LEARNING WHEN TO TRUST PDE COUPLING PRIORS IN NEURAL OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Coupled neural operators must decide how strongly to use interactions suggested by the governing equations. We study this structural decision under finite data using controlled interventions that vary cross-field coupling while holding the operator backbone and evaluation fixed. On Active Matter and a turbulent radiative layer, fixed PDE coupling priors exhibit different data-scaling behavior, and leave-one-pathway-out interventions show that the predictive effect of individual interactions can change sharply with the data regime. Motivated by this observation, we factor each cross-field interaction into a learned operator and a scalar pathway strength. A short warmup provides pathway-specific evidence to initialize these strengths before joint training. On scarce Active Matter data, the resulting model reduces relative error from 7.06% for dense coupling and 6.77% for the fixed PDE graph to 6.44%. Across Active Matter and the turbulent radiative layer, the same method improves on the strongest fixed structural baseline in all four scarce and medium regimes and remains within 0.08 percentage points of the best fixed choice in both abundant regimes. With the method frozen before evaluation on Rayleigh–Bénard convection, it improves scarce-data error from 0.0872 to 0.0831 ± 0.0009 and matches the strongest fixed baseline with abundant data. These results support calibrating PDE coupling structure at the level of individual pathways using target-regime evidence.

1 INTRODUCTION

Neural operators learn mappings between function spaces and now serve as practical surrogates for parameterized partial differential equations (Li et al., 2021; Lu et al., 2021; Kovachki et al., 2023). In coupled multiphysics systems, the learning problem has a distinctive structure: governing equations already expose some dependencies among physical fields through transport, source, and constitutive mechanisms, while the full solution operator must still be learned from data. A model can hard-wire these known dependencies as an interaction graph, or it can leave all cross-field interactions to be inferred. Both choices are common, and the choice is usually made by convention.

The familiar intuition is that structure helps most when data is scarce: a correct prior constrains a small training set, and a flexible model catches up as data grows. Whether this intuition is correct matters for how coupled operators should be built, but it is difficult to test directly. Comparing a physics-guided model against an unconstrained baseline changes the interaction structure, the sparsity, and the parameter count at once, so an observed gap cannot be attributed to the interaction information itself.

We make the interaction information the independent variable. Using a spectral operator in which each cross-field pathway can be independently imposed, removed, or mismatched while the rest of the architecture is held fixed (Sec. 3), we measure how the value of prescribed coupling structure changes with the training-set size on two controlled systems and one prospective one.

The measurements do not support the intuition. On Active Matter, the performance separation between the PDE-derived graph and equally sparse mismatched graphs grows from $+1.7$ to $+15.7$ percentage points as training data increases; on a turbulent radiative layer the same separation shrinks from $+8.2$ to $+3.5$ points (Fig. 1). Data quantity alone therefore does not determine the value of a fixed coupling prior. Leave-one-pathway-out interventions show why: individual pathways differ

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.

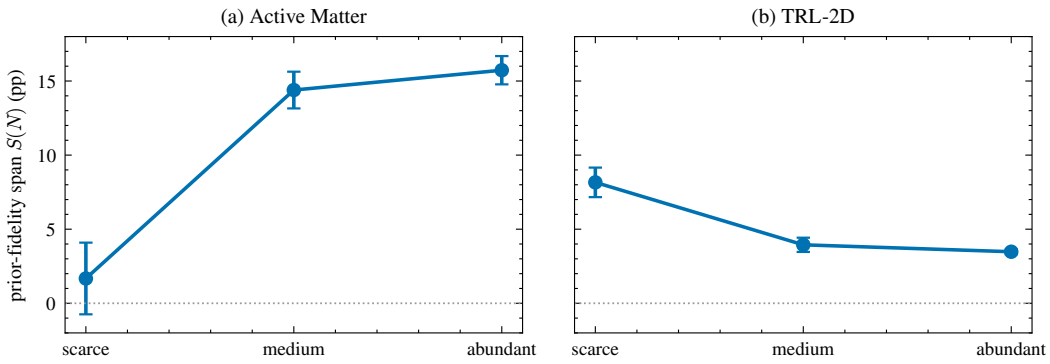


Figure 1: **Fixed PDE coupling priors scale differently with data across systems.** Prior-fidelity span $S(N)$, the validation-selected test-error separation between the PDE coupling graph and equally sparse mismatched graphs (mean $\pm$ s.d. over 3 seeds), against the training-data regime. On Active Matter the span grows with data ($\Gamma = S(\text{scarce}) - S(\text{abundant}) = -14.1$ pp); on the turbulent radiative layer it shrinks ($\Gamma = +4.7$ pp). A universal scarcity rule—structure is worth more when data is scarce—cannot explain both trends.

by more than an order of magnitude in marginal effect, and those effects move with the data regime: one pathway per system grows steadily in importance while others remain inert (Fig. 2).

The validation follows the diagnosis. The gated operator improves the strongest fixed baseline in the most discriminating scarce and medium regimes and remains close to it when abundant data already make a fixed structural choice effective. Frozen before any contact with a third physical system, it transfers to Rayleigh–Bénard convection, where it improves over the strongest fixed structure by 4.7% in the scarce regime. CoDA-NO provides a complementary learned-interaction comparison: it performs best on TRL-scarce, whereas M3 is substantially stronger on TRL-abundant; increasing CoDA-NO from 1.8M to 7.24M parameters does not close the abundant-regime gap.

This observation motivates treating pathway strength as something to be learned, not prescribed. We represent each cross-field interaction as $g_e \mathcal{K}_e$: a scalar structural strength $g_e \in [0, 1]$ multiplying a learned spectral interaction operator $\mathcal{K}_e$ (Sec. 4). The scalars are initialized from a short-warmup probe that measures each pathway's marginal utility at two thousand training steps, then refined jointly with the operator. The method adds six pathway gates (nine scalars allocated) over the fully dense model.

Our contributions are:

1. **Empirical characterization.** A controlled study, with capacity and edge budget held fixed, showing that the finite-data benefit of fixed PDE coupling structure varies across systems and individual pathways, so data scarcity alone does not predict prior value.
2. **Evidence-informed pathway coupling.** A lightweight factorization that separates pathway strength from interaction-function learning and initializes structural strength from short-warmup evidence.
3. **Cross-system validation.** Matched structural controls, prior-mismatch experiments, an external learned-interaction baseline, and prospective transfer to a frozen Rayleigh–Bénard system.

