Supplementary Information for: Mechanism-oriented tendency-integration operator networks for PDE foundation models

Jeongsu Lee[1*]

[1*]School of Mechanical Engineering, Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Yongin-si, 17104, Gyeonggi-do, Republic of Korea.

Corresponding author(s). E-mail(s): leejeongsu@khu.ac.kr;

Supplementary Note 1: detailed evaluation of MOTION

MOTION-S, the scale-reduced model

MOTION-S is a 20.7M-parameter configuration trained and evaluated using the same mechanism vocabulary, two-stage objective, corpus, and protocol as the 156.9M main model. The models differ only in latent width and depth, with 384 channels over eight blocks for MOTION-S and 896 channels over twelve blocks for MOTION. Training requires approximately 155 H100-hours for MOTION-S, compared with 570 H100-hours for MOTION, and yields a class-averaged validation error of 8.22% (Supplementary Table 1). This comparison distinguishes architectural properties retained under a 7.6-fold parameter reduction from performance dependent on model capacity.

Per-family accuracy

Supplementary Table 1 reports per-family accuracy for both model sizes, evaluated in physical space on up to 128 held-out validation trajectories per family, or the full split when fewer are available. Evaluation follows the training protocol with ten observed and ten predicted frames under free-running autoregression or the corresponding family-specific temporal mode. The class-averaged errors are 7.03% and 8.22%. All entries, including the three-dimensional families, are unaveraged results, consistent with Fig. 1.

Three-dimensional inference is evaluated using either a single encoder point sample (K=1), as reported in Supplementary Table 1, or sample averaging over K independent point samples before solution synthesis. Supplementary Table 2 compares both settings under the coefficient-lattice readout. Absolute errors depend on the readout resolution, whereas the effect of sample averaging is assessed through the within-protocol comparison.

Supplementary Table 1 Per-family relative L_2 error (%) in denormalized physical space on held-out validation trajectories. n denotes the number of trajectories evaluated for each equation family.

family	n	MOTION (156.9M)	MOTION-S (20.7M)
Incompressible NS	128	**2.64**	2.83
NS-Sines	128	**8.96**	11.46
NS-Gauss	128	**7.29**	9.76
Allen–Cahn	128	**0.75**	0.92
Diffusion–reaction	128	**4.09**	4.39
Shallow water	128	**0.26**	0.34
Wave (layered)	128	**9.05**	11.12
Poisson	128	**1.55**	2.09
CE-RP	128	**7.62**	10.46
CE-RM	128	**6.46**	7.63
Compressible NS	128	2.32	**2.13**
CE-CRP	128	**14.62**	16.55
CE-Gauss	128	**3.78**	5.02
CE-KH	128	**3.73**	4.03
PDEArena (cond.)	128	**12.00**	12.86
PDEArena (uncond.)	128	**8.61**	12.42
3D CNS M0.1	10	**5.63**	5.65
3D CNS M1.0	10	**11.73**	13.49
3D CNS Turb	63	**22.45**	23.08
class average		**7.03**	8.22

Supplementary Table 2 Sample-averaged inference for the three-dimensional families using the coefficient-lattice readout. Errors are relative L_2 values (%) on the full validation split. K denotes the number of independent encoder point samples whose latent rates are averaged before solution synthesis.

	MOTION (156.9M)			MOTION-S (20.7M)		
family	K=1	K=8	gain	K=1	K=8	gain
3D CNS M0.1	16.16	**8.48**	−47.5%	15.29	**9.54**	−37.6%
3D CNS M1.0	16.82	**14.44**	−14.1%	20.12	**17.72**	−12.0%
3D CNS Turb	36.54	**28.23**	−22.7%	37.39	**29.21**	−21.9%

Full mechanism-suppression matrix

A mechanism is suppressed by setting its gate to zero at evaluation while leaving all other parameters unchanged. Knockout damage is normalized by the available prediction headroom to account for differences in intrinsic difficulty across equation families. Specifically, the error increase is divided by the difference between the persistence error and the full-model error and clipped to $[0, 1]$. A value of one therefore indicates loss of the full improvement over persistence.

Family-specific contribution is evaluated separately from the raw tendency magnitude. For head k and family f, the tendency magnitude is $c_{k,f} = \langle |\sigma_k \odot H^{(k)}(z)| \rangle$, the mean absolute gated contribution averaged over lattice sites, tendency channels, and temporal panels on held-out validation trajectories of family f, with the metadata mask applied. The share of head k is $s_{k,f} = c_{k,f} / \sum_{k'=1}^{19} c_{k',f}$, taken over the gated heads, and the reported contribution is the fold enrichment

$$C_{k,f} = \left[\log_2 \frac{s_{k,f}}{\max(\mathrm{median}_{f'}\, s_{k,f'},\, 0.01)} \right]_+ ,$$

where the denominator is the head's median share across families, floored at a 1% share to avoid inflation from near-inactive heads, $[\cdot]_+$ clips at zero, and the values are normalized within each family by their maximum. Knockout damage and family-specific contribution provide complementary causal and observational measures of mechanism use, respectively.

