

Moju: A Physics Admissibility Auditing Framework for Scientific Machine Learning Surrogates

Abiodun Olaoye
Ifimo Lab, Ifimo Analytics
lab@ifimoanalytics.com

Abstract

Scientific machine learning surrogates are often judged by how closely they satisfy governing partial differential equations. Constitutive relations receive less attention, yet a model can violate them while keeping PDE residuals near zero and still be physically inadmissible. We introduce **Moju**, an open-source library for physics admissibility auditing that supports JAX and PyTorch natively. Moju separates governing-law residuals from constitutive consistency and scores predictions across four tiers: HIGH, MODERATE, LOW, and NON-ADMISSIBLE. Applied to a one-dimensional slab cooling problem, seven fully-connected networks with hidden widths from 2 to 128 are trained for 14,000 steps. Every model satisfies the governing Fourier conduction law to at least 99.99%, yet constitutive admissibility ranges from 17.8% to 93.2%, a 75.4 percentage-point spread that governing-law residuals do not capture. Results suggest an inverted-U between model capacity and constitutive admissibility. Training trajectories show that every architecture passes through the LOW tier before converging, so the full tier range appears within a single benchmark. Train and evaluation scores agree within 0.5 percentage points, which suggests the audit reflects model behavior rather than evaluation grid density. Moju is available at <https://github.com/IfimoAI/moju>.

Keywords: physics-informed neural networks, constitutive relations, physics admissibility, surrogate model validation, scientific machine learning, capacity-admissibility relationship.

1. Introduction

Physics-informed neural networks embed governing equations directly into the training loss, producing solutions that are consistent with known physical laws by construction (Raissi et al., 2019). The same approach underpins neural operators (Li et al., 2021; Lu et al., 2021) and a growing class of engineering surrogates. In all these settings, training success is measured by PDE residual reduction: how closely predicted fields satisfy the governing equation on a set of collocation points.

Reducing governing-law residuals is a useful training signal, but it is not sufficient on its own. Every governing equation rests on constitutive assumptions about material behavior (Fourier’s law in heat conduction, Newton’s law in viscous flow, Darcy’s law in porous media), and a surrogate can drive the PDE residual toward zero while still violating those assumptions. Two distinct constitutive failure patterns appear in practice: under-capacity networks that

lack the expressiveness to maintain a coherent implied material model, and overparameterized networks that fit the governing-law loss through field derivative combinations inconsistent with the specified relation. Governing-law residual scores alone do not detect either pattern.

We present **Moju**, a physics admissibility auditing library that closes that gap with four tiers, separating governing-law satisfaction from constitutive consistency. The library provides:

- Governing-law residuals for 22 physics laws, nondimensionalized to $O(1)$ by construction.
- Automatic inference of constitutive relations implied by 11 governing law linkages.
- A four-tier scoring framework (HIGH, MODERATE, LOW, NON-ADMISSIBLE) from overall admissibility scores, with geometric-mean rollup for governing laws and worst-point audit scoring for closure keys (training loss always uses RMS).
- Native support for JAX and PyTorch workflows via a common `state_pred` interface.
- Constitutive divergence space-time heatmaps and worst-slice consistency plots that locate violations in space and time.

We demonstrate Moju on a one-dimensional slab cooling benchmark, comparing seven architectures that span the full admissibility tier range.

2. Background

2.1. Physics-Informed Neural Networks

PINNs (Raissi et al., 2019) solve forward and inverse PDE problems by minimizing a loss that combines data fidelity with PDE residuals at collocation points. Subsequent work addressed training instabilities through adaptive loss weighting (Wang et al., 2021) and causality-respecting collocation (Wang et al., 2022). Evaluation has remained focused on PDE residual magnitude and mean squared error against reference solutions, neither of which probes constitutive consistency.

2.2. Constitutive Relations and Implied Auditing

A PINN trained on the heat equation implicitly assumes that the ratio of the time derivative of temperature to its Laplacian equals the thermal diffusivity used in training. Violations of this implied relation indicate that the network satisfies the PDE through field derivative combinations that are physically inconsistent with the specified material. PDE residual metrics alone cannot detect this failure, whether it arises from insufficient capacity, overparameterization, or training pathology.