2 Related Work

Neural operators and coupled systems. Fourier neural operators and their extensions learn resolution-independent solution operators for PDE families (Li et al., 2021; Kovachki et al., 2023; Li et al., 2023; Liu-Schiaffini et al., 2024). Coupled and multiphysics variants must additionally represent interactions among fields. CoDA-NO couples variables through codomain attention, learning an input-conditioned interaction weight per variable pair (Rahman et al., 2024); we treat it as the

strongest learned-interaction baseline and compare directly in Sec. 5.4. CMWNO decouples coupled integral kernels into per-pair multiwavelet blocks (Xiao et al., 2023), giving each field pair its own learned interaction operator but no explicit, evidence-adjustable strength. Work on DeepONet organization shows that the preferred branching structure depends on coupling strength (Park et al., 2026). These approaches choose or learn an interaction form; we instead measure when prescribed interactions help and make their per-pathway strength explicit and learnable. The two views are complementary: codomain attention decides *how* fields exchange information inside a flexible architecture, while our factorization decides *how much* each physically nominated pathway should be trusted at the available data scale.

Physical structural priors. Physics-informed neural operators add governing-equation residuals to the training objective (Li et al., 2024); other work embeds Green-function structure (Tan et al., 2026) or fundamental PDE knowledge (Ma et al., 2026) to improve data efficiency. These methods encode structure as soft constraints or architectural motifs and generally report that structure helps, particularly at low data. Our results add a qualification: whether a given structural pathway helps depends on the system, the pathway, and the data regime, measured under controls that hold capacity and edge budget fixed.

Learned interaction structure. Attention and sparse-gating mechanisms can adapt interaction strength to data (Rahman et al., 2024); causal discovery over trajectories infers effective coupling graphs for graph-network simulators (Wang et al., 2026). The present work differs in what is learned: not the interaction function itself, nor a discovered graph, but a low-dimensional trust weight per physically nominated pathway, initialized from cheap training evidence.

3 Fixed Coupling Priors Across Data Regimes

3.1 Controlled setup

We study one-step prediction of coupled field dynamics on three systems: Active Matter (AM) and a two-dimensional turbulent radiative layer (TRL) from The Well (Ohana et al., 2024), and Rayleigh–Bénard convection (RBC) as a prospective system. Fields are partitioned into physically meaningful groups (AM: velocity U, concentration C, active-stress-derived field D; TRL: radiation R, pressure P, velocity U; RBC: buoyancy B, pressure P, velocity U). The governing equations nominate a directed *explicit-mechanism graph* $G_{\mathrm{PDE}} = (V, E_{\mathrm{PDE}})$, where an edge $j \to i$ records that group j enters the modeled dynamics of group i (Appendix A). Table 1 summarizes the three systems.

Table 1: Systems, field groups, and pathway structure. E_{PDE} counts the explicit-mechanism edges; candidates are all $K(K-1)$ cross-field pathways.

System	Groups	E_{PDE}	Candidates	Regimes (configs)
Active Matter	U, C, D	4/6	6	5 / 18 / 45
TRL-2D	R, P, U	5/6	6	2 / 4 / 9
Rayleigh–Bénard	B, P, U	3/6	6	1 / 4

The backbone is a typed spectral operator: each group is lifted to width 32, processed by four Fourier layers retaining 12×12 modes, and projected back (Li et al., 2021). Within each layer, a self-field spectral block is always active, and each prescribed cross-field pathway contributes its own spectral block. All cross-field blocks share an identical spectral shape, so the edge count is an exact parameter budget; the fully connected variant (Dense, all $K(K-1)$ pathways) and a width-matched Dense variant differ by at most 0.08 pp at every regime, so capacity does not confound structural comparisons. NoCross removes all cross-field pathways. Mismatched priors are drawn from the feasible symmetric-difference class at corruption level $q \in \{\frac{1}{2}, 1\}$: a mismatched graph keeps the same number of edges as the PDE graph but substitutes pathways, which holds the interaction budget fixed while destroying its identity.

Training sets are nested prefixes of a frozen permutation (AM: 5/18/45 configurations for scarce/medium/abundant; TRL: 2/4/9), so the scarce regime is a strict subset of the abundant one

and regime comparisons share test configurations. Validation and test splits are at the level of independent configurations, not frames. All reported numbers are validation-selected test relative L2 errors in physical units over 3 seeds (5 for the headline scarce cells), and every test set was evaluated exactly once per run, after checkpoint selection. Full protocols are in Appendices B–D.

3.2 Fixed coupling priors scale differently with data

Let $\epsilon_g(N)$ be the test error of the model prescribed graph g at training-set size N. We define the prior-fidelity span

$$S(N) \;=\; 100 \cdot \frac{\overline{\epsilon_{q=1}(N)} - \epsilon_{\mathrm{PDE}}(N)}{\epsilon_{\mathrm{Dense}}(N)}, \tag{1}$$

the mean excess error of the maximally mismatched graphs over the PDE graph, relative to Dense. If scarce data amplified the value of correct structure, $S(N)$ would decrease in N.

It does so on only one of the two controlled systems (Fig. 1). On AM, $S(N)$ increases from $+1.7 \pm 2.4$ pp at 5 configurations—statistically consistent with zero—to $+14.4 \pm 1.2$ pp at 18 and $+15.7 \pm 1.0$ pp at 45 ($\Gamma = -14.1$ pp; all 18 seed$\times$graph combinations negative; the trend is carried by the medium and abundant regimes). On TRL, $S(N)$ decreases from $+8.2$ to $+3.5$ pp ($\Gamma = +4.7$ pp). A fixed prior's value is therefore not a function of data quantity alone; it depends on which system the prior describes. The NoCross control shows the cross-field information itself is valuable in both systems at medium and abundant data (on AM, removing all coupling raises error by 62–103%), while on TRL at scarce data coupling mildly hurts: NoCross (10.49%) beats both the PDE graph (10.64%) and Dense (11.56%).

3.3 Pathway effects change across data regimes

To localize these differences, we remove one pathway at a time from the PDE graph and measure the marginal removal effect $I_e(N)$, the relative error increase over the full PDE model (Fig. 2). Three patterns repeat across systems. First, each system has one dominant pathway whose effect grows strongly with data: removing AM $D \to U$ (active-stress forcing) costs $+10.6\%$ at scarce data and $+64.4\%$ at abundant; removing TRL $U \to R$ (advection of velocity information into the radiation field) costs -0.8% at scarce data and $+31.4\%$ at abundant. Second, inert pathways exist inside the physically valid graph: AM $U \to C$ never exceeds $+2.1\%$ (the concentration field is nearly constant), and TRL $R \to U$, $R \to P$ are negligible at all data levels. Third, the functional harm of a mismatched graph is predicted by which pathways it removes, not by how many: on TRL, harm correlates with the summed removed-pathway effects at Spearman 1.00 (n=5 corruption variants, permutation $p \approx 0.017$), while graph-edit distance q carries no signal (Appendix G).

