LAW VIOLATION DETERMINES HOW STRONGLY PHYSICS SHOULD CONSTRAIN LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical constraints can reduce estimation error but introduce systematic bias when the assumed law is violated. We use measured law violation to choose hard constraints, soft penalties, or unconstrained learning before fitting a target predictor. The rule applies two invariance-specific thresholds calibrated on reference dynamics. An initial survey spans 12 partial differential equation (PDE) systems and five conservation or symmetry classes. Confirmatory experiments on held-out conservation and mechanism-aligned translation systems vary observation noise and training-set size. The selector reduces rollout error by 12.1% relative to a fixed convolutional neural network configuration selected by calibration-system validation. With backbone-specific recalibration, it also improves a Fourier neural operator while training only the selected target candidate. A linear–Gaussian analysis explains the three regimes for a fixed soft penalty and the inverse-square scaling of the continuously optimized penalty. These results turn constraint strength into a decision informed by observed dynamics and reusable calibration.

1 INTRODUCTION

Physical laws are often accurate without being exact. Sources and sinks weaken conservation, spatially varying coefficients break translation symmetry, and anisotropic transport selects a preferred direction. In each case, the same prior can reduce estimation error when the violation is weak and create systematic error when it is strong. Exact projection or equivariant layers reduce variance when the prior is accurate (Beucler et al., 2021; Cohen & Welling, 2016; Kondor & Trivedi, 2018; Weiler & Cesa, 2019; Finzi et al., 2021b), yet impose bias when the violation is real. Penalties and partially equivariant pathways interpolate between exact and unconstrained models (Lagaris et al., 1998; Raissi et al., 2019; Karniadakis et al., 2021; Finzi et al., 2021a; Wang et al., 2022; Romero & Lohit, 2022). The usual remedy is to train several variants and compare validation rollouts. Yet the trajectories used for training already contain a more direct piece of evidence: whether the proposed law is nearly exact, moderately broken, or no longer a useful restriction. This suggests that constraint strength can be chosen from the data before the competing predictors are fitted.

We formulate hard, soft, or unconstrained enforcement as a pre-training selection problem. For each invariance class I, a scalar $\hat{\delta}_I$ measures violation from noisy input–target pairs. Two class-specific thresholds turn this measurement into a model choice: small defects favor hard enforcement, large defects favor the unconstrained model, and the interval between them favors a soft penalty. Thresholds are calibrated on reference dynamics and frozen before transfer. Confirmatory experiments also fix the soft penalty using reference-system validation before training the held-out model. The selector never sees the hidden breaking strength, clean target trajectories, analytic evolution operator, or candidate-model errors on the target system.

The statistic must reflect the proposed invariance, not merely correlate with the data-generating parameter. We measure mass and L^2-energy violation by signed relative drift, and parity and fourfold rotation violation by the nonsymmetric energy of a fitted increment kernel. These quantities directly measure the candidate invariance. Translation instead uses positional non-stationarity of a local transfer operator to capture spatially varying coefficients, together with a complementary sufficiency statistic for strong reaction-type effects. These direct and operator-based measurements support a common pre-training decision across conservation and symmetry classes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The evidence combines breadth, robustness, and a cross-backbone test. An initial survey covers 12 controlled-breaking PDE systems, five invariance classes, and 2D incompressible Navier–Stokes. On the five calibration systems, $\hat{\delta}$ tracks the hard-versus-none error gap; on held-out systems, shuffling the statistic substantially reduces mode agreement. In a held-out convolutional neural network (CNN) confirmation for conservation and mechanism-aligned translation across nine noise–sample-size regimes, the unchanged selector attains a geometric paired error ratio of 0.879 to a fixed configuration selected by validation on the calibration system; all five data-and-training replicates favor selection and the 95% upper bound is 0.916. Repeating this test with a Fourier neural operator (FNO) confirms the same data-side ordering on a modern operator backbone. FNO-specific calibration attains aggregate ratio 0.872, with all 18 class-by-regime cells within the 1.05 ratio tolerance, and matches the aggregate error of a same-core learned-relaxation baseline while reducing marginal training cost. A joint-calibration bootstrap gives an upper ratio of 0.891 to fixed. The resulting selector preserves the hard/soft/unconstrained decision while avoiding 87–88% of exhaustive target-candidate training after calibration. Across all stages, $19{,}005$ trained models enter the analysis. Statistical uncertainty is measured over independent repetitions with fresh trajectories and optimization randomness, not over the individual fits within a repetition.

The contribution is a pre-training rule that connects measurable physical-law violation to constraint strength. Direct conservation and symmetry defects transfer across several dynamics, while translation requires a statistic that matches the breaking mechanism. A linear–Gaussian model makes the three-way structure explicit: it yields separate hard–soft and soft–none boundaries and predicts how the preferred penalty weakens as violation grows. The experiments test this decision across five invariance classes, 12 PDE systems, and two backbones, measuring both generalization and the candidate training avoided by choosing early.

2 Related Work

Exact and approximate equivariance. Geometric learning provides exact equivariance through group convolutions and steerable or general equivariant layers (Bronstein et al., 2021; Kondor & Trivedi, 2018; Weiler & Cesa, 2019; Finzi et al., 2021b). Approximate-equivariance methods instead learn how much symmetry to retain, using relaxed group convolutions (Wang et al., 2022), equivariant and free residual pathways (Finzi et al., 2021a), non-stationary filters (van der Ouderaa et al., 2022), or learned partial-equivariance levels (Romero & Lohit, 2022). Approximate equivariance can also be expressed through fixed auxiliary inputs (Ashman et al., 2024), while architecture search and constrained optimization adapt the retained symmetry during training (Maile et al., 2023; Manolache et al., 2025). Other methods discover invariances from data (Benton et al., 2020; Dehmamy et al., 2021). These works construct or search a model class. Our problem is complementary: given existing hard, soft, and unconstrained implementations, we choose which one to train on a new system from a measured defect in its transitions.

Physical structure in learned dynamics. Conservation can be imposed through projection, probabilistic correction, or a surrogate penalty (Hansen et al., 2023; Beucler et al., 2021). Neural PDE solvers and physics-informed neural networks encode governing equations through soft residuals (Lagaris et al., 1998; Raissi et al., 2019; Karniadakis et al., 2021), while Hamiltonian networks and group-equivariant architectures build conserved quantities or symmetries into the model (Greydanus et al., 2019; Cohen & Welling, 2016). These approaches define the constraint modes studied here. Our criterion addresses the preceding question: whether the observed violation justifies exact, penalized, or unconstrained learning.

Operator learning and simulation. Fourier and general neural operators learn maps between function spaces (Li et al., 2021; Kovachki et al., 2023); DeepONet provides a parallel operator construction (Lu et al., 2021); and graph or message-passing simulators learn rollouts on meshes and PDE discretizations (Pfaff et al., 2021; Brandstetter et al., 2022). Learned discretizations provide another route for incorporating numerical structure (Bar-Sinai et al., 2019), while recent FNO variants explicitly align inputs to symmetry frames (Xu et al., 2026). Our selection problem is orthogonal to backbone design: it is evaluated by multistep rollout generalization and tested on both convolutional and Fourier-operator backbones, with calibration performed for each backbone.

Pre-training selection. The closest methods make related decisions from data but use different signals and decision times. Lawrence et al. (2026) use a two-sample classifier before training to measure

distributional symmetry breaking and assess whether augmentation is warranted; our statistics instead target violations of the evolution law. van der Linden et al. (2025) select among pretrained models with different symmetry biases using uncertainty measures. Controlled comparisons also show that the useful symmetry restriction depends on task geometry (Vadgama et al., 2025). Our rule is three-way, precedes candidate training on the held-out system, and covers conservation as well as symmetry. Its rationale is consistent with theory relating the useful equivariance of a model to the equivariance error of the data (Petrache & Trivedi, 2023) and with strict generalization gains available under exact equivariance (Elesedy & Zaidi, 2021). We make the violation magnitude operational as a hard/soft/unconstrained rule. Unlike architecture search or learned relaxation, its output is an explicit choice among existing model families; unlike post-training uncertainty selection, the choice is made before those target models exist.

3 Problem Setup and Controlled-Breaking Benchmark

For each dynamical system, we learn a one-step flow map $G : u(t) \mapsto u(t + \Delta t)$ from noisy trajectories. The primary metric is the relative ℓ_2 error of a $K = 10$ step autoregressive rollout on clean, held-out trajectories. Errors are capped at 1 for aggregation so that a few divergent rollouts do not dominate comparisons; uncapped errors are retained as a directional check. The grids contain 64 points for the 1D periodic PDEs, 32×32 points for the 2D fields, and 48×48 points for Navier–Stokes. Separating noisy training from clean evaluation tests whether a constraint improves generalization rather than the fit to corrupted one-step pairs.

Each system has a candidate invariance I and a nonnegative breaking parameter δ in the data-generating process. The invariance is exact at $\delta = 0$ and is violated continuously as δ increases. The learner does not observe δ; given only its noisy training trajectories, it must choose hard, soft, or no enforcement of I before fitting the selected candidate.

For the CNN studies, we instantiate the three modes with a shared residual architecture, $u_{t+\Delta t} = u_t + f(u_t)$, using circular convolutions in 1D and 2D. The hard mode enforces I by construction, the unconstrained mode leaves f free, and the soft mode adds an I-defect penalty with $\lambda \in \{0.01, 0.1, 1.0\}$. Specifically, hard enforcement uses a zero-mean increment projection for mass conservation, removes positional channels for translation equivariance, ties reflection-related kernel weights for parity, rescales the predicted next state to preserve quadratic energy, and averages the learned update over C_4, the rotations by multiples of 90°. These three penalty strengths form the CNN soft candidates used by validation and oracle comparisons; the FNO study uses the wider grid specified in Section 6.1.