2.3. Related Validation Work

Domain-specific surrogate validation frameworks exist for aerodynamics (Brunton et al., 2020) and turbulence modeling (Beck and Kurz, 2021). Brunton et al. (2020) survey data-driven reduced-order models in fluid mechanics, where validation emphasizes reconstruction error and dynamical fidelity rather than per-point closure of material properties implied by the learned fields. Beck and Kurz (2021) review ML-augmented turbulence closures evaluated against DNS/RANS benchmarks; those workflows are problem-specific rather than a reusable audit layer across governing-law residuals and implied coefficients. Moju targets

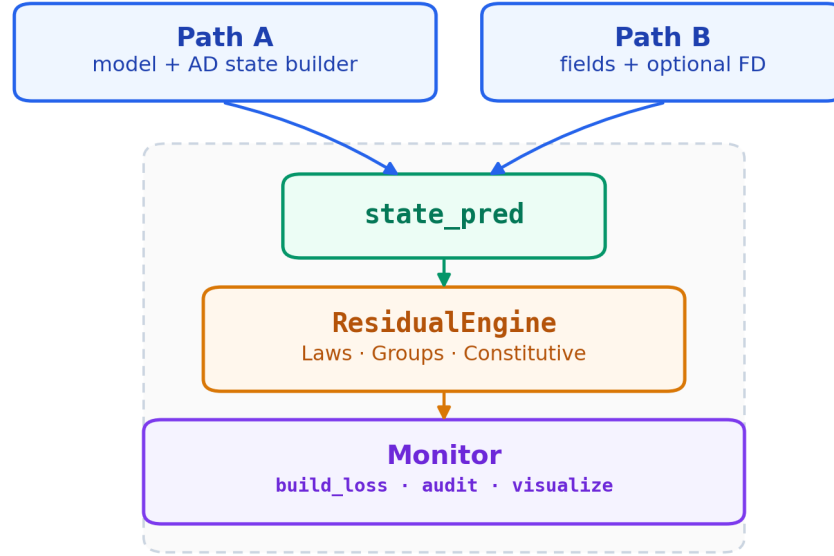


Figure 1: Moju processing pipeline. Path A (model with autodiff state builder) and Path B (pre-computed field variables with optional finite differences) both produce a `state_pred` dictionary. The `ResidualEngine` applies governing laws, dimensionless groups, and constitutive audits; downstream functions `build_loss`, `audit`, and `visualize` consume the resulting log.

the cross-cutting SciML setting in Section 2.2, where a surrogate can satisfy a governing-law residual while violating the material relation implied by its own field derivatives; to our knowledge, no general constitutive auditing framework of this kind has been published for scientific machine learning.

3. The Moju Framework

3.1. Architecture

Moju is a physics supervision layer, not a training framework. It operates on a `state_pred` dictionary of predicted field variables, derivatives, material properties, and coordinates, and works with any model or training stack. The core physics engine (Laws, Groups, Models, Operators) is implemented in JAX for differentiability and JIT compilation; PyTorch users access equivalent functionality through `TorchResidualEngine` and torch-native loss helpers in `moju[torch]`. Path A accepts a model and collocation points directly; Path B accepts `state_pred` directly, supporting CFD snapshots, digital twin pipelines, and any system that can export field variables.

3.2. Residual Computation

Governing-law residuals are defined for 22 physics laws and nondimensionalized to $O(1)$ by construction. For 11 of these laws, Moju automatically infers a law-linked constitutive residual, the fractional deviation between the material property in `state_pred` and the value

implied by field derivatives:

$$\delta = \frac{\alpha_{\text{model}} - \alpha_{\text{implied}}}{|\alpha_{\text{model}}| + \varepsilon}. \quad (1)$$

For Fourier conduction, α_{model} is the thermal diffusivity in `state_pred` and $\alpha_{\text{implied}} = (\partial T / \partial t) / (\partial^2 T / \partial x^2)$ at each collocation point, obtained by rearranging the governing law. Ill-conditioned points where the denominator vanishes are masked before scoring. This normalization makes δ directly interpretable as a percentage deviation from the expected material property, independent of problem scale.

3.3. Scoring and Tiers

Each logged residual key k receives a per-key admissibility score

$$a(k) = \frac{1}{1 + R_{\text{norm}}(k)}, \quad R_{\text{norm}}(k) = \frac{\text{score}(k)}{\text{scale}_k}, \quad (2)$$

with $a(k) \in (0, 1]$ (higher is more admissible) and $R_{\text{norm}}(k) = 1 \Rightarrow a(k) = 50\%$. The denominator scale_k is key-specific. For governing laws, Moju’s default auto_scale_k uses term balance: scale_k is the largest RMS among the main terms in that law’s nondimensional residual (for Fourier conduction, $\partial T / \partial t$ and $\alpha \nabla^2 T$), floored at 10^{-2} . Closure residuals use a fixed $\text{scale}_k \approx 10^{-2}$, which sets the tier cutoffs in the list below.