The pathway interventions refine the system-level picture. The dominant pathways exhibit different finite-data profiles across the two systems: the removal effect of AM $D \to U$ grows from 10.6% to 64.4%, whereas TRL $U \to R$ is already consequential in the scarce regime and grows further with data. Other physically valid pathways have much smaller marginal effects. Thus, a single graph-level decision hides substantial variation across individual interactions and data regimes.

The system-level span $S(N)$ is only a summary of these pathway-level effects. On AM, the PDE graph's advantage over Dense changes modestly across the three main regimes ($+4.1 \to +2.9 \to +2.2$ pp), while the penalty from particular mismatched graphs grows much more sharply. The aggregate reversal therefore reflects which interactions are retained or removed, motivating an adaptive pathway-level representation.

4 Evidence-Informed Pathway Coupling

4.1 Structural-strength factorization

Sec. 3 shows that the missing object is a per-pathway, regime-dependent strength. We therefore enlarge the candidate set to all $K(K-1)$ cross-field pathways and give each pathway $e = j \to i$ a factored form:

$$\mathcal{K}^{\mathrm{eff}}_{j\to i} \;=\; g_{j\to i}\, \mathcal{K}_{j\to i}, \qquad g_{j\to i} = \sigma(\theta_{j\to i}) \in [0,1], \tag{2}$$

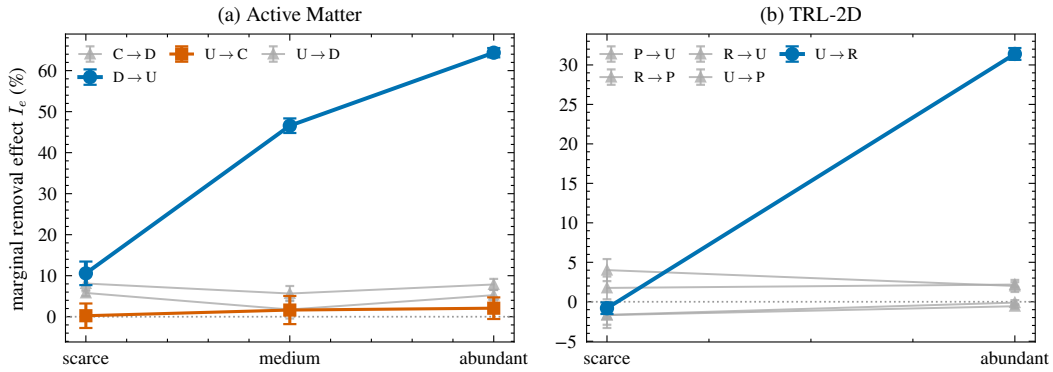


Figure 2: **Individual pathway effects differ sharply and change with data.** Marginal removal effect I_e: the error increase over the full PDE model when pathway e is removed (mean $\pm$ s.d. over seeds). On Active Matter, removing $D \to U$ grows from $+10.6\%$ to $+64.4\%$ with data while $U \to C$ never exceeds $+2.1\%$; on TRL, $U \to R$ moves from -0.8% to $+31.4\%$. Structural decisions therefore live at the level of individual interactions, which motivates the pathway-strength factorization of Sec. 4.

where $\mathcal{K}_{j\to i}$ is the pathway's spectral interaction operator (unchanged from the fixed-structure blocks) and $g_{j\to i}$ is a scalar structural strength. The factorization separates *how much* a pathway is used from *what* the pathway computes: the high-dimensional interaction function is learned in $\mathcal{K}_e$, and the low-dimensional structural decision is made explicit in g_e. The parameterization spans the fixed structures exactly: all gates at one recover Dense, all at zero recover NoCross, and gates at the PDE mask recover the hard prior, with zero numerical discrepancy in each case (unit tests in Appendix C). Structural selection is thereby turned from an architecture-level choice into a continuous, learnable, low-dimensional quantity.

Why expose pathway strength explicitly? A dense operator can in principle attenuate an unhelpful interaction by learning a small kernel $\mathcal{K}_e$. This couples two distinct estimation problems, however: whether a pathway should contribute and what function it should compute are both encoded in the same high-dimensional operator. Equation 2 exposes the first decision as a single scalar g_e, allowing inexpensive pathway-level evidence to act directly on structural strength while leaving $\mathcal{K}_e$ to model the interaction itself. The factorization changes the inductive bias rather than the representational envelope: Dense remains a limiting case when all $g_e = 1$.

4.2 Short-warmup evidence

For each candidate pathway, we train a reduced model containing the self-field blocks and that single cross-field pathway for 2,000 steps, one tenth of full training. We define its short-warmup utility u_e as the validation improvement over the self-field-only model. The six pathway probes require no additional data and can be run independently. Among three inexpensive signals evaluated on the development systems, short-warmup utility gives the most consistent agreement with converged pathway-removal effects (Appendix H), and we use it to initialize the pathway strengths.

4.3 Initialization and joint training

We min–max normalize the six utilities to $m_e \in [0, 1]$ and initialize the gate logits as

$$\theta_e(0) = 2(2m_e - 1),$$

which maps the lowest- and highest-utility pathways to initial strengths of approximately 0.12 and 0.88. The warmup length and mapping are fixed across all experiments. Gates then remain learnable and are optimized jointly with all spectral kernels under the same objective and budget as the fixed-structure models. Because the gates start near the evidence-supported configuration and remain differentiable, training can both exploit the probe where it is right and correct it where it is wrong (Fig. 6 in Appendix I).

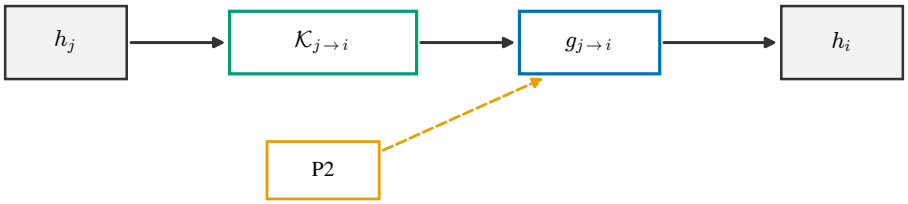


Figure 3: **Evidence-informed pathway coupling.** Each cross-field pathway carries a learned spectral interaction operator $\mathcal{K}_{j\to i}$ scaled by a scalar pathway strength $g_{j\to i}$. A short-warmup probe (P2) measures each candidate pathway's early utility and initializes the strengths; joint training then refines them. The factorization separates the structural decision (how much a pathway is used) from interaction-function learning (what the pathway computes).

4.4 Training procedure

Evidence-informed pathway coupling (M3).