The two measures capture distinct aspects of mechanism use and are not directly comparable in scale. Family-specific contribution measures enrichment relative to a head's typical use across families, whereas knockout damage measures the error increase caused by removing that head from the tendency sum. A head may therefore show high family-specific contribution despite a modest knockout effect, as observed for the gravity-gradient head in the buoyant families. Interpretation is further limited by overlap among head features; for example, the directional-gradient bank contains both gravity-axis and first-axis derivatives. In addition, heads used uniformly across families remain near their median contribution and are consequently suppressed in the enrichment map. Blank cells therefore indicate either negligible use or use that is not exceptional relative to other families.

The state head $H^{(0)}$ bypasses mechanism gating and serves as a general pathway for state propagation. Its role is therefore quantified directly from its contribution to the summed tendency, rather than by mechanism knockout. Supplementary Table 3 reports the state head's share of the total tendency magnitude and the corresponding ratio of gated-mechanism to state-head contributions. The gated mechanisms collectively account for the majority of the tendency in seventeen of nineteen families, with a median state-head share of 0.28. The two exceptions are the steady elliptic problem (0.60) and shallow water (0.51), both dominated by comparatively simple state evolution. The state-head share decreases for more complex dynamics, reaching 0.15 for compressible Navier–Stokes and 0.17–0.25 for the three-dimensional compressible families. Among the gated heads, the largest contributions are physically consistent: nonlinear reaction dominates Allen–Cahn, the wave head is co-dominant for the

layered-wave family, upwind transport becomes prominent in shock-carrying CE-RM, and directional-gradient and mean-curvature heads dominate the three-dimensional families, consistent with the suppression analysis.

Supplementary Table 3 State-head share of the summed tendency magnitude. $|H^{(0)}|$ is the mean absolute gated contribution of the state head and $\sum_{k\geq 1}|H^{(k)}|$ the sum over the nineteen gated mechanism heads, measured on four held-out validation trajectories per family (two for the three-dimensional families) through one forward segment of the evaluation path.

family	$\lvert H^{(0)}\rvert/(\lvert H^{(0)}\rvert+\sum_k\lvert H^{(k)}\rvert)$	$\sum_k\lvert H^{(k)}\rvert\,/\,\lvert H^{(0)}\rvert$
Compressible NS	0.15	5.85
3D CNS Turb	0.17	4.78
3D CNS M0.1	0.18	4.49
CE-CRP	0.24	3.20
3D CNS M1.0	0.25	2.97
NS-Gauss	0.26	2.88
CE-KH	0.26	2.82
NS-Sines	0.26	2.80
CE-Gauss	0.28	2.63
Wave (layered)	0.28	2.58
CE-RP	0.29	2.51
Diffusion–reaction	0.29	2.49
Allen–Cahn	0.30	2.39
CE-RM	0.32	2.10
PDEArena (cond.)	0.35	1.90
Incompressible NS	0.40	1.52
PDEArena (uncond.)	0.42	1.36
Shallow water	0.51	0.96
Poisson	0.60	0.67

Supplementary Note 2: orientation of the steady elliptic response

The Poisson-Gauss solutions in Fig. 3c exhibit anisotropic spatial structure. Although each Gaussian source is isotropic, the superposed Poisson response inherits a preferred orientation from the spatial arrangement of the sources. This orientation is quantified using the principal axis of the inertia tensor of the median-subtracted positive field. Across the sixteen held-out samples, the median elongation is 1.27, with a range of 1.06–2.71, and the principal axes are generally oblique to the computational grid. The example shown in Fig. 3c has an orientation of 47.3° and an elongation of 1.56.

The same measurement is applied to each prediction to quantify the recovered orientation of the elliptic response. The angular deviation from the ground truth provides a measure of global structural accuracy beyond field amplitude alone. Supplementary Table 4 reports these deviations for the sixteen held-out samples, with MOTION yielding the smaller error in eleven cases. For the example in Fig. 3c, the predicted orientations are 48.2° for MOTION and 52.0° for Poseidon-B, compared with 47.3°

for the ground truth. Supplementary Figure 2 illustrates the corresponding orientation measurement.