The benchmark contains 12 systems in five invariance classes. One system in each class calibrates the class-specific rule; all other systems are held out from threshold calibration. Mass conservation uses linear advection with decay for calibration and viscous Burgers with decay for transfer. Translation uses variable-coefficient diffusion for calibration, then transfers to advection–diffusion with the same spatially varying diffusion mechanism and to diffusion with a spatially varying reaction term. Parity uses advection–diffusion for calibration and Burgers dynamics for transfer. Quadratic energy uses advection for calibration and advective–dispersive dynamics for transfer. Finally, fourfold rotation uses anisotropic diffusion for calibration and transfers to anisotropic advection and 2D incompressible Navier–Stokes in vorticity form. The Navier–Stokes solver is pseudo-spectral; isotropic viscosity is C_4-equivariant, whereas anisotropic viscosity breaks this symmetry by an amount controlled by δ. Appendix A gives the equations, solvers, and architecture details.

The two translation transfers deliberately separate mechanism alignment from class identity. The primary statistic measures positional non-stationarity of a local transfer operator and captures spatially varying diffusion. A complementary cross-validated sufficiency gap responds to strong reaction-type breaking. Together, the aligned and complementary transfers test how operator-derived violation statistics support constraint selection across distinct translation-breaking mechanisms.

The initial survey uses seven breaking levels, $\delta \in \{0, .02, .05, .1, .2, .4, .8\}$, trajectories of length $T = 12$, and five independent data-and-training replicates. Here n_{train} counts training trajectories, and noise levels are standard deviations relative to the global standard deviation of clean training data. Mass-conservation and translation systems use noise 0.12 and $n_{\text{train}} = 96$; parity uses noise 0.12 and $n_{\text{train}} = 48$; energy uses noise 0.20 and $n_{\text{train}} = 24$; rotation uses noise 0.20 and $n_{\text{train}} = 12$,

except Navier–Stokes, which uses noise 0.15 and $n_{\text{train}} = 24$. Every violation statistic is computed from the same noisy training trajectories available to the learner, never from clean test data, the hidden δ, the analytic flow map, or a fitted candidate's error.

4 A Pre-Training Selection Criterion

Given a candidate invariance I, deployment on a held-out system selects among hard, soft, and unconstrained enforcement rather than learning an adaptive mixture jointly with the predictor. A scalar statistic $\hat{\delta}_I$ is computed from noisy training pairs before any candidate is fit on that system. Two class-specific thresholds then map small, intermediate, and large measured violations to hard, soft, and unconstrained models, respectively. This construction follows the principle that useful model equivariance should track the equivariance error of the data (Petrache & Trivedi, 2023), but makes the principle operational as a pre-training decision.

Direct conservation statistics. Let $(u_p^{\text{in}}, u_p^{\text{out}})_{p=1}^{P}$ be all consecutive pairs from the training trajectories and $\Delta u_p = u_p^{\text{out}} - u_p^{\text{in}}$. Throughout this section, $\varepsilon > 0$ is a fixed numerical stabilizer. For mass $M(u) = \sum_i u_i$, where i indexes grid sites, we use the magnitude of the signed mean relative drift,

$$\hat{\delta}_{\text{cons}} = \left| \frac{1}{P} \sum_{p=1}^{P} \frac{M(u_p^{\text{out}}) - M(u_p^{\text{in}})}{|M(u_p^{\text{in}})| + \varepsilon} \right|. \tag{1}$$

The absolute value is applied after averaging. When mass is conserved, sign-fluctuating observation noise can therefore cancel across pairs; averaging the absolute drift of each pair would instead introduce a positive noise floor. For quadratic energy or enstrophy, $E(u) = \sum_i u_i^2$, we analogously define

$$\hat{\delta}_{\text{energy}} = \left| \frac{1}{P} \sum_{p=1}^{P} \frac{E(u_p^{\text{out}}) - E(u_p^{\text{in}})}{E(u_p^{\text{in}}) + \varepsilon} \right|. \tag{2}$$

Both statistics are dimensionless and directly measure drift of the proposed conserved quantity.

Direct kernel-symmetry statistics. We fit a shared local kernel w by ridge regression from u^{in} to the increment Δu. Reflection equivariance requires an even kernel, so the parity statistic is its antisymmetric energy fraction, with k indexing kernel offsets,

$$\hat{\delta}_{\text{par}} = \frac{\|w^{\text{anti}}\|_2^2}{\|w\|_2^2 + \varepsilon}, \qquad w^{\text{anti}} = \tfrac{1}{2}(w - \widetilde{w}), \quad \widetilde{w}_k = w_{-k}. \tag{3}$$

The rotation statistic is the two-dimensional analogue. If $\Pi_{C_4} w = \frac{1}{4} \sum_{g \in C_4} g \cdot w$ is the fourfold rotational average, then $\hat{\delta}_{\text{rot}} = \|w - \Pi_{C_4} w\|_2^2 / (\|w\|_2^2 + \varepsilon)$. These quantities measure the symmetry-breaking component of the fitted local operator rather than the hidden physical parameter that generated it. For nonlinear dynamics, they characterize the fitted linear probe rather than the full nonlinear flow.

A translation statistic and complementary sufficiency gap. Translation equivariance is not an additive invariant. We instead test whether the local input–output relation is stationary. On a grid with N positions, we fix a window half-width W. At each position i, ridge regression fits $u_i^{\text{out}} \approx w_i^\top u_{[i-W:i+W]}^{\text{in}}$ using circular windows, and we define

$$\hat{\delta}_{\text{trans}} = \frac{N^{-1} \sum_{i=1}^{N} \|w_i - \bar{w}\|_2^2}{\|\bar{w}\|_2^2 + \varepsilon}, \qquad \bar{w} = N^{-1} \sum_{i=1}^{N} w_i. \tag{4}$$

A stationary local operator gives similar coefficients across positions, while spatially varying coefficients increase the dispersion measured by Equation (4).

We complement this statistic with a two-fold cross-validated sufficiency gap. A shared local kernel and unconstrained position-specific kernels are fit on one training fold and evaluated on the other. With their held-out residual sums of squares denoted by RSS_{eq} and RSS_{free}, respectively,

$$\hat{\delta}_{\text{trans}}^{\text{gap}} = \max\left\{0, \frac{\text{RSS}_{\text{eq}} - \text{RSS}_{\text{free}}}{\text{RSS}_{\text{eq}} + \varepsilon}\right\}. \tag{5}$$

Evaluation on held-out folds limits the in-sample advantage of the position-specific probe, and the truncation makes the reported statistic nonnegative. The primary selector uses positional non-stationarity, while the sufficiency gap provides an auxiliary measurement for strong reaction-type effects.

Calibration and deployment. For each class and $t_1 < t_2$, the thresholds define

$$\pi_I(\hat{\delta}_I) = \begin{cases} \text{hard}, & \hat{\delta}_I \le t_1, \\ \text{soft}, & t_1 < \hat{\delta}_I < t_2, \\ \text{none}, & \hat{\delta}_I \ge t_2. \end{cases} \tag{6}$$

If the fitted thresholds coincide, we instead define hard for $\hat{\delta}_I \le t_1$ and none for $\hat{\delta}_I > t_1$; the soft region is empty. The thresholds are class-specific because the statistics have different scales and exact enforcement removes different amounts of model capacity. On the single calibration system for each class, we grid-search threshold locations to minimize cumulative clean rollout error across the seven breaking levels. The broad five-class study fits this rule on its calibration system and then fixes it before evaluation on the other dynamics.

The selector returns an enforcement family, not a penalty coefficient. In the initial five-class study, we profile the three soft strengths to isolate the family-level question. In the confirmatory studies, one λ is selected for each class and observation regime using only calibration-system validation trajectories, then held fixed on the corresponding held-out system. The fixed comparator is selected from the same candidate pool by the identical calibration-validation rule: three soft strengths for the CNN and five for the FNO, alongside hard and unconstrained models. In these confirmatory studies, no held-out-system test error selects a threshold or penalty.

At deployment, $\hat{\delta}_I$ uses the invariance class, noisy training trajectories, and Δt. The true δ, clean trajectories, analytic flow map, and candidate-model errors remain unavailable; the folds in Equation (5) split only training trajectories. A shuffled-statistic control tests whether the association between measured violation and system instance carries the result.

Cross-validation provides a post-training comparator: it trains every candidate and selects by validation rollout. Our rule first measures violation and trains the selected candidate. We report predictive error and measured training cost for both decision procedures.

5 Why Three Constraint Regimes Emerge

We use a linear model to isolate the bias–variance mechanism behind the three-way rule. Let $\delta \ge 0$ denote violation magnitude, $\mathcal{S} \subset \mathbb{R}^p$ an s-dimensional constraint subspace, and $r = p - s$. The target parameter is

$$\theta^\star = \theta_{\mathcal{S}} + \delta\theta_\perp, \qquad \theta_{\mathcal{S}} \in \mathcal{S}, \quad \theta_\perp \in \mathcal{S}^\perp. \tag{7}$$

Given n observations, let $X \in \mathbb{R}^{n\times p}$ and $y = X\theta^\star + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\sigma > 0$. The unconstrained estimator fits all p coordinates; the hard estimator is restricted to $\mathcal{S}$; and the soft estimator solves

$$\widehat{\theta}_\lambda = \arg\min_{\theta\in\mathbb{R}^p} \|y - X\theta\|_2^2 + \lambda\|P_\perp\theta\|_2^2, \tag{8}$$

where $P_\perp$ is the orthogonal projection onto $\mathcal{S}^\perp$. For an independent test covariate x with $\mathbb{E}[xx^\top] = I_p$, define excess prediction risk as $R(\widehat{\theta}) = \mathbb{E}[(x^\top(\widehat{\theta} - \theta^\star))^2]$, where the expectation is over training noise and the test covariate.