1. Train a self-field-only model briefly (2,000 steps).
2. For each candidate pathway e: activate e alone, train 2,000 steps, record validation improvement u_e.
3. Map utilities to gate logits, $\theta_e(0) = 2(2m_e - 1)$, m_e the min-max-normalized utility.
4. Train the full gated operator (Eq. 2) with gates and kernels optimized jointly until the validation-selected checkpoint.

4.5 Complexity

The method adds six active pathway gates over Dense (the implementation allocates a $K\times K$ gate tensor, so the parameter audit shows nine scalars, three of them unused self-pair slots: 21,246,831 vs 21,246,822) and no measurable compute overhead (1,111 s vs 1,276 s per run on one RTX 3090). The probe costs six short single-pathway runs per regime, each about a tenth of a full run.

5 Experiments

5.1 Setup and baselines

We compare **M3** (Eq. 2 with probe initialization) against the fixed structures Dense, PDE, and NoCross, two gate ablations (**M1**: one global gate; **M2**: per-edge gates, uniform 0.5 initialization), and **M3-Frozen** (probe-initialized gates held fixed). The external baseline is **CoDA-NO** (Rahman et al., 2024) trained from the official implementation under the same splits, budget, and evaluation. All models share the training protocol of Sec. 3.1; per-seed values for every table are in the appendix.

5.2 Evidence-informed pathway strengths improve finite-data adaptation

Table 2 and Fig. 4 compare M3 with three fixed structural choices: Dense, PDE, and NoCross. The strongest fixed choice changes across systems and regimes. On AM scarce, the PDE graph is the strongest fixed baseline at 6.77%, compared with 7.06% for Dense; M3 improves this to 6.44%. On TRL scarce, NoCross is the strongest fixed baseline at 10.49%, while Dense reaches 11.56%; the same M3 model reaches 10.36%. M3 also improves the strongest fixed result on AM medium and TRL medium. In the two abundant regimes, where the strongest fixed alternatives are already close, M3 remains within 0.08 percentage points of the best fixed result.

For secondary analysis, we report regret to the best fixed structural baseline selected in hindsight in Appendix I. The main comparisons below use the raw test errors.

Table 2: Main AM/TRL results: test relative L2 error (%, mean ± s.d.). Fixed structural baselines use dense cross-field coupling (Dense), the PDE-derived coupling graph (PDE), or self-field interactions only (NoCross). M3 uses the same dense candidate set with evidence-informed pathway strengths. Full per-seed results, structural corruptions, and retrospective best-fixed regret are reported in the appendix.

Sys	Regime	Dense	PDE	NoCross	**M3**
AM	scarce	7.06	6.77	10.93	**6.44**±0.04
AM	medium	4.18	4.06	8.23	**3.97**±0.004
AM	abundant	3.28	**3.21**	6.39	3.28±0.01
TRL	scarce	11.56	10.64	**10.49**	10.36±0.18
TRL	medium	8.77	8.48	9.93	**8.47**±0.06
TRL	abundant	7.25	**7.06**	9.25	7.14±0.04

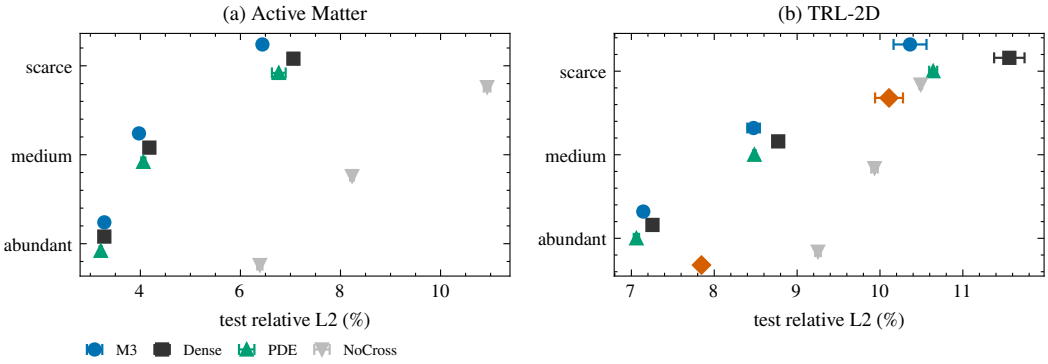


Figure 4: **Performance across data regimes.** Test relative L2 error (mean ± s.d.) for the fixed structural baselines, M3, and CoDA-NO where available. M3 improves the strongest fixed baseline in the scarce and medium regimes of both development systems and remains close to it at abundant data. On TRL, CoDA-NO obtains the lowest scarce-data error, whereas M3 is substantially stronger in the abundant regime.

The ablations isolate the contribution of pathway evidence. M1 uses one global cross-field gate, whereas M2 uses independent pathway gates with uniform initialization. Their performance is similar in the most diagnostic scarce regimes: on TRL both reach 10.76%, and neither reproduces the gain of M3. On AM scarce, M3 reaches 6.44% compared with 6.89% for M2. Thus, adding pathway-specific gate parameters alone is insufficient; the short-warmup pathway evidence provides the main additional signal.

5.3 Robustness to prior mismatch

We next test structural misspecification by replacing the PDE graph with mismatched graphs of the same edge budget. These perturbations can substantially increase error. For example, on AM abundant data the most damaging tested mismatch increases error to 5.24%, compared with 3.21% for the PDE graph and 3.28% for M3. Similar behavior appears across the corruption matrix: M3 is trained over the dense candidate set and avoids most of the degradation caused by committing to a poorly matched fixed graph. Appendix G reports the complete corruption matrix and a normalized recovery analysis.

There is a design lesson in which mismatches hurt. The corruption classes are defined by graph-edit distance, but the damage is not: on TRL, all five maximally mismatched graphs differ from the PDE graph by exactly one edge, yet their abundant-regime errors spread from 7.18% to 7.53% in the order predicted by which pathway was removed, and on AM the two maximally mismatched graphs differ by 1.6 pp in the same predicted order (Appendix G). Structural priors should therefore be evaluated by the marginal value of their individual pathways, not by how many edges they share with a reference graph. This is the practical content of the pathway view: because harm is carried

Table 3: Prospective RBC results (test rel-L2, mean ± s.d., 3 seeds) and method complexity. The method and probe protocol were frozen before RBC evaluation.

Regime	Dense	PDE	NoCross	**M3**
scarce	0.0872	0.0880	0.0919	**0.0831**±0.0009
abundant	0.0753	0.0801	0.0887	**0.0751**±0.0004
parameters		16.5M (PDE)	7.1M (NoCross)	21.25M + 9 (M3)

by individual pathways rather than by the graph as a whole (Sec. 3.3), per-pathway strengths can remove it selectively.

