POISSON-GENERIC NEURAL OPERATORS: EXACT METRIPLECTIC STRUCTURE IN FUNCTION SPACE VIA CASIMIR ENTROPIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Every existing construction of thermodynamically consistent neural operators imposes the two GENERIC degeneracy conditions on a learned energy and entropy, but degeneracy is always obtained by projecting the reversible operator onto the complement of the entropy gradient. The consequence is an operator that depends on state and a Jacobi identity that is forfeited, so the resulting structure is metriplectic-degenerate rather than metriplectic. Our approach closes this gap by taking degeneracy from the same place GENERIC itself takes it. For nonlinear transport, the reversible operator L is the compatible Lie–Poisson pencil $\alpha D + \lambda(uD + Du)$; otherwise it is a constant Fourier multiplier, which is trivially Poisson. On an augmented state (u, s) carrying a latent entropy density, the entropy $S = \int s$ is a Casimir of L, so $L\,\delta S/\delta z = 0$ holds identically, with no projection required. A fixed mechanical quadratic, a learned gauge-free potential, and a convex internal energy together make up the energy. The friction operator $M = AA^\top$ satisfies $M\,\delta E/\delta z = 0$ pointwise, and its Onsager parity structure permits diffusion and damping while provably ruling out transport. For any parameters, skewness, positivity, both degeneracies, and the Jacobi identity (on the resolved band for the Lie–Poisson term) all hold to machine precision. Heat conduction and damped waves have closed-form friction operators that are representable exactly, the second law bounds physical energy under a checkable curvature condition, and exact discrete first and second laws follow from a discrete-gradient integrator. The model was tested on four PDEs in one and two dimensions with three backbones (FNO, Transolver, CNO): with the same backbone it wins 61 of 72 seed-level comparisons against unconstrained baselines, learns the exact transport and wave symbols, reproduces the true dissipation rate to within 13% on heat and Burgers, and dissipates nothing on advection. Whereas a constant-L ablation isolates the price of exact Jacobi as the loss of Burgers, a learned-entropy ablation injects energy on every reversible–irreversible problem.

1 INTRODUCTION

Neural operators do not know anything about energy or entropy. When solving PDEs with neural operator networks, there is nothing inherent in the loss function that tells the system to conserve energy or produce entropy. The errors that accumulate over a long rollout are therefore physical as well as numerical: the total energy of a heat equation surrogate grows, a wave surrogate gains amplitude, a Burgers surrogate stops dissipating and blows up. To address this, we employ the GENERIC formalism of non-equilibrium thermodynamics (Grmela & Öttinger, 1997; Öttinger, 2005).

Any closed system with an energy E and an entropy S evolves as

$$\partial_t z = L\,\frac{\delta E}{\delta z} + M\,\frac{\delta S}{\delta z}, \qquad L\,\frac{\delta S}{\delta z} = 0, \qquad M\,\frac{\delta E}{\delta z} = 0, \tag{1}$$

with L a Poisson operator and M symmetric positive semi-definite. The two degeneracy conditions make the first and second laws hold, and the Jacobi identity on L makes the reversible part a true

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.

Hamiltonian flow, rather than merely a skew flow. Applying this formalism makes the neural operators fully thermodynamically consistent—that is, they are guaranteed to obey both the first and second laws of thermodynamics.

Building a thermodynamically consistent neural operator. GENERIC-FNO (Sulskis & Ravi, 2026) carried the two degeneracy conditions into function space for the first time. It learns E and S as neural operators and sandwiches diagonal Fourier multipliers between rank-one projections, $L = (I - P_S)D_L(I - P_S)$, so that $L\,\delta S/\delta z = 0$ holds identically. However, this was done in a limited way, as the projection makes L depend on the state through $\delta S/\delta z$. Additionally, that work neither imposed nor checked the Jacobi identity for the projected bracket, so its structure was metriplectic-degenerate rather than a full Poisson–metriplectic pair. This same limitation runs through all the current finite-dimensional learned-bracket literature (Lee et al., 2021; Zhang et al., 2022; Gruber et al., 2023; 2025), where Jacobi is guaranteed only for a state-independent L.