The numerator $\text{score}(k)$ depends on the residual key:

- **Governing laws** (`laws/*`): $\text{score}(k) = R_{\text{eff}}(k)$, the RMS over collocation points.
- **Closure residuals** (`constitutive/.../implied_delta` and `ref_delta`): $\text{score}(k) = \max_i |r_i|$, the worst-point magnitude.
- **Other keys** (custom constitutive terms, `data/*`): $\text{score}(k) = R_{\text{eff}}(k)$ (RMS).

Governing-law admissibility asks about the typical violation; closure admissibility asks about the worst point, so a favorable spatial average cannot hide a localized failure. Training loss always aggregates RMS R_{eff} ; admissibility auditing uses RMS for governing-law keys and worst-point scoring for closure keys.

Category scores aggregate per-key admissibilities: geometric mean for **laws**; minimum for **constitutive** when closure keys are present (otherwise geometric mean). Overall admissibility is the minimum of the governing-law and constitutive category scores; this scalar assigns the tier label.

Tier labels (**HIGH**, **MODERATE**, **LOW**, **NON-ADMISSIBLE**) are set from the overall admissibility score using fixed cutoffs ($\approx 90.9\%$, 66.7% , and 50%) derived from the same mapping at default closure scale. The paired $|\delta|_{\text{max}}$ bands below are not a separate classifier; they interpret each band in constitutive terms (Eq. 1) and align with the consistency plots in Section 4.5:

- **HIGH**: overall $> 90.9\%$ (interpreted as $|\delta|_{\text{max}} \leq 0.1\%$).
- **MODERATE**: $66.7\% - 90.9\%$ ($|\delta|_{\text{max}} \in (0.1\%, 0.5\%]$).
- **LOW**: $50\% - 66.7\%$ ($|\delta|_{\text{max}} \in (0.5\%, 1.0\%]$).
- **NON-ADMISSIBLE**: $< 50\%$ ($|\delta|_{\text{max}} > 1.0\%$).

3.4. Visualization

Moju produces Plotly dashboards including admissibility training trajectories, category breakdowns, spatial residual heatmaps, a constitutive divergence space-time heatmap showing $\delta(x, t)$ across the full domain, and a constitutive consistency line plot at the worst time slice with $\pm 0.1\%$, $\pm 0.5\%$, and $\pm 1.0\%$ tolerance band overlays. The worst time slice is selected by maximum absolute divergence.

4. Slab Cooling Benchmark

4.1. Problem Setup

We consider one-dimensional transient heat conduction in an aluminum slab:

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}, \quad x \in [0, 0.1 \text{ m}], \quad t \in [0, 60 \text{ s}], \quad (3)$$

with $\alpha = k/(\rho c_p)$ using aluminum properties ($\rho = 2700 \text{ kg/m}^3$, $c_p = 900 \text{ J/(kg K)}$, $k = 200 \text{ W/(m K)}$), giving $\alpha = 8.23 \times 10^{-5} \text{ m}^2/\text{s}$. A Neumann (zero-flux) condition is applied at $x = 0$ and a Robin (convective) condition at $x = L$, both enforced as weighted soft constraints. Predicted temperature fields agree with the analytical solution for this geometry. Full implementation details are provided in the accompanying Colab notebook.

4.2. Architectures and Training

Seven fully-connected networks with three hidden layers are evaluated, varying hidden width from 2 to 128 and spanning three orders of magnitude in parameter count. Optimization uses L-BFGS, a limited-memory quasi-Newton method that approximates curvature of the physics-informed loss and typically drives governing-law residuals lower per iteration than first-order methods once the landscape is sufficiently smooth. At 14,000 steps, governing-law admissibility is no longer the limiting factor across architectures, leaving constitutive admissibility as the discriminating signal in Table 1. All models train for 14,000 L-BFGS steps with Moju audit logs recorded every 200 steps (71 checkpoints including the random initialization at step 0). Training uses $64 \times 48 = 3,072$ temporal-by-spatial collocation points. Evaluation uses an $8 \times$ denser grid in each direction ($512 \times 384 = 196,608$ points).

Results are organized as follows: training trajectories (Figure 2), endpoint audit summary (Table 1 and Figure 3), constitutive diagnostic gallery (Figure 4), training and evaluation consistency (Figure 5), and the relationship between capacity and admissibility.