5.4 Comparison with CoDA-NO

CoDA-NO (Rahman et al., 2024) provides an external learned-interaction baseline. On TRL scarce, the 1.8M-parameter CoDA-NO configuration obtains 10.11%, compared with 10.36% for M3, making CoDA-NO the strongest method in that regime. With abundant data, M3 reaches 7.14%, compared with 7.85% for the same CoDA-NO configuration and 7.06% for the fixed PDE graph.

We additionally increase CoDA-NO to 7.24M parameters on the abundant regime. The larger configuration obtains 7.93%, compared with 7.85% for the 1.8M configuration. Increasing capacity over this range therefore does not close the observed abundant-regime gap. Across the tested regimes, the two approaches have different performance profiles: CoDA-NO is strongest on TRL-scarce, whereas M3 is substantially stronger on TRL-abundant. Configuration and resource details are reported in Appendix L.

5.5 The frozen method transfers to Rayleigh–Bénard convection

The method, short-warmup procedure, and training configuration were frozen before RBC evaluation. The fixed PDE graph is not the strongest fixed structure on this system: Dense reaches 0.0872 at scarce data and 0.0753 at abundant data, compared with 0.0880 and 0.0801 for the PDE graph. Frozen M3 improves the scarce-data result to 0.0831 ± 0.0009, a 4.7% reduction relative to Dense, and preserves the abundant-data performance at 0.0751 ± 0.0004 (Table 9, Fig. 5).

This prospective evaluation tests the complete procedure beyond the systems used to develop it: neither the RBC pathway evidence nor its resulting gate initialization was available during method development on AM and TRL. This prospective evaluation tests the complete procedure beyond the systems used to develop it: neither the RBC pathway evidence nor its resulting gate initialization was available during method development on AM and TRL. The scarce regime uses one training configuration and the abundant regime uses four; both are evaluated on the same two held-out configurations. Full per-seed results, the frozen protocol, short-warmup pathway utilities, and learned gate values are reported in Appendix M.

5.6 Method decomposition

M3-Frozen isolates the contribution of evidence-informed initialization. On AM scarce, freezing the P2-initialized gates gives 6.39%, compared with 6.44% for M3 and 6.89% for uniformly initialized M2. On TRL scarce, M3 reaches 10.36%, M3-Frozen 10.44%, and M2 10.76%. Thus, short-warmup initialization provides the common improvement in both systems, while joint refinement contributes an additional gain on TRL.

Because M2 and M3 have the same pathway-specific parameterization and differ only in their initialization, this comparison separates the contribution of pathway evidence from that of adding gate parameters. The scarce-regime gains therefore cannot be explained by the additional scalar degrees of freedom alone.

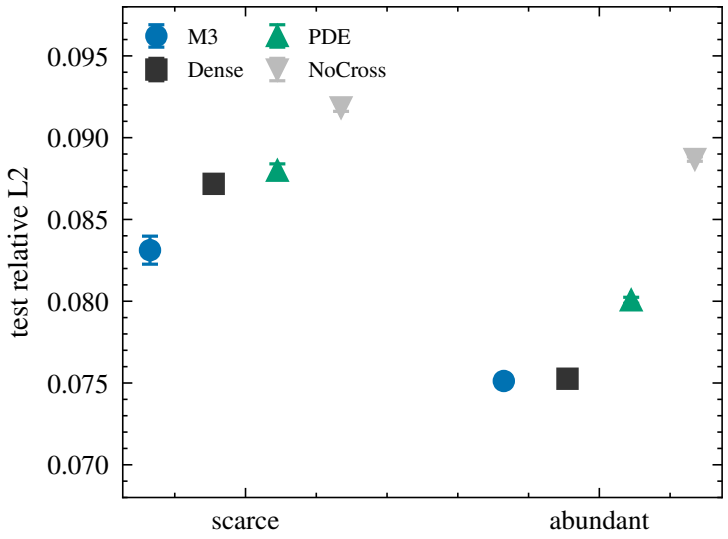


Figure 5: **Prospective transfer to Rayleigh–Bénard convection.** The method, short-warmup procedure, and training configuration were frozen before RBC evaluation. M3 reduces scarce-data test error from 0.0872 for the strongest fixed baseline to 0.0831 ± 0.0009, while matching the strongest fixed result at abundant data (0.0751 ± 0.0004 versus 0.0753). Error bars show standard deviation over three training seeds.

6 Discussion

The experiments shift the structural question from the coupling graph to individual pathways. Across the development systems, physically valid interactions have markedly different marginal effects, and the effect of the same pathway can change with data availability. A single fixed graph therefore compresses several distinct finite-data decisions into one structural choice.

The results also show that adapting these decisions does not require a large auxiliary architecture. Exposing pathway strength as a low-dimensional variable and initializing it from short-warmup evidence improves several scarce and medium regimes in which fixed structural choices differ substantially in performance, while preserving performance when a strong fixed choice already exists. The prospective RBC experiment shows that the frozen procedure transfers without system-specific method redesign in the tested third system.

Our experiments focus on coupled systems with a small number of field groups, where pathway-level probing remains inexpensive. Scaling the same procedure to systems with many interacting fields will require cheaper pathway screening or structured candidate sets. Within the systems studied here, the results support combining PDE knowledge with target-regime evidence at the level of individual interactions.

7 Conclusion

We studied how strongly neural operators should use PDE-derived coupling pathways under finite data. Across coupled physical systems, fixed coupling structures exhibit different data-scaling behavior, and pathway-removal interventions show that individual interactions can have markedly different finite-data effects. This motivates an evidence-informed factorization that separates pathway strength from the interaction operator. The resulting model improves several regimes in which fixed structural choices differ substantially in performance, remains competitive when a strong fixed choice already exists, and transfers without retuning to Rayleigh–Bénard convection. These results support using PDE knowledge as a source of candidate interactions whose strength is calibrated from evidence in the target data regime.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

Anonymous code accompanies the submission: the typed spectral operator, graph definitions for each system, frozen regime and coupling configurations, pathway-intervention code, and the analysis scripts and source tables used for the reported results. Seeds are fixed for Python, NumPy, PyTorch, and CUDA, and per-seed results are reported for the principal comparisons.

The Active Matter and turbulent radiative layer datasets are from The Well (Ohana et al., 2024). The Rayleigh–Bénard source, preprocessing, and the pre-result protocol amendments are documented in the frozen protocols included with the supplementary material. The released artifacts reproduce the main data-regime, pathway-intervention, adaptive-coupling, and prospective transfer analyses.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for designing and providing feedback on research methodology and experiments, providing feedback on mathematical arguments, and interpreting results. We have not used generative AI tools for generating synthetic data sets, formulating mathematical claims or implementing methods, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for literature discovery, brainstorming, drafting and editing parts of the manuscript, creating and editing software code and figures, and simulated peer-review feedback. We have reviewed all AI-assisted work and AI-assisted code was re-run with its outputs checked against the experimental artifacts reported in this paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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.