Proposition 1 (Three-way risk regions)**.** *Suppose $X^\top X = nI_p$, $r > 0$, and $q := \|\theta_\perp\|_2^2 > 0$. Write $B = \delta^2 q$, $V = \sigma^2 r/n$, $C = \sigma^2 s/n$, and $a_\lambda = n/(n+\lambda)$. Then*

$$R_{\text{hard}} = C + B, \qquad R_\lambda = C + (1-a_\lambda)^2 B + a_\lambda^2 V, \qquad R_{\text{none}} = C + V. \tag{9}$$

For any fixed $0 < \lambda < \infty$, the risk-minimizing choice is

$$\arg\min\{R_{\text{hard}}, R_\lambda, R_{\text{none}}\} = \begin{cases} \text{hard}, & B \le \dfrac{a_\lambda}{2-a_\lambda} V, \\ \text{soft}, & \dfrac{a_\lambda}{2-a_\lambda} V < B < \dfrac{1+a_\lambda}{1-a_\lambda} V, \\ \text{none}, & B \ge \dfrac{1+a_\lambda}{1-a_\lambda} V, \end{cases} \tag{10}$$

up to ties at the boundaries. If λ can vary continuously, its risk-minimizing value is

$$\lambda^\star = \frac{\sigma^2 r}{\delta^2 q}, \tag{11}$$

with $\lambda^\star = \infty$ at $\delta = 0$. Thus $\lambda^\star \propto \delta^{-2}$, with hard and unconstrained learning as the limiting solutions when δ approaches zero and infinity.

Proof. The subspace coordinates contribute the common variance C. In the orthogonal block, hard enforcement has squared bias B, no enforcement has variance V, and soft enforcement has bias $-(1 - a_\lambda)\delta\theta_\perp$ and scales the estimation noise by a_λ, giving Equation (9). Comparing the soft risk first with B and then with V yields Equation (10). Minimizing the same quadratic over $a_\lambda \in [0, 1]$ gives $a^\star = B/(B + V)$ and hence Equation (11). □

For a fixed soft penalty, the proposition gives the selector its three-way form. At negligible violation, hard enforcement removes variance at little bias cost. At intermediate violation, partial shrinkage has the lowest risk. At large violation, retaining the full signal matters more than the variance reduction. The boundaries also depend on the capacity r removed by exact enforcement and on the scale q of the violating component. A constraint that removes little capacity can therefore have only a narrow neighborhood around exact invariance in which hard enforcement offers a meaningful advantage.

Distance from a threshold is informative as well. Let δ_{hs} and δ_{sn} denote the two boundaries obtained from Equation (10). For an error tolerance $\epsilon \geq 0$, if an estimate obeys $|\widehat{\delta} - \delta| \leq \epsilon$, its plug-in decision agrees with the risk-minimizing family whenever $\min_{j \in \{\mathrm{hs},\mathrm{sn}\}} |\delta - \delta_j| > \epsilon$. Near a boundary, two families have similar risk and validation is useful; away from both boundaries, the measured defect alone determines the choice. The empirical statistics need not share the latent scale of this model, which is why their thresholds are calibrated by invariance class.

Figure 1 illustrates the calculation by Monte Carlo simulation with $p = 12$, $r = 4$, $n = 40$, $\sigma = 0.5$, and $\|\theta_\perp\|_2 = 1$. For a fixed penalty $\lambda = 40$, the predicted hard–soft and soft–none boundaries are 0.091 and 0.274. Monte Carlo risks follow these regions, while the empirically optimal penalty follows the inverse-square trend. When the shrinkage fraction and other factors are fixed, both boundaries move as $\sigma/\sqrt{n}$. This motivates the theory-inspired rescaling evaluated in Section 6.1, alongside the retained empirical calibration.

6 Experiments

We evaluate three questions. Within each invariance class, does the measured violation order the relative value of hard and unconstrained models? Does a rule calibrated on one dynamical system transfer to another? And does the selection gain persist across observation regimes and model backbones? The initial CNN survey addresses the first two questions on all 12 systems while

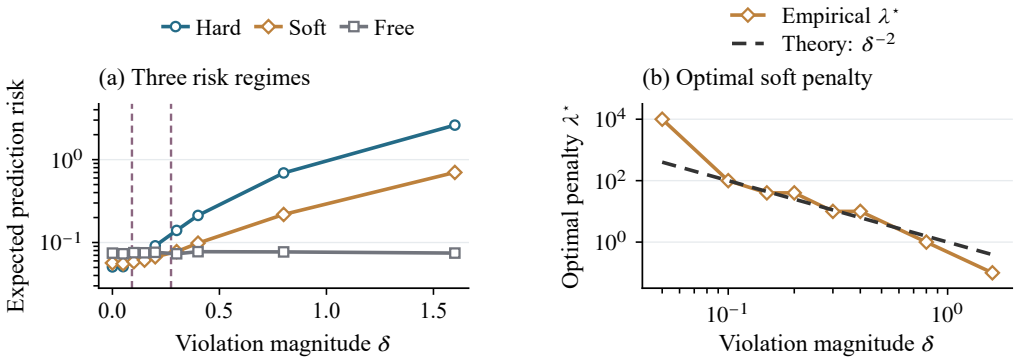

Figure 1: Risk regimes in the linear–Gaussian model. **(a)** Monte Carlo risk with $\lambda = 40$; dashed boundaries separate hard, soft, and unconstrained optima. **(b)** Optimized penalties follow the predicted $1/\delta^2$ scaling.

profiling the three soft penalties. The stricter confirmations choose the soft penalty using only calibration-system validation and evaluate conservation and mechanism-aligned translation over a fixed noise–sample-size grid. The full evaluation comprises 19,005 trained models: 2,100 in the initial survey, 6,825 in the CNN confirmation, and 10,080 in the FNO cross-backbone test. The primary metric throughout is capped 10-step relative ℓ_2 rollout error on clean held-out trajectories.

Figure 2 shows the motivating crossover. As the breaking strength increases, the relative error of hard enforcement grows and the best family moves toward soft or unconstrained learning. On the five calibration systems, the Pearson correlation between $\hat{\delta}_I$ and the hard-minus-none error gap is 0.968 for mass conservation, 0.969 for energy, 0.973 for parity, 0.980 for rotation, and 0.994 for translation. Thus the class-specific statistics consistently order when exact enforcement becomes costly, despite their different numerical scales.

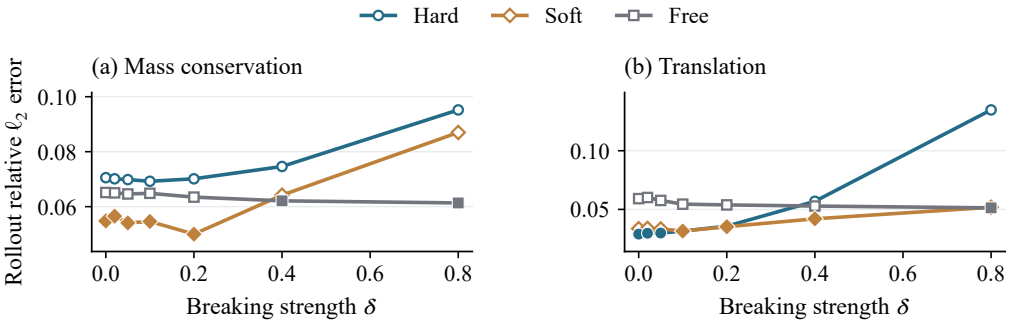

Figure 2: Constraint crossovers for **(a)** mass conservation and **(b)** translation equivariance. Filled markers identify the lowest-error family at each breaking strength δ; curves average five replicates.

Table 1 reports the family-level transfer result. On calibration systems, the rule matches the profiled family oracle. On the seven held-out systems, it attains mean error 0.1171, compared with 0.1375 for the best fixed family and 0.1070 for the test-error oracle. Across all 84 system–breaking-level pairs, its excess error over the oracle is 0.0059. This experiment isolates the hard/soft/none family decision by profiling the three soft penalties. The confirmations below select one penalty using only calibration-system validation.

Table 1: Five-class CNN survey. Soft-family error is profiled over $\lambda \in \{.01, .1, 1\}$. Entries are mean capped rollout errors; the oracle uses clean test error as a reference.

	criterion	best fixed	oracle
calibration (35 pairs)	0.0544	0.0562	0.0544
held out (49 pairs)	0.1171	0.1375	0.1070
all (84 pairs)	0.0910	0.1143	0.0851
mode match	calibration 100%; held out 71%; all 83%		

The per-system results reinforce the aggregate transfer pattern. Held-out gaps to the family oracle are 0.0004 for parity, 0.0004 for rotational advection, and 0.0006 for Navier–Stokes. For translation, positional non-stationarity transfers from varying diffusion to advection with the same varying-diffusion mechanism, while the complementary sufficiency gap captures strong reaction-type violation. Appendix B.3 reports all system-level gaps and mode decisions.

A randomization control tests whether class labels alone explain the result. Across 200 shuffles of $\hat{\delta}_I$ among held-out system–breaking pairs, mean mode match falls from 0.71 to 0.43 and its 95th percentile is 0.55; replacing the measurements by random values also gives mean match 0.43. Appendix A reports the fitted thresholds.

6.1 Robustness across noise and sample size

The CNN confirmation crosses noise $\{0.06, 0.12, 0.20\}$ with training sizes $\{48, 96, 192\}$ for mass conservation and mechanism-aligned translation. It uses the two calibration systems and their

Table 2: Constraint selection across observation regimes and backbones. The CNN comparison contains 6,300 aligned fits plus a 525-fit mechanism control; the FNO comparison adds 10,080 fits. Error entries are geometric ratios to the calibration-validation-selected fixed configuration, and lower is better. "Fits" is the number of candidates trained per held-out instance. "Upper" is the 95% upper bound versus fixed; † denotes the joint calibration–transfer bootstrap that refits all calibration choices. FNO times use the identical 630 transfer instances, and "saved" is relative to training all seven FNO candidates. Bold identifies the retained CNN selector and the strongest calibration-only FNO selector.

			rollout-error ratio to fixed				measured deployment cost	
backbone	policy / ablation	fits	all	CONS	TRANS	upper	A40 GPU h	saved
CNN	fixed configuration	1	1.000	1.000	1.000	1.000	–	–
CNN	**unchanged thresholds**	1	**0.879**	**0.798**	0.968	**0.916**	–	–
CNN	theory rescaled	1	0.891	0.814	0.976	0.926	–	–
CNN	calibration refit	1	0.861	0.827	0.897	0.887	–	–
CNN	per-instance CV	5	0.697	0.572	0.850	0.724	–	–
CNN	test-error oracle	5	0.693	0.569	0.843	0.719	–	–
FNO	fixed configuration	1	1.000	1.000	1.000	1.000	–	–
FNO	unchanged CNN thresholds	1	0.964	0.969	0.960	0.978	5.145	88.0%
FNO	**calibration refit**	1	**0.872**	**0.842**	**0.903**	$0.891^{\dagger}$	**5.414**	**87.4%**
FNO	matched learned relaxation	1	0.873	–	–	–	6.910	83.9%
FNO	all seven candidates	7	–	–	–	–	43.048	0.0%

aligned held-out systems, seven breaking levels, five fresh data-and-training replicate blocks, and five candidates (hard, unconstrained, and soft with $\lambda \in \{.01, .1, 1\}$), totaling 6,300 fits. A complementary reaction-mechanism control at three regimes adds 525 fits. Within each class and regime, calibration-system validation selects one soft penalty. The fixed comparator similarly selects one complete candidate on calibration validation and uses it at every breaking level of the held-out system.