Supplementary Table 4 Angular deviation of the predicted principal axis from the ground-truth orientation across the sixteen held-out samples (degrees).

	median	mean	max
MOTION	**1.62**	**1.87**	**4.56**
Poseidon-B	2.66	5.64	29.9

The principal difference appears in the tail of the error distribution rather than in the median, with both operators recovering similar orientations for most samples but differing in their largest deviations. The error maps in Fig. 3c show a corresponding spatial contrast. For the displayed sample, row- and column-coherent structure accounts for 0.72 of the residual energy for Poseidon-B and 0.45 for MOTION. This difference is sample-specific; the corresponding medians over the held-out set are 0.52 and 0.48, respectively.

Supplementary Note 3: development of the physics-bias mechanism vocabulary

The mechanism vocabulary in Fig. 1 reflects both the choice of physical primitives and their numerical realization. Supplementary Table 5 summarizes the development experiments used to establish these choices. All twenty-seven variants were evaluated in the earlier six-family pretraining setting under matched parameter budgets and a common evaluation protocol. These experiments were not repeated on the nineteen-family corpus and should therefore be interpreted as design studies rather than controlled ablations of the final model. The results indicate three distinct trends. Approaches that prescribe the functional content of the physics, including explicit operator banks, fixed analytic propagators, and residual physics losses, provide no consistent improvement over their baselines. Architectural rearrangements that preserve the effective function class, including structured latent slots, disjoint experts, readout separation, integrator variants, additive transport formulations, and learnable stencil weights, likewise produce limited change. Performance improves primarily when the available function class or its numerical realization is altered, as with upwind rather than central transport, semi-Lagrangian rather than additive transport, physics-typed rather than generic readout, and autoregressive rather than single-shot prediction. These studies motivated the design principle used for Fig. 1: a physics bias is characterized by the function class and numerical realization imposed on computation, rather than by the semantic label assigned to a mechanism head.

Supplementary Table 5 Physics-bias mechanisms evaluated during model development. "Pretrain" denotes mechanisms incorporated during multi-family pretraining, while "Transfer" denotes mechanisms introduced during finetuning of a pretrained operator on the initial-value tasks of the earlier study. Outcomes are classified as improvement (+), no significant change (**0**), or degradation (−) relative to the corresponding matched baseline.

mechanism	setting	outcome	one-line finding
Explicit operator bank (Δ, ∇, advection, fixed coefficients)	pretrain	−	Learned mechanism heads consistently outperform fixed operator banks across families
Learned mechanism heads (this paper)	pretrain	+	Forms the central mechanism architecture used in MOTION
Heads before vs. after the transformer	pretrain	order	Encoder-first ordering performs materially better
Learnable stencil weights	pretrain	0	Freeing the finite-difference stencil weights provides no measurable benefit
Physics residual loss (global-norm MSE)	pretrain	−	Degrades all families; objective-level physics bias is unstable in joint training
Central-difference self-advection	pretrain	−	Numerically unstable, with overflow during forward evaluation
Upwind self-advection	pretrain	+	Stabilizes self-advection and improves velocity and tracer prediction
$\partial_t u$ state feed	pretrain	0/+	Benefits the non-recursive variant but remains inferior to recursive state feedback
Slot-structured latent / disjoint experts	pretrain	0	Explicit channel-identity constraints provide no measurable benefit
Per-group readout un-sharing	pretrain	0	Separate readout weights within the same function class provide no measurable benefit
Integrator-scheme variants	pretrain	0	Provide no improvement over analytic tendency integration
History augmentation (K input frames)	pretrain	0/+	Provides a modest benefit, smaller than that from state feedback
Gradient surgery (PCGrad)	pretrain	0	No substantial gradient conflict is observed; model capacity is the dominant limitation
Overcomplete (expanded) encoder	pretrain	$+^*$	Improves accuracy but adds 41M parameters, confounding the comparison
Loose transport bank	pretrain	+	Time-modulated semi-Lagrangian transport improves the transport representation
Seven residual transport arms	pretrain	0	Additive transport corrections provide no benefit beyond the existing model capacity
ω-based velocity decoding (vorticity potential)	pretrain	−	Enforces solenoidality only approximately under forcing and degrades performance
Physics-typed transport decoder	pretrain	+	Changing the readout function class provides a non-redundant improvement
Segment autoregression	pretrain	+	Improves accuracy progressively with rollout depth in turbulent dynamics
Rigid wave propagator (Taylor cos /sinc)	transfer	0	Gates remain inactive, indicating mismatch between the prescribed form and downstream dynamics
Rigid Allen–Cahn kinetics	transfer	−	Consistently underperforms the corresponding unconstrained variant
Loose separable bank $\sum_m w_m(t)B_m(x)$	transfer	+/0/−	Benefits the reaction task but cannot represent transport effectively
Loose spectral-phase bank	transfer	+	The only transfer-time mechanism bank to improve consistently over its matched baseline
Basis-endpoint padding	transfer	+	Reduces endpoint-dominated error and improves all four few-shot tasks
Lead-weighted loss	transfer	0	Provides no benefit, indicating that endpoint error arises from basis geometry rather than gradient weighting
Autoregressive inference at transfer	transfer	+/−	Improves in-family prediction but degrades under distribution shift
Head pruning at transfer	transfer	+/0/−	Becomes increasingly beneficial with distribution shift, but is detrimental in-family