Making it Poisson. The projected bracket has a normalized Jacobi residual of 10^{-3} to 1 on random test functionals, whereas our constant Fourier-multiplier operator achieves $\mathcal{O}(10^{-16})$. It satisfies the Jacobi identity exactly because $\partial L = 0$, and degeneracy follows from constructing S as a Casimir. This also reveals a limitation of the single-field formulation: for heat, Burgers, and advection, the Casimirs of a constant multiplier depend only on untouched modes and cannot provide an entropy that drives diffusion. This reflects the absence of nontrivial Poisson–Casimir structure in these equations, not a flaw in the construction. We therefore augment the state with a latent entropy density s, define $S = \int s$ with $\delta S/\delta z = (0, 1)$, and use an energy comprising a mechanical quadratic term, a learned potential, and a learned convex internal energy. Dissipation acts through $M = AA^\top$; its Onsager structure permits diffusion and damping but not transport, while the degeneracy condition $M\,\delta E/\delta z = 0$ is enforced pointwise by removing the local energy flux. This parameterization contains the exact friction operators for heat conduction and the damped wave (Proposition 6) and, together with energy conservation and a lower-bounded internal energy, bounds the mechanical energy for all time (Theorem 2), which a freely learned energy does not. Nonlinear transport is provided by the compatible Lie–Poisson term $\lambda(u\partial + \partial u)$, whose learned Burgers coefficient converges to $-\Delta t/3$. Finally, constant L and quadratic mechanical energy enable exact mode-wise integration of the linear reversible flow, while a discrete-gradient step conserves E and produces S exactly in discrete time.

Our main contributions are:

- A neural operator with full metriplectic structure in function space: a Poisson L, a Casimir entropy on an augmented thermodynamic state, and an Onsager-structured friction operator, with skewness, symmetry, positivity, both degeneracies and the Jacobi identity holding exactly for any parameters (§3, §4).
- Exact representability of the closed-form GENERIC friction operators for heat conduction and the damped wave, and a bound on the physical energy for all time that follows from the first and second laws for any parameters whose projected quadratic has curvature below $\frac{1}{2}$ (Proposition 6, Theorem 2).
- A discrete-gradient integrator that gives exact discrete first and second laws, including with the state-dependent Lie–Poisson term, and an exact unitary integration of the linear reversible flow (§A.5).
- A study across four PDEs, two dimensions and three backbones that reports the wins, the systematic loss on heat, the dissipation calibration against ground truth, and the two ablations that separate a physical energy from a learned one and exact Jacobi from nonlinear transport (§4).

2 Related Work

Neural operators. Neural operators learn discretization-consistent maps between function spaces (Kovachki et al., 2023). FNO parameterizes kernels spectrally (Li et al., 2021), DeepONet uses branch–trunk factorization (Lu et al., 2021), and later architectures introduced wavelet and Laplace kernels (Tripura & Chakraborty, 2023; Cao et al., 2024), alias-free convolutional operators (Raonić

et al., 2023), and transformers for general geometries, including Transolver's physics attention (Wu et al., 2024; Luo et al., 2025) and universal physics transformers (Alkin et al., 2024). None encodes conservation or dissipation laws. As in Sulskis & Ravi (2026), we use FNO, Transolver, and a CNO-style backbone so controlled comparisons change only the thermodynamic structure. Here the backbone supplies only the entropy encoder, friction coefficients, and unconstrained baseline and, unlike GENERIC-FNO, maps fields to fields rather than to differentiable scalar functionals.

Structure-preserving learning in finite dimensions. Hamiltonian/Lagrangian networks and symplectic integrators learn reversible dynamics (Greydanus et al., 2019; Cranmer et al., 2020; Jin et al., 2020), while metriplectic methods address thermodynamic dynamics (Hernández et al., 2021; Yu et al., 2021; Lee et al., 2021; Zhang et al., 2022; Gruber et al., 2023; 2025). They enforce skewness of L, positivity of M, and both degeneracies, but Jacobi only when L is state-independent (Sulskis & Ravi, 2026). Poisson neural networks instead pull back a constant bracket through a learned symplectomorphism via Darboux–Lie (Jin et al., 2022); our constant L is the trivial pullback, with a function-space extension sketched in §A.8. Alternatively, compatible Hamiltonian operators remain Hamiltonian under linear combination, as in KdV's bi-Hamiltonian structure (Magri, 1978; Olver, 1993). We exploit this with the scalar-field pencil $\alpha D + \lambda(uD + Du)$, compatible for all parameters; on Burgers, learned λ converges to the coefficient making $-\partial_x(u^2/2)$ Hamiltonian under quadratic energy. To our knowledge, compatible pencils have not previously been used for learned brackets. Unlike dimension-scaling matrices, our L and M use spectral multipliers and pointwise fields on a fixed band, allowing the same parameters across one and two dimensions and resolutions.