4.3. Training Trajectories

Figure 2 plots governing, constitutive, and worst-point r_{\max} admissibility over 14,000 steps. Governing scores converge within $\sim 2,000$ steps; constitutive trajectories show the capacity-dependent spread that PDE metrics do not resolve.

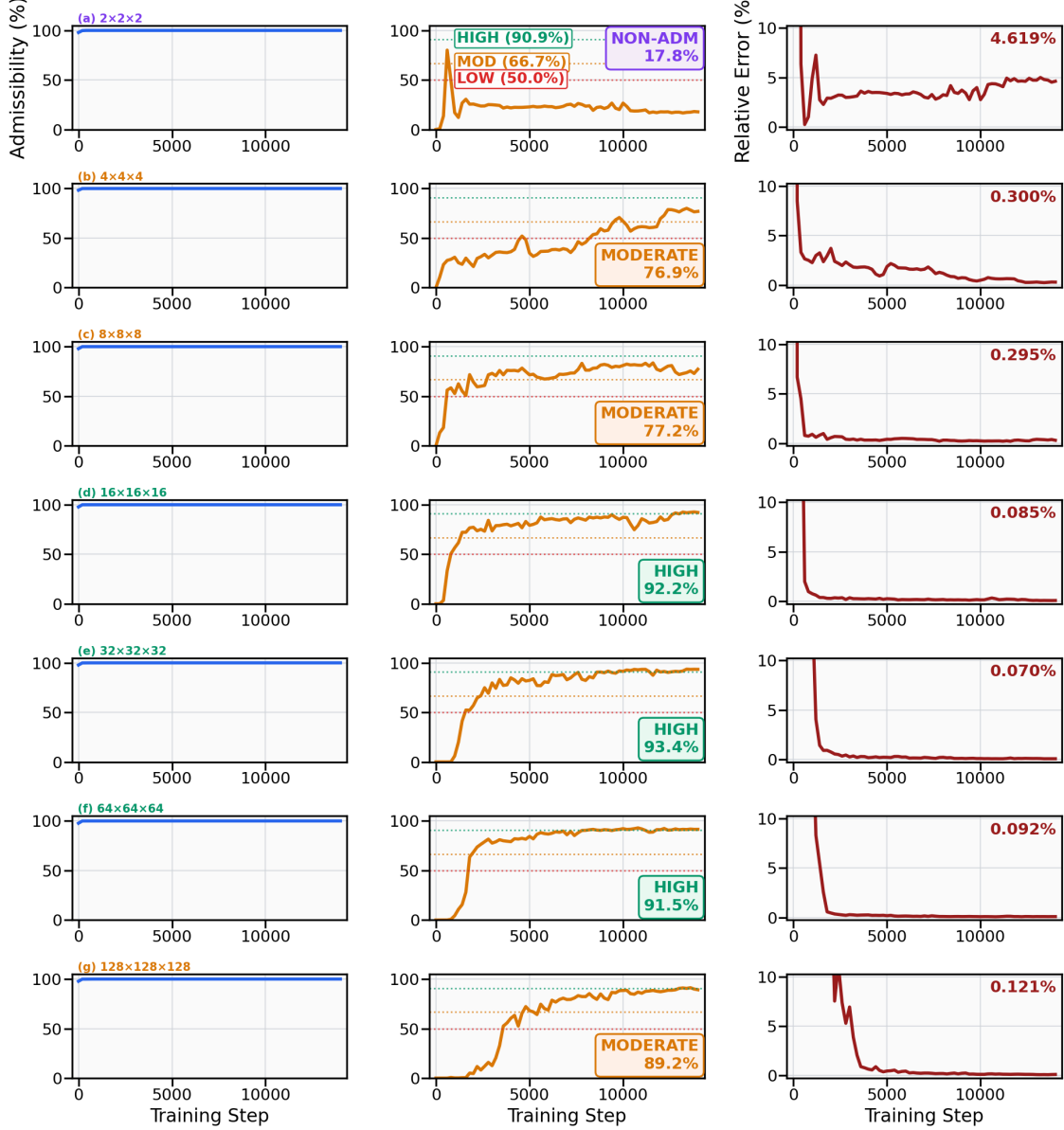


Figure 2: Training admissibility trajectories for architectures (a)–(g): governing admissibility (left), constitutive admissibility with tier thresholds (centre), and worst-point r_{\max} (right). Row labels appear above each panel. All models pass through LOW during early training; (a) remains NON-ADM at the endpoint while (e) reaches peak HIGH (93.4%). Scores along the trajectories are computed on the 3,072-point training collocation; endpoint evaluation-grid scores in Table 1 agree within 0.5 pp.