Prescribe structure, learn content.

The development results further support a separation between prescribed structure and learned content. Successful mechanisms specify the form of interaction—which fields couple, through which primitive, and in which representation—while leaving its quantitative realization to a learned map. Residual physics losses instead impose physical content through the objective and degraded joint training, whereas learning stencil parameters provided little benefit when the prescribed numerical operator was already appropriate. The effectiveness of a mechanism therefore depends on whether its function class spans the observed failure mode. A separable representation cannot capture non-separable transport such as $u(x - ct)$, and accordingly performs poorly on wave and steady problems while remaining effective for pointwise reaction. Spectral-phase and transport representations introduce explicit carriers of phase and displacement and improve the corresponding errors. These results favor matching the computational hypothesis space to the physical failure mode rather than increasing the amount of prescribed physics.

Numerical realization materially affects the effectiveness of a physical inductive bias. Self-advection provides a direct example: the same transport mechanism is unstable with a central-difference stencil but effective with an upwind discretization. A physically appropriate continuous operator can therefore become detrimental when represented by an unsuitable discrete approximation. Several development results were likewise attributable to factors other than physical bias. The overcomplete encoder improved two families but added 41M parameters, while an apparent gain from an early mechanism bank disappeared after correction of a channel-dilution artifact in the evaluation metric. Gradient surgery provided no benefit because measured inter-family gradient cosines were uniformly positive, indicating capacity rather than gradient conflict as the dominant limitation. Positive results in Supplementary Table 5 are therefore evaluated against parameter-matched baselines under identical scoring protocols.

Steady families are represented as relaxation toward the fixed point of the corresponding elliptic balance. Four relaxation steps are evaluated through the same integration framework and supervised against the steady solution, with greater weight assigned to later steps. Mechanisms specific to transient dynamics, including the wave term, upwind shock capturing, and buoyancy, are excluded from the steady architecture, while the remaining vocabulary represents the steady balance through elliptic coupling, diffusion, advection, dynamic pressure, near-wall operators, and curvature. Mechanism heads introduced after pretraining are zero-initialized and loaded by name, preserving the pretrained operator at initialization. Subsequent optimization can nevertheless degrade performance when an added pathway is mismatched to the target problem, so additional mechanisms are evaluated against matched-budget base models rather than assumed to be performance-neutral.

Supplementary Note 4: pretraining protocol and training history

All models are pretrained using the same two-stage objective. The amplitude stage uses a per-channel relative L_2 loss normalized by the full-field norm $\|u\|$, followed by a

fluctuation stage normalized by the mean-removed norm $\|u - \bar{u}\|$. The second stage is initialized from the best amplitude-stage checkpoint and uses a renewed linear warmup with a reduced constant learning rate, from 3×10^{-4} to 5×10^{-5}. Supplementary Figure 3 shows the corresponding training histories for both model sizes. Validation is monitored every 2,000 updates using eight trajectories per family. The final monitor errors are 7.11% for MOTION and 8.56% for MOTION-S, corresponding to 7.03% and 8.22% on the full validation split (Supplementary Note 1).

The transition to the fluctuation objective produces a consistent reduction in validation error for both model sizes. Preservation of the channel mean during input normalization causes the full-field denominator to underweight fluctuations in high-mean channels, whereas the mean-removed denominator restores their contribution to the loss. Both models exhibit the same stage-wise behavior, with model capacity primarily determining the final error level. For the layered-wave family, the sample-wise pairing between the solution and the spatially varying wave-speed field $c(\mathbf{x})$ was additionally verified from the released data. The temporal variance of the solution correlates negatively with the paired speed field (-0.12 ± 0.04) and shows no corresponding relation after shuffling ($+0.02 \pm 0.04$), consistent with the expected physical dependence.