Structure-preserving neural operators. Function-space structure has been imposed through soft physics residuals (Raissi et al., 2019), built-in conserved quantities (Liu et al., 2024), and symmetries (Helwig et al., 2023). GENERIC-FNO (Sulskis & Ravi, 2026) enforced both degeneracies for learned energy and entropy using rank-one projections around diagonal Fourier multipliers, introduced the gauge-invariant dissipation diagnostic r_{mech}, and transferred its guarantees across $4\times$ super-resolution. We retain its testbed and diagnostic, measure the Jacobi residual of its projected operator, and replace its learned functionals and operators with fixed mechanical plus learned internal energy, Casimir entropy, and new operators. Related graph/mesh metriplectic models (Tierz et al., 2025), metriplectic conditional flow matching (Baheri & Lindemann, 2025), and Metriplector (Oprisa & Toth, 2026) are discrete, finite-dimensional, or latent-grid constructions. None enforces Jacobi for a state-dependent bracket, obtains degeneracy through a Casimir, or constrains friction beyond positivity.

Casimirs, internal variables, and friction matrices in GENERIC. GENERIC supplies three ingredients. Field-theoretic entropy is a Poisson Casimir, and dissipative continua require mechanical variables augmented by internal energy or entropy density (Grmela & Öttinger, 1997; Öttinger, 2005), rooted in metriplectic theory (Kaufman, 1984; Morrison, 1986) and classical function-space brackets (Olver, 1993; Marsden & Ratiu, 1999; Morrison, 1998). Hydrodynamic friction has factored form $C\,D'\,C^{\top}$, depends on state through velocity gradients and temperature, and satisfies energy degeneracy through this dependence rather than projection (Öttinger, 2005, Ch. 2), consistent with Onsager theory (Onsager, 1931; de Groot & Mazur, 1962). Learning methods have instead worked in observed coordinates, learned S freely, and imposed degeneracy per gradient direction. We instead use latent entropy density with an initial-state encoder, physical quadratic plus learned pointwise energy, and a friction parameterization containing the exact heat and damped-wave operators while, by derivative-order parity, permitting diffusion and damping but excluding transport. To our knowledge, latent entropy density and this parity constraint are new to operator learning; §4 ablates both.

Structure-preserving integration. Discrete-gradient methods exactly conserve first integrals at any step size (Gonzalez, 1996; McLachlan et al., 1999; Celledoni et al., 2012); metriplectic integrators additionally require discrete gradients respecting degeneracy (Romero, 2009; Öttinger, 2018; Gruber et al., 2025). GENERIC-FNO used explicit Euler because higher-order integration broke exact discrete degeneracy. Here entropy is linear, so its discrete gradient is exact, while $\ker L$ contains every entropy direction at every state; hence the Gonzalez step preserves exact discrete first and second laws even with the state-dependent Lie–Poisson term, unlike brackets whose kernel moves with state. For the reversible part, a constant multiplier with quadratic mechanical energy yields an

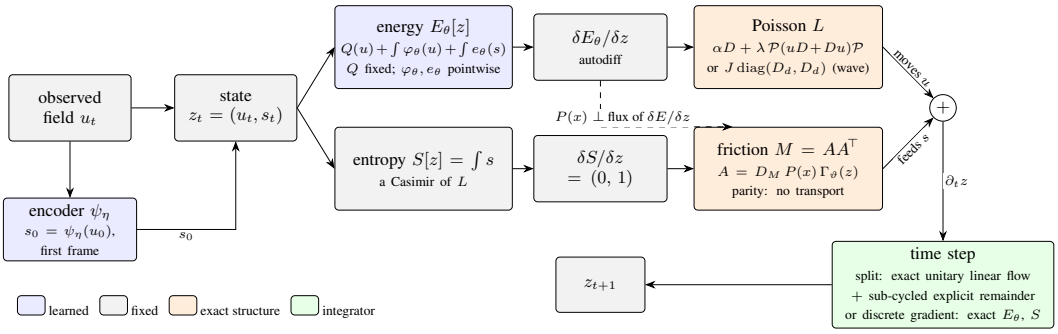


Figure 1: One step of the Poisson-GENERIC neural operator. The state $z = (u, s)$ combines the observed field with a latent entropy density, with $S = \int s$ a Casimir of L. The energy combines a fixed mechanical quadratic with learned pointwise terms. The Poisson operator governs the reversible dynamics, while the projected $AA^\top$ friction operator enforces $M\,\delta E/\delta z = 0$ and excludes transport. The channels are advanced by a split step, or by a discrete-gradient step that conserves E_θ and produces S exactly.

exactly integrable unitary map per mode, applied through Strang splitting (Strang, 1968). Although exponential and splitting methods for linear stiff or oscillatory systems are standard (Hochbruck & Ostermann, 2010), here the learned structure makes the exact flow available; an unconstrained operator must learn that map.