4.4. Endpoint Audit Summary

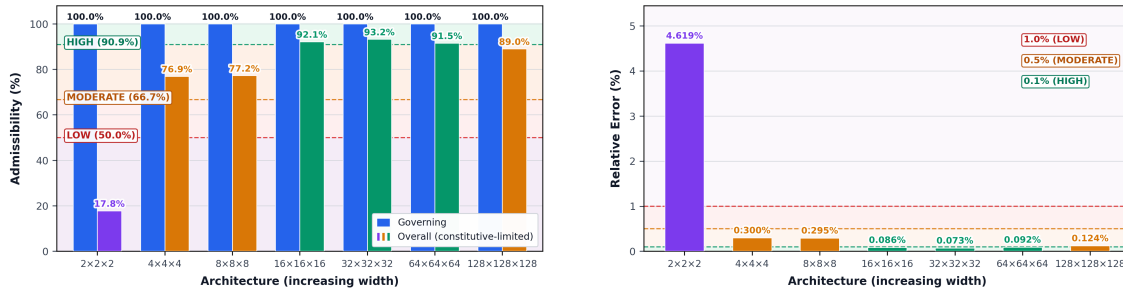
At the training endpoint (step 14,000), Table 1 reports evaluation-grid admissibility for all seven architectures. Governing-law admissibility is uniformly high ($\geq 99.99\%$) across all widths, so the tier column reflects constitutive performance alone: from NON-ADM at the smallest width through MODERATE and HIGH tiers as capacity increases.

Table 1: Evaluation-grid audit summary at step 14,000 (\star = peak constitutive admissibility). All models use three hidden layers with width shown in the architecture label. Constitutive admissibility and worst-point $|\delta|_{\max}$ are from the 196,608-point evaluation grid; governing scores are uniformly high ($\geq 99.99\%$). Tier follows eval overall admissibility.

Architecture	Params	Gov. (%)	Const. (%)	$ \delta _{\max}$ (%)	Tier
$2 \times 2 \times 2$	21	99.99	17.80	4.619	NON-ADM
$4 \times 4 \times 4$	57	100.00	76.93	0.300	MODERATE
$8 \times 8 \times 8$	177	100.00	77.21	0.295	MODERATE
$16 \times 16 \times 16$	609	100.00	92.12	0.086	HIGH
$32 \times 32 \times 32$ \star	2,241	100.00	93.19	0.073	HIGH
$64 \times 64 \times 64$	8,577	100.00	91.54	0.092	HIGH
$128 \times 128 \times 128$	33,537	100.00	88.98	0.124	MODERATE

The $2 \times 2 \times 2$ model remains NON-ADM with 17.80% constitutive admissibility and $|\delta|_{\max} = 4.619\%$, more than four times above the $\pm 1.0\%$ alarm band.

From $4 \times 4 \times 4$ through $8 \times 8 \times 8$, scores cluster in MODERATE; $16 \times 16 \times 16$ – $64 \times 64 \times 64$ reach HIGH, peaking at $32 \times 32 \times 32$ (93.19%). The widest network ($128 \times 128 \times 128$) regresses slightly to MODERATE despite $15\times$ more parameters than the peak width. Figure 3 visualizes this constitutive spread and the corresponding worst-point errors from Table 1.



(a) Overall admissibility by architecture, ordered by increasing hidden width, on the 196,608-point evaluation grid at step 14,000. Blue bars: governing admissibility; tier-coded bars: overall (constitutive-limited) admissibility.

(b) Worst-point fractional thermal diffusivity error $|\delta|_{\max}$ (thermal diffusivity implied δ) with $\pm 0.1\%$, $\pm 0.5\%$, and $\pm 1.0\%$ tolerance bands on the evaluation grid. The $2 \times 2 \times 2$ model exceeds the NON-ADM threshold by a factor of 4.6.

Figure 3: Summary audit metrics at step 14,000 on the 196,608-point evaluation grid (slab cooling benchmark, 14,000 training steps).

4.5. Constitutive Diagnostic Gallery

Figure 4 shows $\delta(x, t)$ heatmaps and worst-slice consistency traces on the 196,608-point evaluation grid. All panels share common δ ($\pm 0.1\%$) and α scales for direct cross-architecture comparison against the $\pm 0.1\%$, $\pm 0.5\%$, and $\pm 1.0\%$ bands.