Supplementary Table 6 Class-averaged relative L_2 (%, physical) at the end of each stage, under the training monitor.

model	parameters	amplitude stage	fluctuation stage
MOTION	156.9M	9.35	**7.11**
MOTION-S	20.7M	10.15	**8.56**

Supplementary Note 5: classification of the transformations probed

Each transformation in Fig. 4 is assigned one of three roles before model evaluation. Exact transformations preserve the governing equations and boundary conditions and act on the stored lattice without interpolation, including reflections and 90° rotations of isotropic square domains and cyclic translations of verified periodic domains. Admissible transformations preserve the physical state only within a measured finite-domain or discretization error, which is reported separately. Transformations that violate the governing equations or boundary conditions are classified as not admissible and excluded from the defect values reported in the main tables.

Translation admissibility is determined using the boundary-seam criterion described in Methods. Seam differences comparable to interior first differences indicate periodic compatibility, whereas substantially larger seam differences exclude translation. Cases in which the disturbance has not reached the boundary are treated separately as trajectory-local admissible transformations, while fixed Dirichlet boundaries exclude translation. All assignments are determined from the governing

equations, boundary conditions, and stored fields without reference to model predictions. As an independent consistency check, Supplementary Table 7 reports the replica accuracy ratio, defined as the transformed-copy error relative to the original-copy error. Exact symmetries are expected to preserve task difficulty, and no systematic discrepancy is observed.

Supplementary Table 7 Classification of all probed transformations according to the criteria described in Supplementary Note 5. Replica accuracy ratios are computed for MOTION using the metric associated with each equation family. For the unconditioned buoyant family, the ratios were obtained from a probe evaluated over all frames rather than the valid-frame subset used in Fig. 4b, and the corresponding absolute errors therefore differ. Entries marked "floor" indicate cases in which the identity error is too small for the ratio to provide a meaningful distinction; these cases are discussed separately in the text.

family	transform	data-level criterion	replica ratio	role
SWE	mirrors	isotropic; boundary constant in-window	1.000	exact
SWE	rotations	isotropic, square lattice	1.18 (floor)	exact
SWE	translation	seam 0.000, edge variation 0.000	1.09–1.78 (floor)	admissible
Comp. NS	mirrors	periodic, isotropic	1.00–1.02	exact
Comp. NS	rotations	periodic, isotropic	1.00–1.02	exact
Comp. NS	translation	seam 0.76–0.84 = interior	1.001	exact
Incomp. NS	mirrors	periodic, isotropic	0.98–1.00	exact
Incomp. NS	rotations	periodic, isotropic	0.99–1.01	exact
Incomp. NS	translation	seam 2.0× interior	1.06	admissible
PDEArena	mirror ⊥ gravity	wall–wall, forcing invariant	1.12	exact
PDEArena	mirror along gravity	gravity fixes the axis	**6.87**	not adm.
PDEArena	rotations	anisotropic	**7.64**	not adm.
PDEArena	translation	seam 5–16× interior	1.60–1.85	not adm.
PDEArena (unc.)	mirror ⊥ gravity	wall–wall, forcing invariant	1.005	exact
PDEArena (unc.)	mirror along gravity	gravity fixes the axis	**3.62**	not adm.
PDEArena (unc.)	rotations	anisotropic	**3.92**	not adm.
CFDBench	mirror ⊥ flow	inflow fixes the axis; channels scalar	1.29 (floor)	admissible
CFDBench	translation	seam 71× interior across the channel	—	not adm.
NS-PwC	mirrors, rotations	periodic, isotropic	≈1	exact
NS-PwC	translation	seam 0.65–0.67 = interior	1.007	exact
ACE	mirrors, rotations	isotropic, same boundary type	≈1	exact
ACE	translation	seam 8.8–10.2×	1.96–1.98	not adm.
Poisson	mirrors, rotations	Dirichlet, square domain	0.99–1.05	exact
Poisson	translation	Dirichlet	**4.27**	not adm.
Wave	mirrors	medium transforms with the state	1.00–1.06	exact
Wave	rotations	layered medium is anisotropic	**1.25–1.33**	not adm.

The buoyant family provides a clear example of the transformation taxonomy. The governing equations permit reflection perpendicular to gravity but not along the gravity axis, and the measured defects follow this distinction. The admitted mirror gives a 7.18% defect against a 7.73% prediction error, whereas reflection along gravity and 90°

rotation increase the defect to 52.7% and 58.8%, respectively. MOTION encodes this anisotropy through role-aware spatial operators and gravity-specific mechanism heads, without symmetry augmentation or explicit equivariance constraints. The resulting near-equivariance is therefore consistent with the prescribed physical structure rather than learned transformation labels.

