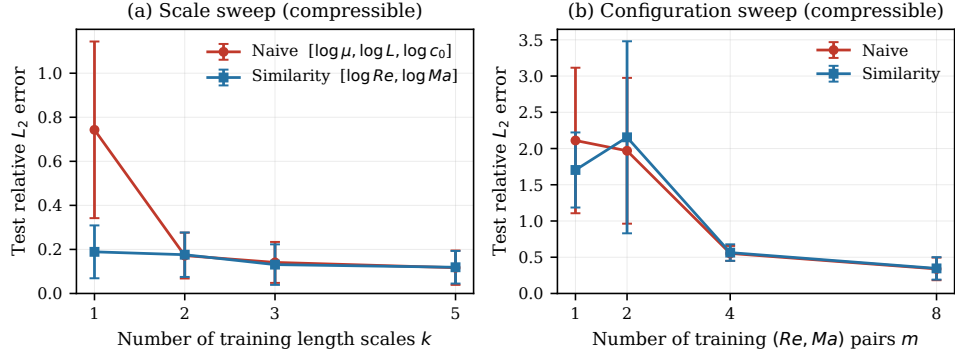


Figure 3: Out-of-band test error for both operators, MLP backbone (mean $\pm$ std, 5 seeds). (a) Scale sweep; (b) configuration sweep. The panels use different y -axis ranges (the configuration sweep starts from a harder, higher-error regime); the relevant comparison is between the two curves *within* each panel.

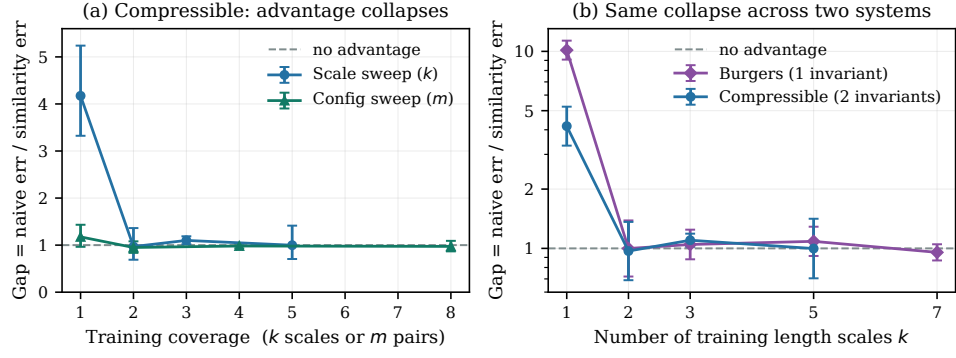


Figure 4: The transfer advantage of exact similarity encoding (gap = naive / similarity error) collapses toward 1. (a) Both compressible sweeps; (b) the same collapse in Burgers and compressible flow. Error bars are 95% CIs on the geometric mean of per-seed ratios, 5 seeds throughout.