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*, 2021.

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Anima Anandkumar. Geometry-informed neural operator for large-scale 3d pdes. In *Advances in Neural Information Processing Systems*, 2023.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024. doi: 10.1145/3648506.

Miguel Liu-Schiaffini, Julius Berner, Boris Bonev, Thorsten Kurth, Kamyar Azizzadenesheli, and Anima Anandkumar. Neural operators with localized integral and differential kernels. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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*, 3:218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Siying Ma, Mehrdad Momeni Zadeh, Mauricio Soroco, Wuyang Chen, Jiguo Cao, and Vijay Ganesh. Learning data-efficient and generalizable neural operators via fundamental physics knowledge. In *International Conference on Learning Representations*, 2026.

Ruben Ohana, Michael McCabe, Lucas Meyer, et al. The well: A large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, 2024.

Jaewan Park, Kazuma Kobayashi, Qibang Liu, Diab Abueidda, Seid Koric, and Syed Bahauddin Alam. Network architecture follows coupling in multiphysics systems: single vs. multiple branches in deeponet and s-deeponet. *Communications Engineering*, 2026. doi: 10.1038/s44172-026-00714-4.

Md Ashiqur Rahman, Robert Joseph George, Mogab Elleithy, Daniel Leibovici, Zongyi Li, Boris Bonev, Colin White, Julius Berner, Raymond A. Yeh, Jean Kossaifi, Kamyar Azizzadenesheli, and Anima Anandkumar. Pretraining codomain attention neural operators for solving multiphysics pdes. In *Advances in Neural Information Processing Systems*, 2024.

Yingjie Tan, Quanming Yao, and Yaqing Wang. Dgnet: Discrete green networks for data-efficient learning of spatiotemporal pdes. In *International Conference on Learning Representations*, 2026.

Zhixue Wang, Xuefeng Ge, Ji Zhang, Hao Niu, and Lei Wu. Causalphysicsnet: Discovering effective coupling structure in multi-physics simulations via physics-informed temporal graph networks. *Computer Methods in Applied Mechanics and Engineering*, 2026. doi: 10.1016/j.cma.2026.119098.

Xiongye Xiao, Defu Cao, Ruochen Yang, Gaurav Gupta, Gengshuo Liu, Chenzhong Yin, Radu Balan, and Paul Bogdan. Coupled multiwavelet neural operator learning for coupled partial differential equations. In *International Conference on Learning Representations*, 2023.

A Governing systems and pathway construction

A.1 Active Matter (AM)

The Active Matter system from The Well (Ohana et al., 2024) describes a suspension of self-propelled particles in a viscous fluid. The state groups are the velocity field $U = (u, v)$, the concentration field C, and the derived active-stress forcing field D. The explicit-mechanism graph records four of the six possible cross-field pathways: $U \to C$ (advection of concentration), $U \to D$, $C \to D$ (concentration-dependent stress), and $D \to U$ (active-stress forcing of the flow). The concentration field is nearly constant in this dataset (channel std 0.003, resolution-independent), which makes the two C-dependent pathways weak a priori.

A.2 Turbulent radiative layer (TRL)

The two-dimensional turbulent radiative layer couples a radiation intensity field R, pressure P, and velocity U. The explicit-mechanism graph contains five of six pathways; all fields are strongly dynamic (density/velocity std $\gg 0$, persistence well below the trained model), so no pathway is trivially inert by construction.

A.3 Rayleigh–Bénard convection (RBC)

Two-dimensional Rayleigh–Bénard convection couples buoyancy B, pressure P, and velocity U. The explicit-mechanism graph is $\{U \to B, P \to U, B \to U\}$: advection of buoyancy, pressure-gradient forcing, and buoyant forcing of the flow. RBC was reserved as a prospective system; the method and probe protocol were frozen before any RBC training run.

A.4 Candidate sets

Every adaptive method uses the full $K(K-1) = 6$ candidate cross-field pathways per system, regardless of how many appear in the explicit-mechanism graph. The graph is used only as a fixed-structure baseline and as the reference for corruption classes.

B Dataset construction and splits

Sources. AM and TRL are from The Well (Ohana et al., 2024) (82 shards, 52.5 GB for AM, integrity-checked). RBC is the two-dimensional Rayleigh–Bénard convection dataset from The Well

(Ohana et al., 2024). The original trajectories contain 200 frames at 512×128 resolution with buoyancy, pressure, and two velocity components. We use eight configurations and downsample the spatial fields to 128×32, giving channels $[b, p, u, v]$.

Preprocessing. AM: compacted to $64\times64 \times 81$ frames $\times$ 6 channels; block-mean downsampling verified by unit tests. TRL: native 128×128. RBC: block-mean from 512×128 to 128×32. All fields are normalized per channel by training-split statistics; metrics are reported in physical units after inverse normalization.

Splits. Configuration-level splits throughout: no trajectory appears in more than one of train/validation/test. AM: scarce $= 5$, medium $= 18$, abundant $= 45$ training configurations; TRL: scarce $= 2$, medium $= 4$, abundant $= 9$ (both nested prefixes of a fixed permutation, seed 20260918), with 2 validation and 2 test configurations held out per system. RBC: 4 train / 2 validation / 2 test configurations; scarce uses 1 training configuration, abundant uses all 4.

Temporal sampling. All main experiments use one-step-ahead prediction at temporal stride 1 (AM $\Delta t = 0.25$). Multi-step rollout behavior is evaluated separately in Appendix K.

C Architecture details

The backbone is a typed spectral operator. Each field group is lifted to width $d_v = 32$, processed by four spectral layers retaining 12×12 Fourier modes on the two spatial corners (low-mode mixing), and projected back to the group's channels. Within each layer, every self-field group has its own spectral block, and every prescribed cross-field pathway has its own spectral block of identical shape. Complex spectral weights are counted as two scalars in all parameter audits.

Fixed structures: Dense = all six cross-field pathways (21,246,822 parameters); PDE = the explicit-mechanism graph (16,528,230); NoCross = self-field blocks only (7,091,046). A width-matched Dense variant matches the four-edge parameter budget within 2% and differs from Dense by at most 0.08 pp test error in every regime, so capacity does not explain structural effects.