We first average errors across the seven breaking levels within each replicate–system–regime block and then form paired ratios, which we aggregate by their geometric mean. As shown in Table 2, the unchanged thresholds reduce error by 12.1% relative to the fixed configuration. All five independent blocks favor selection (descriptive one-sided exact sign test $p = 0.031$), the 95% block-bootstrap upper ratio is 0.916, and 16 of 18 class-by-regime cells have median ratio at most 1.05. The effect is larger for conservation (20.2%) than for translation (3.2%).

A theory-motivated extension multiplies each nonnegative threshold t_0 by $(\sigma/0.12)\sqrt{96/n_{\mathrm{train}}}$. Relative to the fixed configuration, it reaches an aggregate ratio of 0.891, an upper ratio of 0.926, class ratios below 1.05, 15 of 18 cells within tolerance, and an uncapped ratio of 0.892. Yet it is 1.4% worse than the unchanged selector and is favored in only three of five paired blocks. The simple rescaling is therefore not adopted; the unchanged thresholds define the CNN selector.

Table 2 consolidates these selector ablations with post-training CV, the oracle, the FNO cross-backbone test, a matched learned relaxation, and measured FNO deployment cost.

Per-instance cross-validation and the oracle attain ratios 0.697 and 0.693. Both train all five candidates, selecting by validation error and clean test error, respectively. The pre-training selector reaches 0.879 after measuring the violation and fitting one chosen candidate, establishing a favorable accuracy–search-cost tradeoff.

Cross-backbone FNO test. We repeat the aligned conservation–translation grid with a periodic FNO and five new data-and-training replicates (10,080 fits). The unchanged CNN thresholds preserve a favorable ordering (ratio 0.964, upper 0.978, five of five blocks), but miss the prespecified effect-size and regime-stability criteria. Thresholds fitted solely with FNO calibration-system validation reduce aggregate error by 12.8%, with all 18 cells within tolerance, and match the aggregate error of a same-core learned relaxation coefficient (ratio 0.999) while reducing marginal training cost by about 22%. A paired joint-calibration bootstrap repeats every calibration choice inside each draw and gives a one-sided upper ratio of 0.891 to fixed. Across the 630 matched transfer instances, selection costs 5.14–5.41 A40 GPU hours versus 43.05 for seven-candidate search, an 87–88% reduction. At this measured workload, calibration is amortized over two comparable transfer batches. Appendix B.2 gives the complete sensitivity and cost analysis.

7 DISCUSSION

The useful signal is the broken law itself. Mass and energy drift, parity defects, and rotation defects remain informative across the tested dynamics because they measure the proposed invariance directly. Translation is more subtle. Positional non-stationarity transfers under varying diffusion but becomes nearly flat for reaction-driven breaking. The statistic must therefore match the physical mechanism, especially when one symmetry can fail in several ways.

Violation magnitude organizes the constraint families. Small defects favor the variance reduction of hard enforcement, while large defects favor the expressiveness of an unconstrained model. Soft penalties are most valuable between the two crossovers. The linear–Gaussian analysis makes these three regions explicit and shows how the boundaries depend on noise, sample size, removed capacity, and the violating signal. This structure, rather than a universal numerical threshold, is what the empirical rule carries from one dynamics to another.

The selection principle transfers more readily than its thresholds. CNN-calibrated thresholds preserve a favorable ordering on the FNO, but the effect is smaller and several regime cells change modes. Recalibration on one FNO system per class restores a 12.8% aggregate reduction, with all 18 cells within tolerance. This separation is practically useful: the form of the decision transfers across backbones, while its numerical scale reflects architecture and observation noise. The failed $\sigma/\sqrt{n}$ rescaling reinforces the same point. Noise and sample size matter, but so do removed model capacity and the map from a physical defect to the latent violation coordinate.

Distance from a crossover determines the evidence needed for a decision. In the linear model, an estimate with error smaller than its distance to both crossovers selects the same family as the true violation. Near a crossover, neighboring families have similar risk. Threshold distance therefore motivates a focused validation comparison, but its use with empirical statistics is a deployment heuristic. Transfer errors suggest two remedies: replace a measurement that misses the breaking mechanism, or recalibrate thresholds when the measurement preserves the ordering but the backbone or observation regime changes.

The criterion is most valuable before an expensive candidate sweep. Far from a crossover, the statistic can remove implausible model families without training them. Near a crossover, validation should still compare the neighboring choices. After calibration, the discrete selector avoids 87–88% of exhaustive FNO candidate training. At the measured workload, its calibration cost is recovered over two comparable transfer batches. Learned relaxation has a smaller setup cost, whereas the selector becomes attractive when hard, soft, and unconstrained implementations are reused across many systems.

The present evidence establishes this relationship on periodic grids with controlled breaking, five invariance classes, and two model backbones. Irregular geometries and observational data will require defects adapted to their measurement process. The central principle remains concrete: first measure how the proposed law fails, then decide how strongly the learner should obey it.

8 CONCLUSION

A physical prior need not be accepted or rejected wholesale. Across 12 PDE systems and two backbones, invariance-specific defects locate the useful model between exact enforcement and unconstrained learning. The resulting choices reduce held-out error by 12.1% on the CNN and 12.8% on the calibrated FNO evaluation while replacing a target-level candidate sweep with one model fit. Together with the three-regime analysis, these results make constraint strength a measurable property of the learning problem rather than a fixed architectural preference.

Constraint selection has two transferable components: a defect tied to the physical departure and thresholds that trade this defect against variance and model capacity. A new breaking mechanism calls for checking the defect's sensitivity; a new backbone or observation regime calls for checking its calibration. The translation and FNO experiments distinguish these cases. Comparisons should keep the physical defect and evaluation loss fixed, so that an improvement reflects a better model choice rather than a change in the prediction task.

AI use statement. OpenAI Codex was used for coding assistance and language editing. The authors reviewed all AI-assisted content and take full responsibility for the manuscript, implementation, and reported results.

Reproducibility statement. The main text defines the statistics, selection rule, benchmark, evaluation metric, and statistical unit. The appendix and anonymous release provide the equations, solvers, configurations, evaluation grids, replicate results, fit counts, measured compute, code, fixed splits, and verification commands.

REFERENCES

Matthew Ashman, Cristiana Diaconu, Adrian Weller, Wessel Bruinsma, and Richard E. Turner. Approximately equivariant neural processes. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3078. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b00ef390dcd5f147fd7c5c2bb35f09be-Abstract-Conference.html.

Yohai Bar-Sinai, Stephan Hoyer, Jason Hickey, and Michael P. Brenner. Learning data-driven discretizations for partial differential equations. *Proceedings of the National Academy of Sciences*, 116(31):15344–15349, 2019. doi: 10.1073/pnas.1814058116. URL https://doi.org/10.1073/pnas.1814058116.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G. Wilson. Learning invariances in neural networks from training data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17605–17616, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/cc8090c4d2791cdd9cd2cb3c24296190-Abstract.html.

Tom Beucler, Michael Pritchard, Stephan Rasp, Jordan Ott, Pierre Baldi, and Pierre Gentine. Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, 126(9):098302, 2021. doi: 10.1103/PhysRevLett.126.098302. URL https://doi.org/10.1103/PhysRevLett.126.098302.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural pde solvers. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.03376.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. URL https://arxiv.org/abs/2104.13478.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network. *arXiv preprint arXiv:2109.07103*, 2021. URL https://arxiv.org/abs/2109.07103.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969. PMLR, 2021. URL https://proceedings.mlr.press/v139/elesedy21a.html.

Marc Finzi, Gregory Benton, and Andrew Gordon Wilson. Residual pathway priors for soft equivariance constraints. In *Advances in Neural Information Processing Systems*, volume 34, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/fc394e9935fbd62c8aedc372464e1965-Abstract.html.

Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In *Proceedings of the 38th International*

Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pp. 3318–3328. PMLR, 2021b. URL https://proceedings.mlr.press/v139/finzi21a.html.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/26cd8ecadce0d4efd6cc8a8725cbd1f8-Abstract.html.

Derek Hansen, Danielle C. Maddix, Shima Alizadeh, Gaurav Gupta, and Michael W. Mahoney. Learning physical models that can respect conservation laws. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12469–12510. PMLR, 2023. URL https://proceedings.mlr.press/v202/hansen23b.html.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. doi: 10.1038/s42254-021-00314-5. URL https://doi.org/10.1038/s42254-021-00314-5.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2747–2755. PMLR, 2018. URL https://proceedings.mlr.press/v80/kondor18a.html.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

I. E. Lagaris, A. Likas, and D. I. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, 9(5):987–1000, 1998. doi: 10.1109/72.712178. URL https://doi.org/10.1109/72.712178.

Hannah Lawrence, Elyssa Hofgard, Vasco Portilheiro, Yuxuan Chen, Tess Smidt, and Robin Walters. To augment or not to augment? diagnosing distributional symmetry breaking. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZTb4YmHD9n.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

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. doi: 10.1038/s42256-021-00302-5. URL https://doi.org/10.1038/s42256-021-00302-5.

Kaitlin Maile, Dennis G. Wilson, and Patrick Forré. Equivariance-aware architectural optimization of neural networks. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=a6rCdfABJXg.