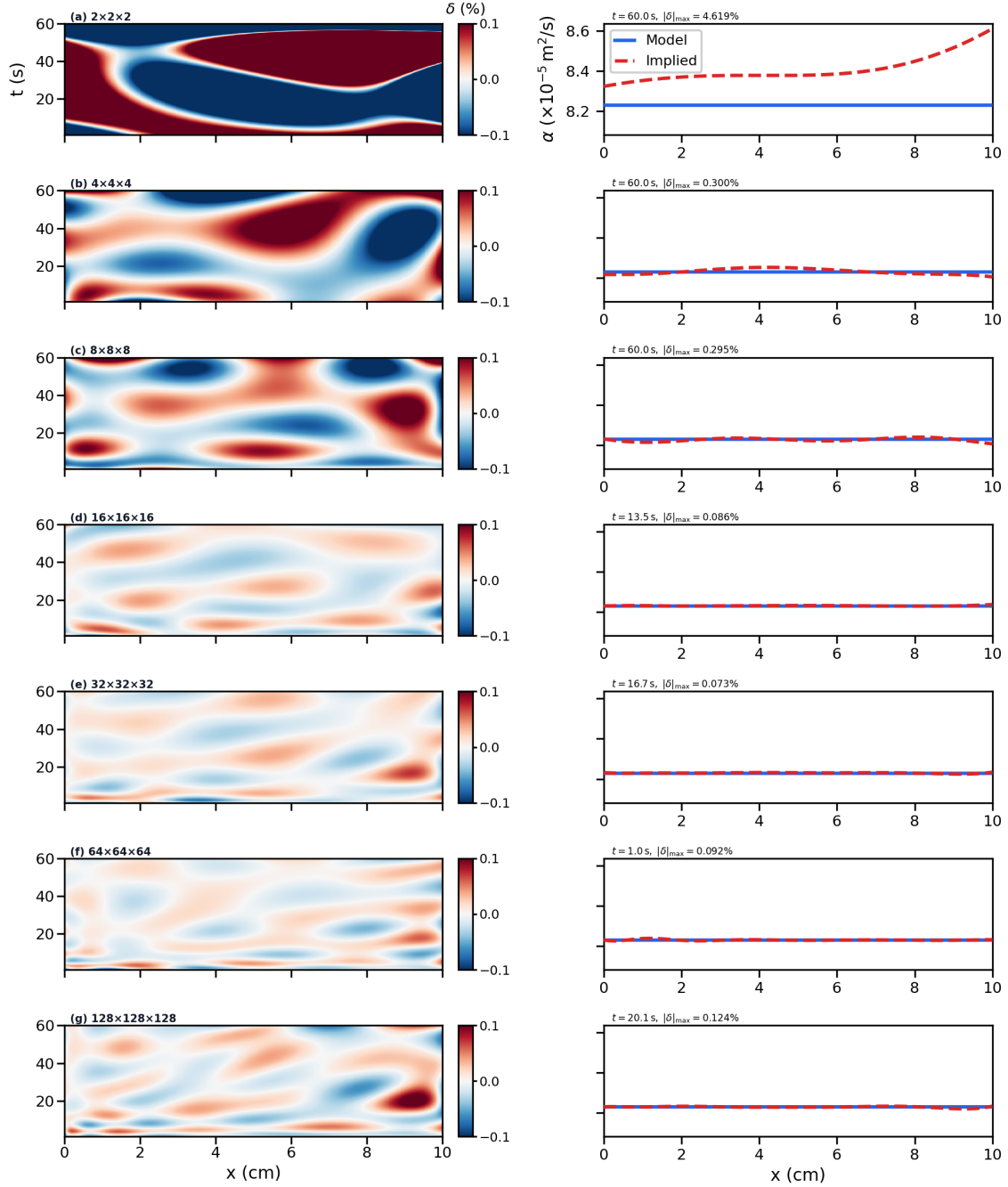


Figure 4: Constitutive diagnostic gallery for architectures (a)–(g): divergence heatmap $\delta(x, t)$ (left) and worst-slice model/implied α consistency (right) on the evaluation grid, with shared δ ($\pm 0.1\%$) and α scales. Row labels appear above each panel. (a)–(c) show capacity-limited drift and oscillatory violations; (d)–(f) show progressively tighter agreement between model and implied α ; (g) localized violations return at the widest width despite $15\times$ more parameters than the peak width (e).

4.6. Train-Evaluation Consistency

Training and evaluation admissibility scores agree within 0.5 pp across all seven architectures, with identical tier classifications at both collocation densities. Figure 5 shows that the audit tracks model behavior rather than grid density alone.