Adaptive variants: **M1** adds one global gate (Dense $+1$); **M2** adds per-edge gates initialized at $\theta = 0$ (Dense $+9$, of which six multiply used pathways); **M3** adds the same gates with probe-informed initialization. Unit tests verify the exact structural limits: gates $= 0$ reproduce NoCross, gates $= 1$ reproduce Dense, gates $=$ PDE mask reproduce the hard prior (maximum output difference 0.0), gradients reach all gates and kernels, and parameter counts are exact (13/13 checks).

D Training protocol

All runs: AdamW, learning rate 10^{-3}, weight decay 10^{-4}, cosine schedule, gradient-norm clipping 1.0, batch size 16, 20,000 steps (P2 probes: 2,000). Random pairs $(\text{trajectory}, t)$ are sampled each step. Validation relative L2 every 500 steps selects the checkpoint; the test set is evaluated exactly once per run, after selection. Seeds 0–2 per cell (headline scarce cells also 3–4). Hardware: $8\times$ RTX 3090 (AM, TRL, RBC) and $1\times$ A100-40GB (CoDA-NO). Total program compute: $\sim$350 training runs, $\sim$60 GPU-hours.

E Full fixed-structure results

Per-seed test relative L2 for every fixed-structure cell is in the released CSVs (`AM_gateA_results.csv`: 101 runs; `AM_TRL_mechanism_results.csv`: 208 runs). Table 4 repeats the regime means for reference.

On TRL the mismatched variants are q10_0..q10_4. Each replaces exactly one PDE edge with the single non-PDE pathway ($P \to R$), so the edge budget (5) is preserved, as on AM. Their abundant-regime test errors are 7.53, 7.30, 7.23, 7.31, 7.18% respectively (deviations from Dense of -3.85, -0.68, $+0.32$, -0.79, $+0.96$ pp), and q10_0—the variant replacing the dominant $U \to R$ pathway—is the most harmful at every regime. All five variants enter the TRL $S(N)$ average; Table 4 shows two for space.

Table 4: Fixed-structure test relative L2 (%, mean over 3 seeds).

Sys	Regime	Dense	PDE	NoCross	q05_0	q10_0	q10_1
AM	scarce	7.06	6.77	10.93	7.52	7.09	6.68
AM	medium	4.18	4.06	8.23	5.95	4.69	4.63
AM	abundant	3.28	3.21	6.39	5.24	3.75	3.70
TRL	scarce	11.56	10.64	10.49	—	11.52	11.78
TRL	medium	8.77	8.48	9.93	—	9.05	8.88
TRL	abundant	7.25	7.06	9.25	—	7.53	7.30

F PATHWAY-REMOVAL INTERVENTIONS

The marginal removal effect $I_e(N) = 100 \cdot (\epsilon_{\setminus e} - \epsilon_{\mathrm{PDE}})/\epsilon_{\mathrm{PDE}}$ is the relative error increase when pathway e is removed from the PDE graph. Full per-seed values are in `pathway_results.csv`; means:

Table 5: Marginal removal effects I_e (%). AM has all three regimes (scarce cells 2 seeds, medium/abundant 3); TRL has scarce and abundant (3 seeds).

System	Pathway	scarce	medium	abundant
AM	$D \to U$	+10.6	+46.6	+64.4
AM	$U \to D$	+8.1	+5.7	+7.9
AM	$C \to D$	+5.8	+1.8	+5.3
AM	$U \to C$	+0.2	+1.6	+2.1
TRL	$U \to R$	-0.8	—	+31.4
TRL	$U \to P$	+4.0	—	+2.0
TRL	$P \to U$	+1.8	—	+2.2
TRL	$R \to U$	-1.7	—	-0.1
TRL	$R \to P$	-1.7	—	-0.6

Negative values indicate slight improvement on removal (within seed noise for the inert pathways). The dominant-pathway pattern—one pathway growing strongly with N, others flat—holds in both systems and recurs in RBC's probe utilities.

G PRIOR-CORRUPTION STUDY

Mismatched priors are drawn from the feasible symmetric-difference class: for a 3-group graph with four PDE edges, $q \in \{0, \frac{1}{2}, 1\}$ normalizes the edge-set distance by its feasible maximum, holding the edge count fixed. On TRL (five PDE edges), each $q{=}1$ variant replaces exactly one PDE edge with the non-PDE pathway $P \to R$ (edge budget preserved), which makes the system a clean test of whether graph-edit distance predicts functional harm. It does not: harm is $+3.85$ pp for the variant replacing $U \to R$ and at most $+0.96$ pp for the others, tracking the replaced pathway's I_e (Spearman 1.00, $n{=}5$, permutation $p \approx 0.017$), while q is constant across variants and carries no signal. On AM the two $q{=}1$ graphs differ in harm (14.4 vs 12.8 pp) in the order predicted by their summed removed-pathway effects (72 vs 66). Graph-edit distance should therefore not be used as a proxy for prior quality.

H PROXY STUDY

Three frozen proxies were evaluated against converged removal effects (Spearman correlation of pathway rankings):

Among the three tested signals, P2 shows the most consistent positive rank agreement with converged removal effects across the four evaluated settings. It is imperfect: on AM, for example, the short warmup ranks $U \to D$ above the pathway with the largest converged removal effect, $D \to U$.

Table 6: Proxy validity: Spearman correlation with converged I_e rankings.

Proxy	AM scarce	AM abundant	TRL scarce	TRL abundant
P1 gradient energy	+0.20	+0.20	−0.10	−0.70
P2 short-warmup (2k)	**+0.80**	**+0.80**	+0.60	**+0.80**
P3 linear-increment	+0.40	+0.40	−0.10	−0.40

P1 is less consistent across regimes, while P3 performs poorly on TRL. These comparisons motivate using P2 as an initialization signal in the method tournament.

I METHOD TOURNAMENT

M1/M2/M3 share the dense candidate set, training protocol, and evaluation; they differ only in gate structure and initialization (Sec. 4, Appendix C). Per-regime means (test rel-L2 %):

Table 7: Method tournament (mean ± s.d.; scarce cells 5 seeds, others 3).

Sys	Regime	M1 (global)	M2 (per-edge)	M3 (probe-init)
AM	scarce	6.82 ± 0.06	6.89 ± 0.08	$\mathbf{6.44 \pm 0.04}$
AM	medium	4.13 ± 0.04	4.10 ± 0.04	$\mathbf{3.97 \pm 0.00}$
AM	abundant	3.34 ± 0.01	3.31 ± 0.01	$\mathbf{3.28 \pm 0.01}$
TRL	scarce	10.76 ± 0.15	10.76 ± 0.15	$\mathbf{10.36 \pm 0.18}$
TRL	medium	8.56 ± 0.05	8.56 ± 0.05	$\mathbf{8.47 \pm 0.06}$
TRL	abundant	7.13 ± 0.02	7.15 ± 0.01	7.14 ± 0.04

M1 and M2 coincide at TRL scarce (10.76): without pathway information the gates collapse to a uniform ≈ 0.51, so per-edge flexibility alone adds nothing. Gate trajectories during training are mechanism-consistent: dominant pathways stay high (AM $U \to D$ at 0.86; TRL $U \to R$ at 0.92), inert pathways stay suppressed (0.1–0.2), and the probe-misranked AM $D \to U$ climbs $0.12 \to 0.27$ as joint training corrects the initialization.