Andrei Manolache, Luiz F. O. Chamon, and Mathias Niepert. Learning (approximately) equivariant networks via constrained optimization. *arXiv preprint arXiv:2505.13631*, 2025. URL https://arxiv.org/abs/2505.13631.

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c35f8e2fc6d81f195009a1d2ae5f6ae9-Abstract-Conference.html.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03409.

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. URL https://doi.org/10.1016/j.jcp.2018.10.045.

David W. Romero and Suhas Lohit. Learning partial equivariances from data. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2110.10211.

Sharvaree Vadgama, Mohammad Mohaiminul Islam, Domas Buracas, Christian A. Shewmake, Artem Moskalev, and Erik J. Bekkers. Probing equivariance and symmetry breaking in convolutional networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/afb4b81632998eed0357939db825aa0d-Abstract-Conference.html.

Putri A. van der Linden, Alexander Timans, Dharmesh Tailor, and Erik J. Bekkers. On equivariant model selection through the lens of uncertainty. *arXiv preprint arXiv:2506.18629*, 2025. URL https://arxiv.org/abs/2506.18629.

Tycho van der Ouderaa, David W. Romero, and Mark van der Wilk. Relaxing equivariance constraints with non-stationary continuous filters. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dafd116ac8c735f149558b79fd48e090-Abstract-Conference.html.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23078–23091. PMLR, 2022. URL https://proceedings.mlr.press/v162/wang22aa.html.

Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/45d6637b718d0f24a237069fe41b0db4-Abstract.html.

Jiaxiao Xu, Changhong Mou, Yeyu Zhang, and Fengxiang He. Physics-aligned canonical equivariant fourier neural operator under symmetry-induced shifts. *arXiv preprint arXiv:2605.18606*, 2026. URL https://arxiv.org/abs/2605.18606.

A Benchmark Construction and Experimental Details

A.1 Data generation and governing equations

All systems are posed on periodic domains of unit length, and every trajectory is sampled at $\Delta t = 0.02$ for $T = 12$ steps. One-dimensional initial conditions are sums of Fourier modes 1–6 with independent Gaussian sine and cosine coefficients scaled by inverse frequency. Each field is normalized to unit standard deviation and shifted by an independent constant drawn uniformly from $[0.5, 1.5]$. Two-dimensional initial conditions contain modes through frequency four, with coefficients decaying in total frequency; Navier–Stokes vorticity is also centered to zero mean. We abbreviate mass conservation and translation equivariance as CONS and TRANS, respectively, and use C_4 for rotations by integer multiples of $90°$.

Table 3 lists the 12 controlled-breaking systems. We write $\kappa_\delta(x) = 0.04[1 + \delta \sin(2\pi x)]$ and $m(x) = \sin(2\pi x)$. The calibration/held-out split is made at the level of the governing dynamics, not trajectories. At $\delta = 0$, the candidate invariance associated with each class is exact. Table 3 links descriptive system names to the identifiers used in the accompanying code.

Table 4 specifies the numerical task associated with each governing equation. The listed training sizes and noise levels define the initial five-class survey. The CNN and FNO confirmation studies instead use the 3×3 noise–sample-size matrix in Appendix B.1. With seven breaking strengths and five paired realizations, the initial survey comprises 84 system–breaking tasks and 420 system–breaking–realization blocks.

Table 3: Controlled-breaking benchmark. "Cal." denotes the single calibration dynamics in each invariance class; all other systems are held out from threshold fitting in the initial survey. CONS, TRANS, PARITY, ENERGY, and ROTATION denote conservation of mass, translation equivariance, reflection equivariance, conservation of L^2 energy, and C_4 rotation equivariance.

class	role	system	code identifier	evolution equation
CONS	Cal.	Damped advection (mass)	`cons_adv`	$u_t + u_x = -\delta u$
CONS	held out	Damped Burgers	`cons_burgers`	$u_t + uu_x = 0.1u_{xx} - \delta u$
TRANS	Cal.	Variable-diffusivity diffusion	`trans_diff`	$u_t = \partial_x(\kappa_\delta(x)u_x)$
TRANS	held out	Variable-diffusivity transport	`trans_advdiff`	$u_t = -0.5u_x + \partial_x(\kappa_\delta(x)u_x)$
TRANS	held out	Reaction–diffusion	`trans_react`	$u_t = 0.04u_{xx} - \delta m(x)u$
PARITY	Cal.	Directed advection–diffusion	`par_advdiff`	$u_t = 0.04u_{xx} + \delta u_x$
PARITY	held out	Parity-breaking Burgers	`par_burgers`	$u_t = 0.1u_{xx} - \delta uu_x$
ENERGY	Cal.	Damped advection (energy)	`energy_adv`	$u_t = -u_x - \delta u$
ENERGY	held out	Damped dispersive transport	`energy_advdisp`	$u_t = -u_x + 0.004u_{xxx} - \delta u$
ROTATION	Cal.	Anisotropic diffusion	`rot_aniso`	$u_t = 0.05u_{xx} + 0.05(1+\delta)u_{yy}$
ROTATION	held out	Directed 2D transport	`rot_advec`	$u_t = 0.05\Delta u + \delta u_x$
ROTATION	held out	Anisotropic Navier–Stokes	`ns2d`	$\omega_t + \mathbf{v}\cdot\nabla\omega = \nu_x\omega_{xx} + \nu_y\omega_{yy}$

Table 4: Numerical specification of the controlled-breaking benchmark. Grid, training size n_{train}, observation noise σ, and solver are fixed for each system. Table 3 gives the governing equations and code identifiers. RK4 denotes fourth-order Runge–Kutta with the indicated number of substeps. Every row uses 32 validation and 64 clean test trajectories.

system	role	grid	n_{train}	σ	solver
Damped advection (mass)	Cal.	64	96	.12	Fourier
Damped Burgers	held	64	96	.12	RK4–60
Variable-diffusivity diffusion	Cal.	64	96	.12	RK4–40
Variable-diffusivity transport	held	64	96	.12	RK4–40
Reaction–diffusion	held	64	96	.12	RK4–40
Directed advection–diffusion	Cal.	64	48	.12	Fourier
Parity-breaking Burgers	held	64	48	.12	RK4–60
Damped advection (energy)	Cal.	64	24	.20	Fourier
Damped dispersive transport	held	64	24	.20	Fourier
Anisotropic diffusion	Cal.	32^2	12	.20	Fourier
Directed 2D transport	held	32^2	12	.20	Fourier
Anisotropic Navier–Stokes	held	48^2	24	.15	RK4–16

The linear constant-coefficient systems use exact Fourier propagators. Variable-coefficient diffusion and reaction use pseudo-spectral spatial derivatives and fourth-order Runge–Kutta with 40 substeps per observed step. The two Burgers systems use 60 substeps and two-thirds dealiasing. For Navier–Stokes, $-\Delta\psi = \omega$, $\mathbf{v} = (\partial_y\psi, -\partial_x\psi)$, $\nu_x = 0.008$, and $\nu_y = 0.008(1 + 4\delta)$; a dealiased pseudo-spectral RK4 solver uses 16 substeps. The other two-dimensional systems have exact Fourier propagators.

One-dimensional fields use 64 grid points, the linear two-dimensional fields use 32×32, and Navier–Stokes uses 48×48. For every configuration we generate independent training, validation, and clean test trajectories; validation and test contain 32 and 64 initial conditions, respectively. The training size is varied as specified in each study. Independent Gaussian noise with standard deviation σ times the global standard deviation of the clean training trajectories is added to both training and validation trajectories. Test trajectories remain clean. Thus validation represents the same noisy observation process used for fitting, whereas reported rollout error is measured against the underlying dynamics.

A.2 Constraint implementations

The convolutional neural network (CNN) flow map is residual. In 1D it has four width-48 circular convolutional layers with kernel size five and GELU activations, followed by a scalar-output convolution. The 2D version uses three width-24 circular layers with the same kernel size. The hard and soft constructions are:

- **Mass:** hard subtracts the spatial mean from the predicted increment; soft penalizes the squared sum of the increment.
- **Translation:** hard uses only the circular CNN. The unconstrained and soft models add a second branch receiving the field and eight periodic position channels. Soft penalizes the squared correction produced by this position-dependent branch.

- **Parity:** hard symmetrizes every convolutional kernel under reflection. Soft penalizes disagreement between the update of a reflected input and the reflected update of the original input.
- **Energy:** hard rescales the predicted next state to preserve $\sum_i u_i^2$. Soft penalizes its squared relative energy change.
- **Rotation:** hard replaces the update f by the group average $|C_4|^{-1} \sum_{g \in C_4} g^{-1} f(gu)$. Soft averages the squared defects for the three nonidentity rotations.

For the initial CNN survey and robustness confirmation, the soft grid is $\{0.01, 0.1, 1\}$. Every model is optimized for 250 epochs with Adam, learning rate 2×10^{-3}, batch size 64, and the sum of one-step mean-squared error and the weighted constraint penalty. There is no early stopping or test-dependent selection of saved model weights.

The cross-backbone model is a 1D periodic Fourier neural operator (FNO) with width 24, four Fourier layers, 12 retained modes, and a pointwise residual path in each layer. Its input has one state channel and eight periodic coordinate channels. All hard, soft, and unconstrained candidates have 59,185 parameters. For conservation, hard projects the FNO increment to zero mean and soft uses the same mass penalty as the CNN. For translation, hard zeros the coordinate channels, the unconstrained model uses them, and soft penalizes the equivariance defect under a fixed quarter-domain shift. The matched learned-relaxation baseline uses $f_{\mathrm{hard}} + \alpha(f_{\mathrm{free}} - f_{\mathrm{hard}})$ with one jointly trained sigmoid mixture weight, for 59,186 parameters. FNO candidates use the soft grid $\{0.001, 0.01, 0.1, 1, 10\}$ and are trained for 250 epochs with AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, and cosine decay to 10^{-5}.