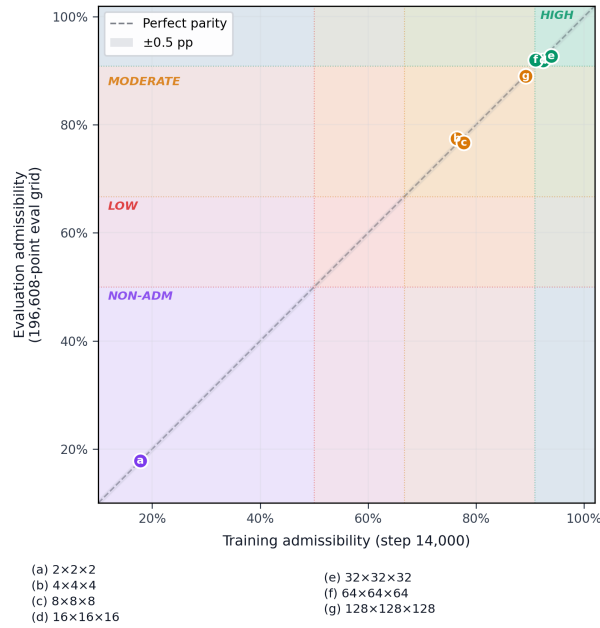


Figure 5: Training versus evaluation admissibility parity at step 14,000 for all seven architectures on the 196,608-point evaluation grid. Each marker is labeled (a)–(g); architecture names appear in the legend below. Closely spaced points are offset perpendicular to the parity line for visibility while remaining within the ± 0.5 pp band.

4.7. Analysis

4.7.1. Governing Law Satisfaction Is Universal but Insufficient

All seven architectures achieve governing-law admissibility above 99.99% at both collocation densities. That near-uniform result is why constitutive auditing matters: governing-law satisfaction alone cannot separate a physically admissible model from a constitutively inadequate one.

4.7.2. The LOW Tier as a Transient Training State

No architecture in this benchmark finishes training in the LOW tier, yet every run passes through it. The training trajectories in Figure 2 show that each architecture enters the LOW band during early training, typically within the first 1,000 to 3,000 steps, before either converging upward into MODERATE or HIGH, or stalling in the NON-ADMISSIBLE region as the $2 \times 2 \times 2$ model does. Here the LOW tier marks a transitional regime rather than a stable endpoint at this training budget. A smaller network or a more demanding physics problem would be more likely to end training in LOW. The four tiers remain useful because they describe both transient and final behavior; which tiers appear at the endpoint depends on capacity and training budget.

4.7.3. The Inverted-U Capacity-Admissibility Relationship

Across the seven architectures, constitutive admissibility suggests an inverted-U with hidden width. Under-capacity networks fail constitutively; an intermediate width achieves peak

admissibility; and the widest network shows a modest decline. Governing-law scores stay near perfect throughout and therefore cannot explain this pattern. Figure 3a makes the shape explicit.

4.7.4. The NON-ADMISSIBLE Signature: Capacity-Limited Drift

The $2 \times 2 \times 2$ architecture scores 17.80% (NON-ADMISSIBLE). Its worst-point thermal diffusivity error of 4.619% exceeds the $\pm 1.0\%$ alarm threshold by a factor of 4.6, and the domain-averaged RMS error of $\approx 2\%$ would also misclassify it under spatial averaging. The large-scale monotonic drift visible in panel (a) of Figure 4 is the characteristic pattern of a network that lacks capacity to satisfy the boundary conditions while maintaining a coherent material model. The $4 \times 4 \times 4$ model (panel b, 76.93%) shows the transition: structured oscillatory patterns replace the monotonic drift, indicating the network has gained expressiveness but cannot yet maintain constitutive consistency across the full domain.

4.7.5. The MODERATE-to-HIGH Progression

From $8 \times 8 \times 8$ through $64 \times 64 \times 64$, constitutive admissibility rises from 77.2% to 93.2%, and the divergence heatmaps transition from amber-band oscillatory patterns to low-amplitude green-band traces. The $32 \times 32 \times 32$ model achieves the highest admissibility at 93.2%, with a worst-point error of 0.073%. The $64 \times 64 \times 64$ model at 91.5% sits at the green-band boundary, suggesting that marginal constitutive degradation begins above width 32 for this problem.