J M3-FROZEN DECOMPOSITION

Table 8: Decomposition (test rel-L2 %, mean ± s.d., 3 seeds).

Sys	Regime	M2 (uninformed)	M3-Frozen (probe, frozen)	M3 (probe, refined)
AM	scarce	6.89	6.39 ± 0.03	6.44 ± 0.04
AM	abundant	3.31	3.28 ± 0.02	3.28 ± 0.01
TRL	scarce	10.76	10.44 ± 0.17	10.36 ± 0.18
TRL	abundant	7.15	7.15 ± 0.05	7.14 ± 0.04

On AM the probe initialization alone carries the low-data gain (Frozen $\approx$ M3 $\ll$ M2); on TRL refinement adds a further increment (M3 $<$ Frozen $<$ M2). Both orderings support the design: initialize from evidence, then let training correct it.

K MULTI-STEP AM ROLLOUT

Forty-five checkpoints from the AM runs were rolled out autoregressively to horizons $h \in \{1, 5, 10, 20\}$. The prior-fidelity span $S(N, h)$ remains increasing in N through $h = 10$ (scarce $+2.4 \to +8.6$ pp, abundant $+10.3 \to +20.0$ pp), so the central scaling reversal is not an artifact of one-step evaluation. At $h = 20$ all models sit at the chaos horizon (errors 67–82%) and the metric saturates; we therefore report the $h \leq 10$ range as the reliable window.

L CoDA-NO comparison

CoDA-NO (Rahman et al., 2024) was trained from the official `neuralop` implementation (CODANO class), same splits, 20k steps, cosine schedule, 3 seeds, validation-selected checkpoints. The 1.8M-parameter configuration (hidden 128, 4 layers, 12×12 modes) yields test rel-L2 10.11% at TRL scarce (best of all compared methods) and 7.85% at TRL abundant (worst of the compared set; M3 7.14%, PDE 7.06%). A capacity check at 7.24M parameters on TRL abundant yields 7.93% test rel-L2 (20k steps, seed 0, one A100-40GB; vs 7.85% for the 1.8M configuration over 3 seeds) (configuration, runtime, and per-seed values in `CODA_CAPACITY_CHECK.md`). On AM (six input channels) the official configuration exceeded the memory of both the 24 GB and 40 GB GPU classes at every viable width; no faithful run completed, so we report no AM CoDA number rather than substitute a weakened variant.

M Prospective RBC evaluation

The RBC protocol—regimes (scarce $= 1$ configuration, abundant $= 4$), graphs, probe budget, and evaluation—was frozen before any RBC training run, and the method hyperparameters were inherited unchanged from AM/TRL. All 24 matrix runs (dense/pde/nocross/M3 $\times$ 2 regimes $\times$ 3 seeds) plus 14 probe runs completed; per-seed values are in `RBC_system3_results.csv`.

Table 9: RBC prospective results (test rel-L2, mean $\pm$ s.d., 3 seeds).

Regime	Dense	PDE	NoCross	M3
scarce	0.0872	0.0880	0.0919	$\mathbf{0.0831 \pm 0.0009}$
abundant	0.0753	0.0801	0.0887	$\mathbf{0.0751 \pm 0.0004}$

Learned gates (mean over seeds): $U \to B$ $0.87/0.83$, $U \to P$ $0.89/0.50$, $P \to U$ $0.42/0.29$, $B \to P$ $0.24/0.21$, $P \to B$ $0.25/0.15$, and $B \to U$ $0.18/0.11$ (scarce/abundant). The corresponding short-warmup utilities are reported alongside these values in Table 10. In particular, $U \to B$ has high short-warmup utility in both regimes ($0.295 \to 0.571$) and remains strongly weighted after full training ($0.87/0.83$).

Table 10: RBC pathway evidence: short-warmup utility u_e (P2) and final learned gate g_e (mean over 3 seeds), per regime.

Pathway	u_e scarce	u_e abundant	g_e scarce	g_e abundant
$U \to B$	0.295	0.571	0.87	0.83
$U \to P$	0.281	0.330	0.89	0.50
$P \to U$	0.133	0.242	0.42	0.29
$B \to P$	0.107	0.203	0.24	0.21
$P \to B$	0.094	0.165	0.25	0.15
$B \to U$	0.069	0.138	0.18	0.11

N Runtime, memory, and parameter accounting

Per 20k-step run on one RTX 3090: Dense 1,276 s, PDE 929 s, M3 1,111 s. Parameter counts: Dense 21,246,822; PDE 16,528,230; NoCross 7,091,046; M3 = Dense $+9$. CoDA-NO: 1,815,513 (main) and $\sim$7M (capacity check). P2 probe cost: six 2k-step runs per regime, $\sim$10% of one full run each. All training used mixed precision off; evaluation batches of 64 on CPU-held tensors.

O Additional figures

Additional diagnostic figures (functional-corruption analysis, rollout $S(N, h)$ curves, complete pathway matrices) are included in the released artifact bundle with their generating scripts.

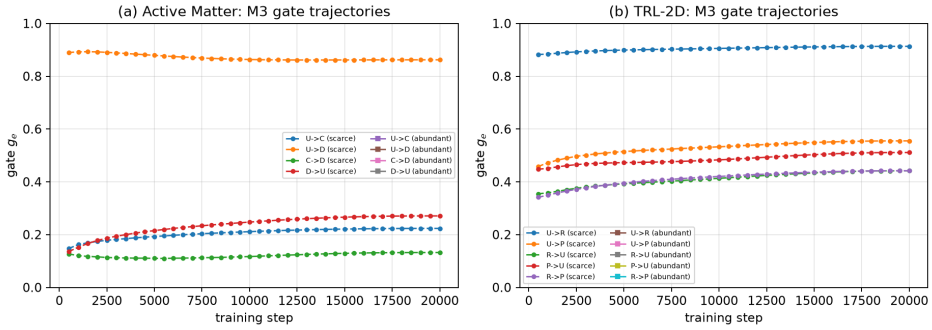


Figure 6: M3 gate trajectories over training on AM and TRL. Dominant pathways remain high, inert pathways remain suppressed, and the probe-misranked AM $D \to U$ gate climbs during joint training.