Table 5: Comparators used in the confirmation studies. "Target fits" is the number of candidates trained for one held-out task after class-level calibration. The criterion and fixed policies select a candidate before target training, whereas cross-validation (CV) and the oracle train the full discrete pool.

backbone	policy	pool	target fits	selection information	training
CNN	fixed configuration	5	1	calibration validation	Adam, 250 epochs
CNN	violation criterion	5	1	training statistic and calibrated thresholds	Adam, 250 epochs
CNN	per-instance CV	5	5	target validation after all fits	Adam, 250 epochs
CNN	test oracle	5	5	clean target test; reference only	Adam, 250 epochs
FNO	fixed configuration	7	1	calibration validation	AdamW, 250 epochs
FNO	violation criterion	7	1	training statistic and calibrated thresholds	AdamW, 250 epochs
FNO	learned mixture	1	1	mixture weight learned with shared core	AdamW, 250 epochs
FNO	exhaustive CV	7	7	target validation after all fits	AdamW, 250 epochs
FNO	test oracle	7	7	clean target test; reference only	AdamW, 250 epochs

CNN confirmation models contain 35,233 parameters except the translation free and soft branches, which contain 72,386. The two-dimensional initial-survey CNNs contain 30,073 parameters. All discrete FNO candidates contain 59,185 parameters; the adaptive model contains 59,186.

A.3 Calibration, evaluation, and statistical unit

For each system and breaking level, the initial-study statistic is averaged over five paired data-and-training replicates. Within an invariance class, candidate thresholds are placed at observed statistic values and at one point below the observed range, allowing the hard region to be empty. The selected pair minimizes cumulative clean rollout error on the calibration system. Family errors are minimized over the three soft penalties, so this experiment characterizes hard/soft/none selection rather than a deployable penalty rule. Table 6 gives the resulting CNN thresholds.

Table 6: CNN thresholds fitted in the initial survey. A dash means that the hard region lies below the observed nonnegative statistic range.

invariance class	t_1	t_2
mass conservation	–	0.003668
translation	0.027631	0.066940
parity	–	0.004788
energy	0.013232	0.028533
rotation	0.0000870	0.004249

In the confirmation studies, noisy calibration-system validation rollout selects both the soft penalty and the fixed comparator. Held-out test error is not used for either choice. The FNO-specific threshold

refit also minimizes calibration validation error alone. Per-instance cross-validation and the clean-test oracle provide reference points that use target validation or test information unavailable to the pre-training selector.

For each replicate, system, class, and observation regime, capped rollout error is averaged over all seven breaking levels before a paired policy ratio is formed. Aggregate ratios are geometric means of these block-level ratios. The 95% upper bound is the 95th percentile of 100,000 bootstrap geometric means obtained by resampling the five entire replicate blocks. A class-by-regime cell is within tolerance when the median of its five block ratios is at most 1.05. Uncapped errors provide a direction check. The cap of one applies to initial-survey family scores and confirmatory policy ratios, but not to the training loss.

Table 7: Metrics and statistical definitions used throughout the analysis.

item	fixed definition
prediction task	one-step flow map, evaluated autoregressively
primary metric	mean relative ℓ_2 error over a 10-step clean-test rollout
observations	independent noisy train/validation trajectories; clean test trajectories
broad task unit	one system, breaking strength, and paired realization
confirmation block	one class, held-out system, noise, training size, and realization, averaged over seven breaking strengths
policy ratio	paired selected-policy error divided by fixed-policy error
aggregate	geometric mean of paired ratios over entire blocks
uncertainty	one-sided 95% block-bootstrap upper bound; calibration is refitted inside the joint FNO sensitivity bootstrap
cell criterion	median paired ratio at most 1.05 within a class-by-regime cell
capping	rollout error capped at one for policy aggregation; uncapped direction retained; training loss never capped

Across all stages, the experiments comprise 19,005 model fits: 2,100 in the initial survey, 6,825 in the CNN confirmation (including the reaction-mechanism control), and 10,080 in the FNO cross-backbone study. Every fit returned finite validation and test metrics.

B Supplementary Experimental Results

B.1 CNN confirmation across noise and sample size

The CNN confirmation follows the validation-based selection and block-level aggregation protocol in Appendix A. Calibration fixes one soft strength from $\{0.01, 0.1, 1\}$ per class and observation regime before transfer to the aligned held-out system.

The aligned design uses damped advection and variable-diffusivity diffusion for calibration, with damped Burgers and variable-diffusivity transport as the corresponding held-out systems. It crosses these systems with noise levels $\{0.06, 0.12, 0.20\}$, training sizes $\{48, 96, 192\}$, seven breaking strengths $\{0, 0.02, 0.05, 0.10, 0.20, 0.40, 0.80\}$, five independent data-and-training replicates, and five candidate configurations. These factors yield 6,300 aligned fits. A mechanism-shift control using reaction–diffusion with a spatially varying reaction term adds 525 fits at $(0.06, 192)$, $(0.12, 96)$, and $(0.20, 48)$.

Table 8 summarizes the aggregate policy comparisons.

Table 8: Aggregate capped-error ratios to the calibration-selected fixed configuration. The upper bound is computed by resampling entire replicate blocks.

policy	all	CONS	TRANS	95% upper	cells
unchanged thresholds	0.87896	0.79836	0.96769	0.91579	16/18
theory rescaled	0.89121	0.81406	0.97568	0.92562	15/18
calibration refit	0.86096	0.82653	0.89683	0.88702	18/18
per-instance CV	0.69705	0.57154	0.85012	0.72373	18/18
oracle	0.69288	0.56940	0.84313	0.71861	18/18

The theory-rescaled policy satisfies all five prespecified comparisons with the fixed configuration: its aggregate ratio is below 0.95, its one-sided upper bound is below 1, both class ratios are at most 1.05, 15 of 18 regime cells satisfy the tolerance, and its uncapped aggregate ratio is favorable (0.89184). The unchanged thresholds nevertheless perform better overall and are therefore retained. Their direct comparison gives a rescaled-to-unchanged ratio of 1.01394, with rescaling favored in three of five

replicate blocks; these data do not support the scaling sub-hypothesis. Because all five blocks favor the unchanged selector over fixed, the descriptive one-sided exact sign-test probability under equal win odds is $2^{-5} = 0.03125$.

Table 9 reports the corresponding class-by-regime results.

Table 9: Median policy-to-fixed error ratio in every aligned class-by-regime cell. Values above the prespecified tolerance 1.05 are bold.

class	noise	n_{train}	unchanged	rescaled
CONS	0.06	48	1.040	1.040
CONS	0.06	96	0.901	0.874
CONS	0.06	192	**1.918**	**1.177**
CONS	0.12	48	0.489	0.522
CONS	0.12	96	0.642	0.642
CONS	0.12	192	0.532	0.532
CONS	0.20	48	0.821	0.767
CONS	0.20	96	0.721	0.721
CONS	0.20	192	0.860	0.808
TRANS	0.06	48	0.974	**1.062**
TRANS	0.06	96	0.734	0.965
TRANS	0.06	192	0.946	0.988
TRANS	0.12	48	1.001	0.926
TRANS	0.12	96	0.920	0.920
TRANS	0.12	192	0.864	0.889
TRANS	0.20	48	**1.241**	1.024
TRANS	0.20	96	1.048	**1.061**
TRANS	0.20	192	1.004	1.011

The unchanged selector has two cell-level exceptions: low-noise/high-data CONS and high-noise/low-data TRANS. Rescaling substantially reduces both, but introduces smaller exceptions at low-noise/low-data TRANS and high-noise/mid-data TRANS. This redistribution explains why rescaling can satisfy the rescaled-versus-fixed primary test without improving the nominal selector overall.

Table 10 gives the paired transfer ratios in each replicate.

Table 10: Geometric aligned-transfer ratio in each entire replicate block. The final column compares the rescaled and unchanged selectors.

replicate	unchanged/fixed	rescaled/fixed	rescaled/unchanged
1	0.8310	0.8541	1.0278
2	0.8428	0.9362	1.1108
3	0.9622	0.9527	0.9902
4	0.9073	0.8724	0.9615
5	0.8580	0.8460	0.9859

For the reaction–diffusion system with spatially varying reaction, the theory-rescaled selector has geometric ratio 1.0043 to the calibration-selected fixed configuration, which is effectively neutral. The violation estimate is less aligned with the hard-minus-none gap for reaction-type breaking than for aligned diffusion-type breaking. This control therefore limits the translation result to mechanisms observed by the non-stationarity statistic.

The globally pooled rank correlation between the estimated violation statistic $\hat{\delta}$ and the hard-minus-none gap is -0.101, reflecting the mixture of two estimator scales and nine observation regimes. A post hoc stratified analysis gives positive within-cell correlations in all 90 aligned replicate–system–regime blocks, with median 0.964. Because the stratification was introduced after the primary aggregation, this result is interpreted descriptively and does not replace the prespecified pooled summary.

All 6,825 planned fits returned finite validation and test metrics, using 27.61 NVIDIA A40 GPU hours. Thus the aggregate comparisons include the full design without replacement or omission of model fits.

B.2 Cross-backbone confirmation with FNO

The cross-backbone confirmation uses the periodic 1D FNO and matched adaptive baseline specified in Appendix A, with a shared FNO core, optimizer, and training schedule across candidates.

The design crosses conservation and mechanism-aligned translation with calibration and held-out systems, three noise levels, three training sizes, seven breaking strengths, five new data-and-training replicates (201–205), and eight candidate models: hard, none, five soft penalties $\{.001, .01, .1, 1, 10\}$, and the adaptive model. All 10,080 planned fits returned finite metrics. The primary selector applies the CNN thresholds unchanged. FNO calibration-system validation selects the soft strength and fixed comparator; the secondary FNO threshold refit uses the same calibration data. No threshold or penalty is selected using held-out test results.