Figure 4b normalizes the equivariance defect by prediction error to permit comparison across families of different intrinsic difficulty. Supplementary Figure 4 additionally reports replica accuracy ratios and absolute defect magnitudes. The normalized ratio alone can overemphasize violations for highly accurate models because a fixed absolute defect is divided by a smaller prediction error. Absolute defect and normalized ratio are therefore reported together to distinguish representation sensitivity from the scale of the underlying prediction error.

Two cautions apply in reading the table. The replica ratio loses resolving power where the identity error sits at the floor of the metric; for channel flow the decisive evidence is instead the channel convention—its two stored channels are speed-like scalars (correlation 0.99), and transforming them as scalars gives a defect of 0.170% within the triangle bound of 0.293%, whereas treating them as vector components inflates the same defect to 23.7%. And the magnitude of a defect alone says nothing about symmetry: the elliptic mirror defect of 5.67% is large only because the accuracy it is measured against is 6.26%, with every group element leaving the replica 0.99–1.05 times as hard as the original—which is why the error column sits beside the defect columns in Fig. 4.

Supplementary Note 6: transform response profiles

Figure 4 reports the worst-case defect for each transformation class, whereas the full response profile provides additional information about its origin. For shallow water, the sequence baseline shows a strong dependence on alignment with the patch grid rather than on translation magnitude. A one-cell shift produces a 0.487% defect, while an eight-cell shift, equal to the patch size and therefore an exact permutation of the token grid, increases the defect to 1.211% (Supplementary Table 8). This behavior is inconsistent with a simple displacement-magnitude effect and instead indicates sensitivity to the absolute patch representation. MOTION shows the opposite trend, with defects of 0.080% and 0.022% for one- and eight-cell shifts, respectively. A similar inversion appears for compressible flow, while the incompressible case follows the boundary-seam effect described in Supplementary Note 5 rather than patch-grid alignment.

Supplementary Note 7: transform-resolved ensemble averaging

Figure 4 reports the ensemble gain over the full admissible transformation set for each family. Supplementary Table 9 resolves this gain by transformation subgroup. Four averaging sets are considered: K_4, comprising the identity and three lattice reflections; K_8, the full dihedral group of the square lattice including 90° rotations; T_3, comprising the identity and cyclic translations by one and eight cells; and $K_8 \times T_3$, used for

Supplementary Table 8 Periodic-translation equivariance defect (%) resolved by roll magnitude. An eight-cell roll corresponds to the patch size of the sequence baselines and therefore permutes the token grid exactly, whereas a one-cell roll shifts the field within each patch. The main table reports the larger defect of the two translations for each family.

family	roll	MOTION	BCAT
SWE	1 cell (within a patch)	0.080	0.487
SWE	8 cells (patch-aligned)	**0.022**	1.211
Comp. NS	1 cell	0.223	0.317
Comp. NS	8 cells	0.203	0.490
Incomp. NS	1 cell	0.409	5.456
Incomp. NS	8 cells	0.652	5.387

isotropic periodic families. All models are evaluated using the same admissible set and the standard class-averaged relative L_2 metric.

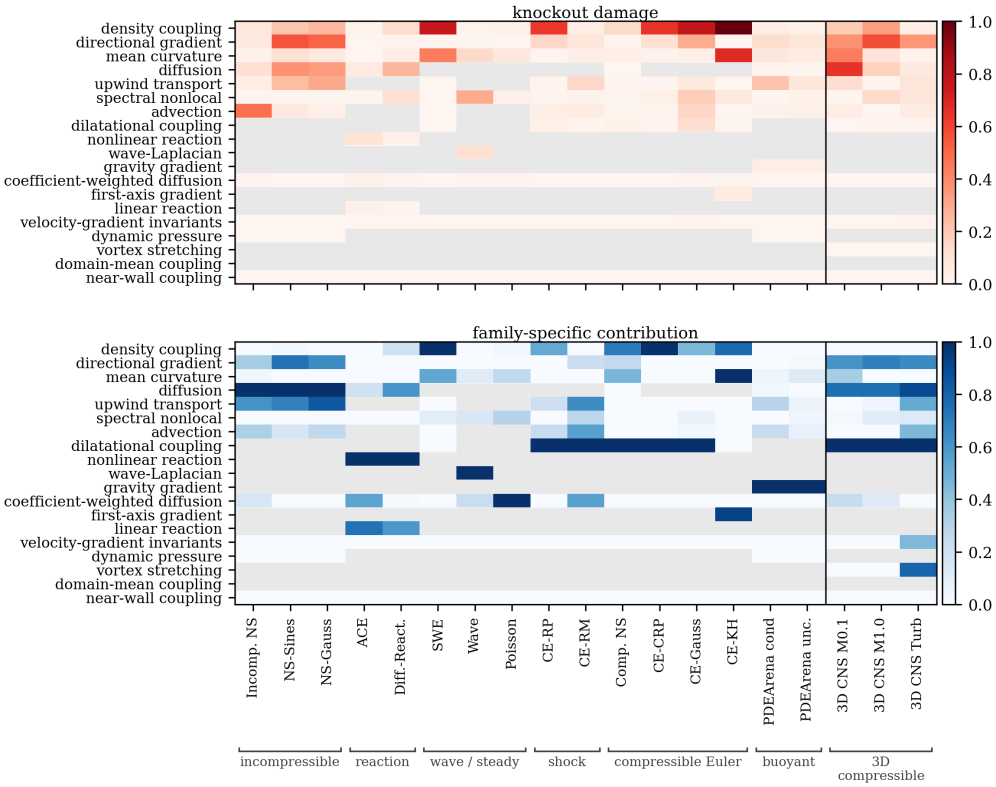
Supplementary Table 9 Prediction error under each transformation average and its change relative to the unaveraged result. Blank entries denote unavailable measurements, and dashes indicate transformations not admitted by the corresponding family. Translation is excluded for the elliptic and reaction tasks because these domains are non-periodic. The joint benchmark is compared with BCAT and PROSE-FD, while the transfer benchmark is compared with Poseidon-B.