4.7.6. Capacity Regression at $128 \times 128 \times 128$

The $128 \times 128 \times 128$ architecture scores 88.98% (MODERATE), 4.2 pp below the $32 \times 32 \times 32$ peak despite having $15 \times$ more parameters than the $32 \times 32 \times 32$ model. Panel (g) of Figure 4 shows localized violation patches with higher amplitude than the $64 \times 64 \times 64$ model. Larger networks can minimize the governing-law training loss through field derivative combinations that depart from the constitutive constraint.

5. Discussion

5.1. Tier-to-Failure-Mode Correspondence

The gallery in Section 4.5 (Figure 4) shows that Moju’s four tiers correspond to distinct failure modes. NON-ADMISSIBLE models exhibit large-scale monotonic constitutive drift from insufficient capacity. MODERATE models at the low end show high-amplitude oscillatory violations; those at the high end (e.g. $128 \times 128 \times 128$) show lower-amplitude localized violations when the network is overparameterized. HIGH models maintain $|\delta|_{\max}$ within the $\pm 0.1\%$ green band. This correspondence helps guide remediation: increase capacity for NON-ADM and low-MODERATE cases; adjust regularization, learning rate, or loss weighting when overparameterization drives regression.

The absence of a permanent LOW-tier result in this benchmark does not limit Moju’s usefulness. As the training trajectories in Section 4.3 show, every architecture passes through the LOW tier in early training. Here the tier is transient because each architecture has enough capacity to leave it within 14,000 steps. A smaller network or a stiffer physics

problem would be more likely to finish in LOW. The band definitions and detection logic are unchanged.

5.2. Architecture Selection

Since governing-law admissibility is above 99.99% for all seven architectures, loss-driven architecture selection cannot distinguish among them. Moju’s constitutive audit offers a practical criterion: choose the architecture that maximizes constitutive admissibility at fixed training budget. For the slab cooling benchmark, $32 \times 32 \times 32$ is the best choice on that basis.

5.3. Worst-Point Constitutive Scoring

Using worst-point r_{\max} rather than RMS for closure-key *audit* admissibility is what enables the NON-ADMISSIBLE classification of the $2 \times 2 \times 2$ model (training loss still aggregates RMS R_{eff}). A spatially averaged metric would mask the large positive drift visible at high x in panel (a) of Figure 4. In deployment, localized material property violations can matter most where the field gradients are steepest, even when a spatial average looks acceptable.

5.4. Limitations and Future Work

This study is limited to a one-dimensional, single-physics benchmark with fully-connected architectures (one run per width; no seed variance reported). We plan to extend Moju to two- and three-dimensional problems, multiple coupled governing laws, and neural operator architectures (FNO, DeepONet). The capacity–admissibility trend is consistent with an inverted-U here but not a statistically established scaling law; the width at which overparameterization begins to harm constitutive admissibility is likely problem-dependent. The Path B digital twin workflow, which supports CFD exports via OpenFOAM, VTK, and HDF5 adapters, is the subject of ongoing development and will be reported separately.

6. Conclusion

Moju provides a systematic framework for physics admissibility auditing of scientific machine learning surrogates. A benchmark study of seven architectures on one-dimensional slab cooling shows that Moju spans the full admissibility tier range, from NON-ADMISSIBLE (17.80%) to HIGH (93.19%), even though governing-law metrics treat all seven models as equivalent. The constitutive diagnostic gallery links distinct divergence heatmap signatures to distinct failure modes, from capacity-limited monotonic drift to overparameterization-induced oscillatory violations, in a pattern consistent with an inverted-U capacity–admissibility trend. Agreement within 0.5 pp between training and evaluation grids shows that the audit is stable under refinement of the collocation mesh.

Moju is available under the MIT license at <https://github.com/IfimoAI/moju> and via `pip install moju`. A reproducible Colab notebook for the slab cooling benchmark is provided in the repository examples.

Acknowledgments

The author thanks the open-source JAX, PyTorch, and Plotly communities whose infrastructure underlies Moju’s implementation.

References

- Andrea Beck and Marius Kurz. A perspective on machine learning methods in turbulence modeling. *GAMM-Mitteilungen*, 44(1):e202100002, 2021. doi: 10.1002/gamm.202100002.
- Steven L. Brunton, Bernd R. Noack, and Pierre Sagaut. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 52:477–508, 2020. doi: 10.1146/annurev-fluid-010719-060214.
- 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 E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021. doi: 10.1038/s42256-021-00302-5.
- 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.
- Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021. doi: 10.1137/20M1318043.
- Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022. doi: 10.1016/j.jcp.2021.110768.