Table 11: FNO aggregate capped-error ratios. "Fixed" is the calibration-selected fixed configuration; "adaptive" is the matched one-parameter mixture.

policy	fixed	CONS	TRANS	95% upper	cells	adaptive
unchanged CNN thresholds	0.96449	0.96867	0.96032	0.97789	12/18	1.10537
FNO calibration refit	0.87191	0.84162	0.90328	0.87733	18/18	0.99927

Applying the CNN thresholds unchanged reduces aggregate FNO error by 3.6%, with an upper ratio of 0.97789 and favorable results in all five replicate blocks. The magnitude is smaller than the prespecified 5% effect criterion, 12 rather than 15 of 18 regime cells meet the cell tolerance, and the adaptive comparison is 1.10537. This pattern supports transfer of the family ordering, but not direct reuse of the numerical CNN thresholds. By contrast, the calibration-only FNO refit reduces error by 12.8%, covers all 18 regime cells, and matches the adaptive model in aggregate. The statistic-based ordering therefore remains informative after the backbone change when the numerical thresholds are recalibrated.

Calibration and task-mixture sensitivity. The dependence of the secondary FNO-specific refit on the finite calibration sample is assessed retrospectively. All $5^5 = 3{,}125$ nonparametric resamples of the five paired calibration–transfer blocks are enumerated. Within each draw, the soft penalty, fixed comparator, and two thresholds are re-estimated on resampled calibration blocks before evaluation on the corresponding transfer blocks. The resulting intervals propagate calibration uncertainty and are interpreted as retrospective sensitivity evidence.

Table 12: Retrospective joint calibration–transfer bootstrap. "Upper" is the one-sided 95% nearest-rank bound after refitting all calibration choices in every draw. The 1.05 value is the existing adaptive-comparison tolerance.

comparison	point ratio	95% upper	reference
FNO refit / fixed	0.87191	0.89054	1.00
FNO refit / adaptive	0.99927	1.00834	1.05

Refitting within each draw increases the refit-to-fixed upper bound from the conditional value 0.87733 to 0.89054, which remains below one. The refit-to-adaptive upper bound is 1.00834, below the 1.05 reference margin. The reduction relative to fixed and the adaptive-level aggregate accuracy are therefore stable to resampling of the calibration blocks.

The primary aggregation weights the seven breaking levels equally. A complementary retrospective analysis reweights the same errors to represent different task mixtures. Table 13 gives integer weights in order of increasing breaking strength; only their normalized values affect the ratios.

Table 13: Retrospective deployment-prior sensitivity. Every illustrative prior favors the FNO refit over the calibration-selected fixed configuration; its comparison with adaptive relaxation depends on the task mixture.

task mixture	weights over increasing δ	refit/fixed	refit/adaptive
uniform	$1, 1, 1, 1, 1, 1, 1$	0.872	0.999
near-exact	$40, 22, 14, 10, 7, 4, 3$	0.845	0.897
middle-heavy	$5, 10, 20, 30, 20, 10, 5$	0.924	1.018
strong-heavy	$3, 4, 7, 10, 14, 22, 40$	0.861	1.081
endpoints	$35, 5, 5, 5, 5, 10, 35$	0.808	0.983

Across these task mixtures, refit/fixed ranges from 0.808 to 0.924. Relative to adaptive relaxation, the discrete selector is strongest when tasks concentrate near exact invariance or the two ends of the breaking range, while adaptive relaxation is stronger under middle- or strong-breaking mixtures.

Tables 14 and 15 report the class-by-regime and replicate-level results; Table 16 records the measured computational cost.

Table 14: Every FNO class-by-regime cell. t_1, t_2 are the secondary thresholds selected on FNO calibration-system validation; λ is the calibration-selected soft penalty. The last three columns are median ratios to the calibration-selected fixed configuration. A dash means that the fitted hard region lies below the observed statistic range.

class	noise	n_{train}	t_1	t_2	λ	unchanged	refit	adaptive
CONS	.06	48	.001029	.001029	.001	1.1613	.9453	.8438
CONS	.06	96	.000932	.000932	.001	1.1831	.9437	.9136
CONS	.06	192	.000355	.000355	.001	1.1890	.9394	.7672
CONS	.12	48	.003901	.003901	.001	1.0608	.9280	.8884
CONS	.12	96	.001704	.001704	.001	1.0315	.9099	.8783
CONS	.12	192	.001745	.001745	.001	1.0525	.9208	.7934
CONS	.20	48	.007486	.007486	.01	.8239	.7142	.7122
CONS	.20	96	.007129	.007129	.01	.6477	.6756	.6715
CONS	.20	192	–	.007187	.1	.6939	.6939	.8150
TRANS	.06	48	.05109	.12658	.01	.9930	.9120	.9077
TRANS	.06	96	.02469	.37055	.001	.9148	.8894	.8805
TRANS	.06	192	.01366	.36371	.001	1.1187	.9035	.8734
TRANS	.12	48	.05921	.25422	.01	1.0045	.8508	.9316
TRANS	.12	96	.02932	.22191	.01	.8661	.8655	.9289
TRANS	.12	192	.01693	.21485	.01	.9132	.8921	.9118
TRANS	.20	48	–	.07728	1	.9135	.9081	1.0246
TRANS	.20	96	.03361	.11370	.1	.9898	.9369	1.0484
TRANS	.20	192	.01990	.03142	.1	.9427	.9368	1.0391

Table 15: FNO geometric paired ratio in each fresh replicate block.

replicate	unchanged/fixed	unchanged/adaptive	refit/fixed	refit/adaptive
201	.95658	1.06465	.88111	.98066
202	.94411	1.10521	.87481	1.02409
203	.98603	1.10880	.87622	.98531
204	.98237	1.14662	.86857	1.01379
205	.95406	1.10312	.85898	.99318

Table 16: Measured FNO cost on NVIDIA A40. Calibration is a one-time cost; the transfer rows compare deployment policies on the identical 630 instances.

quantity	measured time
discrete calibration (seven candidates)	46.302 A40 GPU h
held out: selected, unchanged thresholds	5.145 A40 GPU h
held out: selected, FNO calibration refit	5.414 A40 GPU h
held out: adaptive baseline	6.910 A40 GPU h
held out: all seven discrete candidates	43.048 A40 GPU h
violation-statistic estimation	0.132 CPU h
selected-policy evaluation	0.00503 GPU h
all FNO jobs, including overhead	106.031 A40 job h

Relative to seven-candidate transfer search, selected-model training is reduced by 88.0% with unchanged thresholds and by 87.4% with the FNO refit. These times cover all 630 transfer instances. Let m count comparable batches of 630 instances at the measured throughput, with calibration reused. The refit then costs $46.302 + 5.414m$ A40 GPU hours, whereas exhaustive search costs $43.048m$. Calibration is amortized at the second batch when the invariance classes, observation regimes, and backbone permit threshold reuse. The 87.4% saving concerns candidate training after calibration. Relative to adaptive relaxation, marginal cost is about 22% lower, and calibration is amortized at batch 31. Below this workload, adaptive relaxation is less expensive when a compatible mixed architecture is available.

B.3 Initial-survey transfer and translation sensitivity

The broad five-class study minimizes over three soft strengths when scoring the hard/soft/none family choice. Table 17 gives system-level results together with validation-selected cross-validation and the clean-test oracle.

Table 17: System-level results from the initial survey. Errors are capped 10-step rollout means over all seven breaking levels. "Fixed" is the best single hard, soft, or unconstrained family for the entire system. CV validates a family separately at each breaking level after training all candidates; the oracle uses clean test error. Ratio is criterion/fixed, and bold criterion entries improve upon or match the best fixed family.

system	role	criterion	fixed	CV	oracle	ratio	match
Damped advection (mass)	Cal.	**.0562**	.0602	.0562	.0562	.934	100%
Damped Burgers	held	.0292	.0218	.0184	.0184	1.337	86%
Variable-diffusivity diffusion	Cal.	**.0355**	.0373	.0355	.0355	.951	100%
Variable-diffusivity transport	held	**.0368**	.0387	.0366	.0366	.951	86%
Reaction–diffusion	held	.0369	.0359	.0330	.0330	1.026	71%
Directed advection–diffusion	Cal.	**.0204**	.0206	.0204	.0204	.990	100%
Parity-breaking Burgers	held	**.0157**	.0162	.0153	.0153	.972	57%
Damped advection (energy)	Cal.	**.1427**	.1442	.1449	.1427	.990	100%
Damped dispersive transport	held	.6376	.5831	.5831	.5831	1.094	86%
Anisotropic diffusion	Cal.	**.0172**	.0181	.0172	.0172	.955	100%
Directed 2D transport	held	**.0197**	.0202	.0193	.0193	.976	57%
Anisotropic Navier–Stokes	held	**.0438**	.0447	.0433	.0433	.980	57%

The criterion improves on the best system-wide fixed family in all five calibration systems and four of seven governing-equation-level transfers. In the latter group it lowers the aggregate error from 0.1375 to 0.1171 while remaining close to the profiled oracle on translation with varying diffusion, parity, rotational advection, and Navier–Stokes. Exact mode agreement is not required for low excess error: the two held-out rotation systems have 57% mode match but absolute oracle gaps of only 0.0004 and 0.0006.

Translation is harder because its non-stationarity statistic measures one mechanism by which shift equivariance can fail. It transfers cleanly from variable-diffusivity diffusion to variable-diffusivity transport, where the mechanism remains spatially varying diffusion, but it has limited sensitivity to the spatially varying reaction in the reaction–diffusion system. Figure 3 shows that the complementary sufficiency-gap estimator responds to reaction-type breaking at large δ while remaining near zero on shift-equivariant data; it is less sensitive in the soft-transition regime.

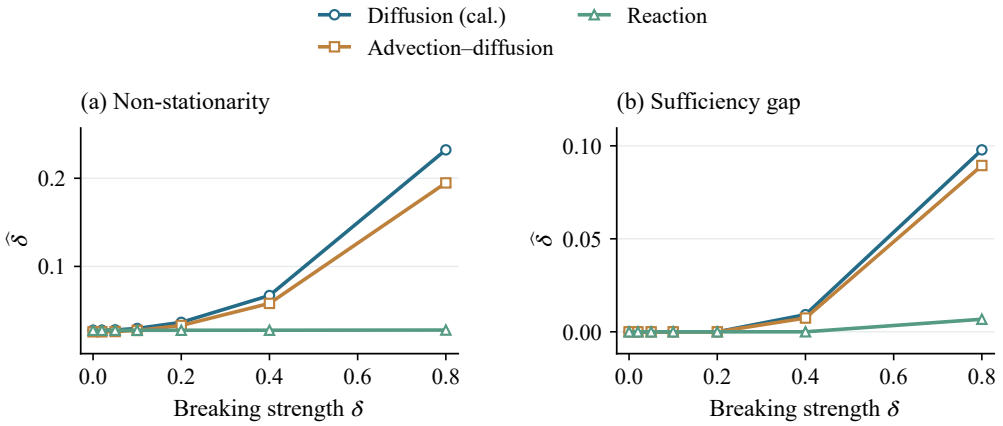

Figure 3: Translation estimators under diffusion- and reaction-type symmetry breaking. The sufficiency gap captures strong reaction-type violation.