model	family	unaveraged	K_4	K_8	T_3	$K_8 \times T_3$
MOTION	shallow water	0.055	0.045	0.047	0.063	**0.053**
BCAT	shallow water	0.207	100.0	100.0	0.499	**100.0**
PROSE-FD	shallow water	0.203	100.0	100.0	1.461	**100.0**
MOTION	compressible NS	1.565	1.371	1.335	1.559	**1.332**
BCAT	compressible NS	3.893	3.831	3.817	3.879	**3.815**
PROSE-FD	compressible NS	3.770	3.742	3.735	3.758	**3.729**
MOTION	incompressible NS	1.976	1.761	1.720	1.992	**1.726**
BCAT	incompressible NS	5.174	3.767	3.446	4.535	**3.267**
PROSE-FD	incompressible NS	7.483	6.584	6.418	7.330	**6.669**
MOTION	NS-PwC	3.283	2.264	2.116	3.188	**2.059**
Poseidon-B	NS-PwC	4.171	3.523	6.446	3.990	**6.437**
MOTION	Poisson-Gauss	6.257	5.207	**5.032**	—	—
Poseidon-B	Poisson-Gauss	11.170	7.784	**7.798**	—	—
MOTION	ACE	0.544	0.503	**0.494**	—	—
Poseidon-B	ACE	0.700	0.523	**0.453**	—	—

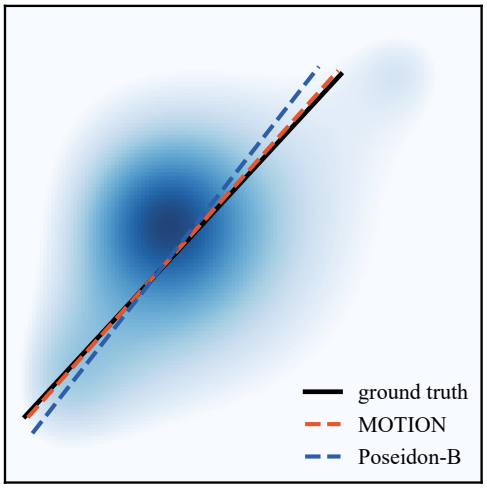
The transform-resolved results identify which symmetries determine the ensemble behavior. For shallow water, reflection averaging causes both sequence baselines to approach the 100% error ceiling, while translation averaging increases BCAT and PROSE-FD errors by factors of 2.4 and 7.2, respectively. MOTION remains below its unaveraged error under the same transformations. On the NS-PwC vector task,

Poseidon-B benefits from reflections and translations individually, but inclusion of 90° rotations changes the gain to a 54% loss. Under the same full group, MOTION improves from 3.283% to 2.059%. The equivariance defects show the same distinction: Poseidon-B reaches an 11.3% defect under rotation against an approximately 4% prediction error, while its reflection and translation defects remain below 4%. These results motivate averaging over the full physically admissible transformation set rather than a selectively chosen subgroup.

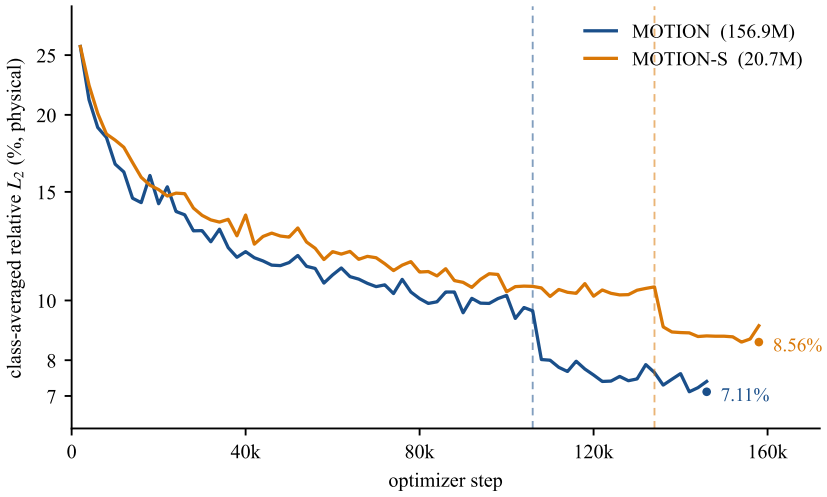
A larger transformation set does not necessarily improve ensemble performance. For shallow water, translation averaging increases MOTION's error by 14.5%, as its unaveraged error (0.055%) is smaller than the discrepancy introduced by a one-cell translation. The same behavior appears in the equivariance-defect analysis, where this transformation produces MOTION's largest defect-to-error ratio. Thus, ensemble benefit depends not only on the validity of the transformation but also on whether transformation-induced discrepancies remain below the intrinsic prediction error.



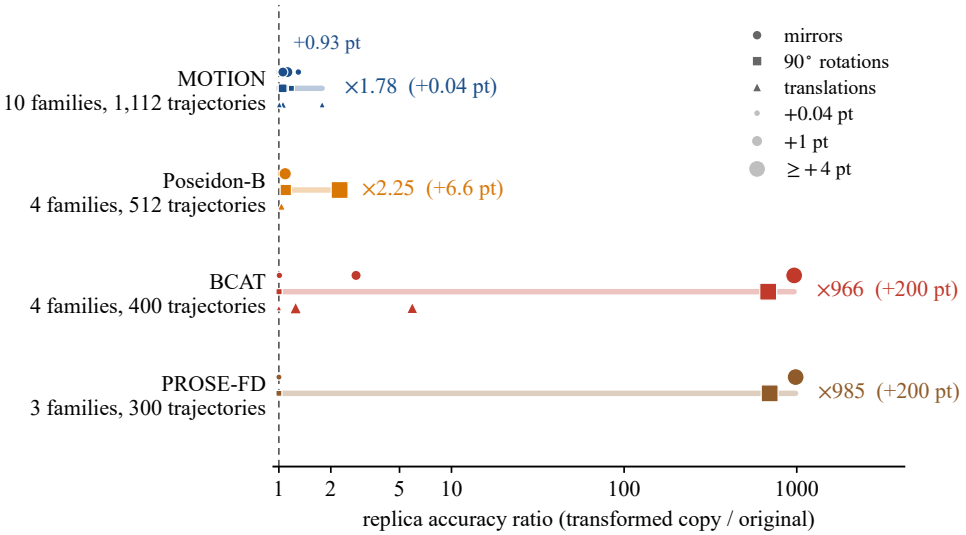
Supplementary Figure 1 Knockout damage and family-specific contribution for the full mechanism vocabulary, using the ordering and normalization of Fig. 1c. Grey cells denote inactive mechanisms, either through closed metadata gates or structural zeros. In two-dimensional flow, vortex stretching is identically zero because planar motion cannot tilt or stretch vorticity. The domain-mean head is inactive for all families in the corpus.



Supplementary Figure 2 Median-subtracted positive field for the example shown in Fig. 3c. Principal axes of the ground truth and model predictions are drawn through the corresponding weighted centroids.



Supplementary Figure 3 Two-stage pretraining of MOTION (156.9M) and MOTION-S (20.7M) across the nineteen equation families. Class-averaged relative L_2 error in physical space is shown against optimizer step and evaluated every 2,000 steps on eight held-out trajectories per family. Dashed lines indicate the transition from the amplitude stage to the fluctuation stage for each model.



Supplementary Figure 4 Replica accuracy ratios corresponding to Fig. 4b. Each point denotes the error of the inverse-transformed replica normalized by the error of the original prediction, evaluated only for admitted transformations. The worst case in each row is annotated with its ratio and absolute error increase in percentage points, which is also encoded by marker size. Row labels indicate the number of equation families and held-out trajectories evaluated for each model.