C Reference Solution Geometry

The following visualizations show the underlying PDE solutions, complementing the aggregate prediction errors. For each of five fixed systems, we take the first clean test initial condition at public seed zero and compare $\delta = 0$ with $\delta = 0.8$. Each pair uses identical initial data. The fields are reference-solver outputs, not learned predictions or additional performance trials. The initial conditions follow the benchmark distribution in Appendix A, with no selection by prediction error.

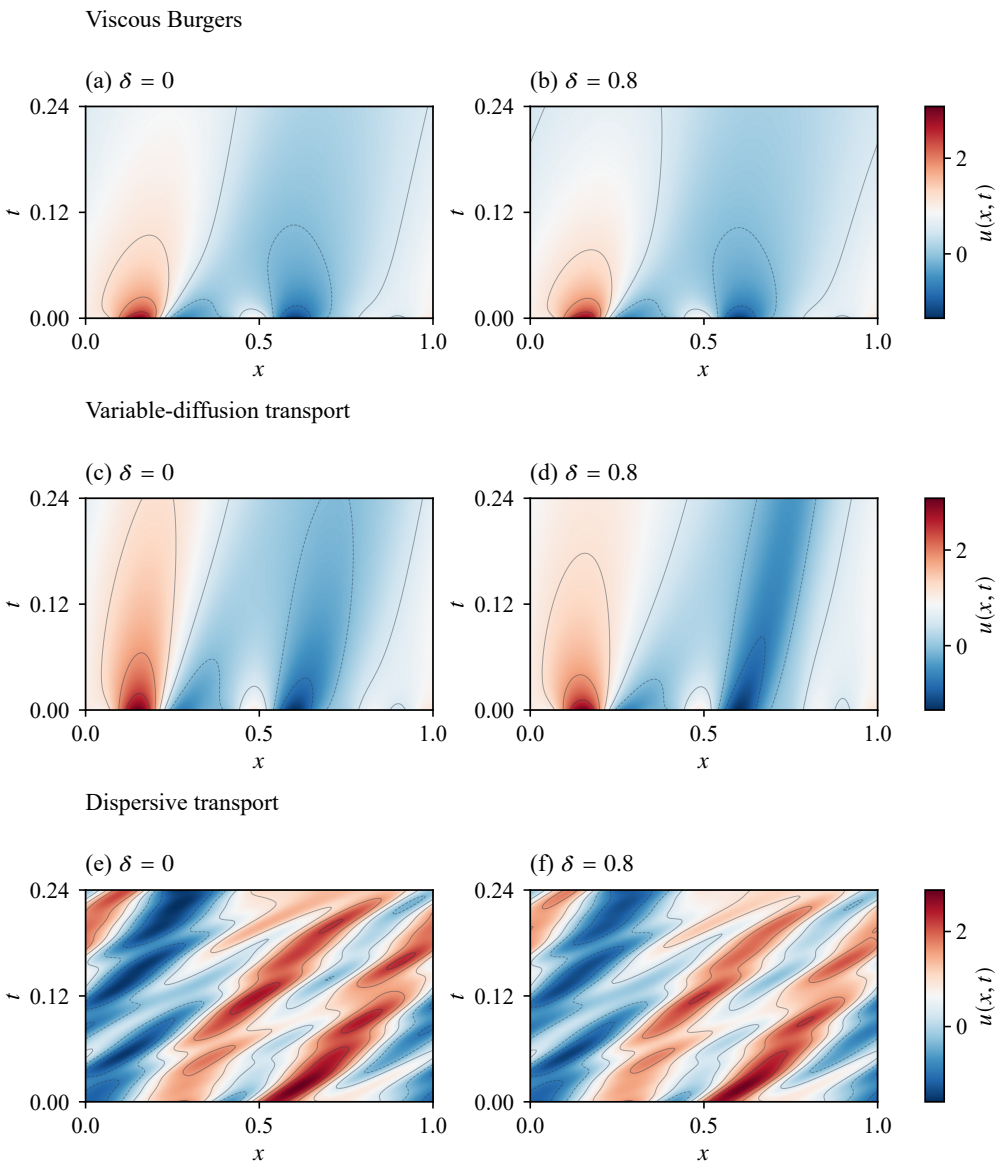

Figure 4: **One-dimensional reference solutions across breaking strengths.** Rows show viscous Burgers, transport with spatially varying diffusion, and dispersive transport. Thin contours resolve field structure, and each row shares one color scale. The left and right columns use $\delta = 0$ and 0.8. The fields are evaluated at 256 spatial points and 121 times over $[0, 0.24]$. Mass decays in Burgers and energy decays in dispersive transport when damping is present. Spatially varying diffusion changes the transport profile.

The initial fields are Fourier-resampled from the benchmark grid. The one-dimensional solvers retain the original equations and evaluate finer output steps of 0.002. Halving this step changes the plotted trajectories by less than 10^{-10} in relative ℓ_2 norm. The Burgers mass and dispersive-transport energy satisfy their respective exponential decay laws to within 10^{-10} in absolute error.

Two-dimensional morphology. Figure 5 compares two rotation-breaking mechanisms at $t = 0.12$. Anisotropic diffusion damps spatial directions differently, whereas the preferred-direction term transports the field along one axis. The surfaces and their projected contours display the same scalar solution. A common color range and vertical scale make changes in amplitude and position comparable across panels.

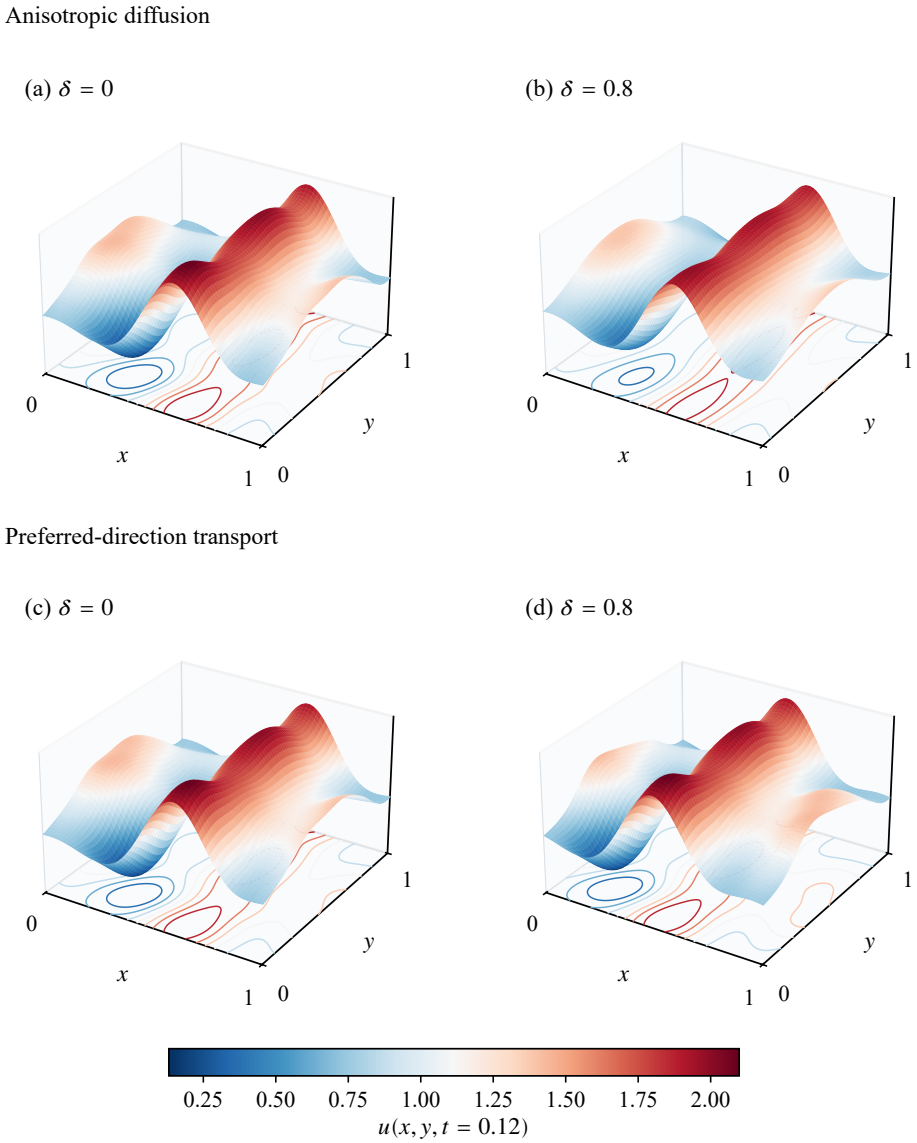

Figure 5: **Two-dimensional reference fields and projected contours.** Top: $u_t = 0.05u_{xx} + 0.05(1 + \delta)u_{yy}$. Bottom: $u_t = 0.05\Delta u + \delta u_x$. Columns compare $\delta = 0$ and 0.8 with identical initial data. Surface height and color encode $u(x, y, 0.12)$, while the contour projection reveals the locations and shapes of extrema. All panels use the same scale. These are reference solutions, not model forecasts.

The two-dimensional fields use the exact Fourier propagators already present in the benchmark, evaluated on a 128×128 grid. Their values at the original 32×32 nodes agree with the original-grid propagator to within 10^{-12}. This agreement checks the finer-grid evaluation against the benchmark propagator. The illustrations describe solution geometry and do not constitute additional training or prediction-error experiments.