3 Method: Poisson-GENERIC Neural Operators

Figure 1 shows the forward map for a single frame. Learning is confined to three components: an encoder that derives the latent entropy density from the first observed frame; two pointwise constitutive functions within the energy; and the pointwise coefficient field of the friction operator, along with a small number of operator symbols. All remaining elements—the entropy, the mechanical energy, the structure of both operators, and the integrator—are fixed by construction, and this is the source of the guarantees.

State, energy, entropy. The pair $z = (u, s)$ defines the state, where u is the observed mechanical field and s is a latent entropy density. For heat, advection and Burgers, u carries $p = 1$ channel; for the damped wave, $u = (q, v)$. Being unobserved, s is trained solely through its influence on u, with a small encoder providing its initialization, $s_0 = \psi_\eta(u_0)$. Nothing is learned in the entropy, which is fixed at $S = \int s$, hence $\delta S/\delta z = (0, 1)$. Two learned pointwise terms join the physical quadratic to form the energy,

$$E_\theta[z] = Q(u) + \int_\Omega \varphi_\theta(u) + \int_\Omega e_\theta(s), \qquad Q = \tfrac{1}{2}\|u\|^2 \text{ or } \tfrac{1}{2}\big(\|v\|^2 + c^2\|\nabla q\|^2\big), \tag{2}$$

where the convex, increasing function e_θ gives a temperature $T = e'_\theta(s) \geq T_0 > 0$, and φ_θ is a softplus network minus its degree-≤ 2 polynomial part, so it need not be nonnegative. A quadratic hidden in φ_θ would otherwise absorb part of Q; projecting it out removes that gauge (§A.1). On the reversible side no projection is needed, since L acts on u alone and $L\,\delta S/\delta z = 0$ is an identity: as in every GENERIC field theory (Öttinger, 2005), S is a Casimir.

Poisson operator. For scalar fields, $L_u = \alpha D + \lambda\,\mathcal{P}(uD + Du)\mathcal{P}$, where $D = \sum_a \partial_a$, the two coefficients are learned scalars, and $\mathcal{P}$ is the 2/3-rule projector. Since $\{D,\, uD + Du\}$ is the compatible pair of the KdV hierarchy, L_u is Poisson for every (α, λ) (Magri, 1978; Olver, 1993); moreover, it is the Lie–Poisson term that makes Burgers Hamiltonian with a quadratic energy ($\lambda = -\frac{1}{3}$ up to the time scale). For the wave, $L_u = J\,\mathrm{diag}(D_d, D_d)$, with J canonical and $d(k)$ a learned even multiplier. Both operators are exactly skew; their constant parts are Poisson on any state, and the Lie–Poisson term is Poisson on the resolved band (Table 1, §A.2). A free learned multiplier $a(k)$, of the kind used in GENERIC-FNO, is not compatible with the Lie–Poisson term and so is not used.

Friction operator. Because $\delta S/\delta z = (0, 1)$, the coupling of u to s is the sole route by which u is dissipated, exactly as in the closed-form friction operators for heat conduction and the damped wave (§A.3, eq. 9), which depend on ∇u, T and $\delta E/\delta z$. Our construction is $M = AA^\top$ with $A = D_M\, P\, \Gamma_\vartheta(z)$: D_M is a multiplier with saturated flux (divergence) and local slots, Γ_ϑ a learned pointwise matrix field whose flux–local blocks are linear in ∇z, and $P(x)$ the pointwise projector off the local flux of $\delta E/\delta z$. Whatever the parameters and the state, M is symmetric positive semi-definite and $M\,\delta E/\delta z = 0$; every path from u back to u carries an even derivative order, so M can diffuse and damp but cannot transport (Proposition 5); and the closed-form heat and wave operators are recovered exactly on the band $\nu k^2 \Delta t < 1$, which contains the data band (Proposition 6, §4).

What is exact. Along $\partial_t z = L\,\delta E_\theta/\delta z + M(z)\,\delta S/\delta z$, E_θ is conserved and S is non-decreasing for any choice of parameters (Theorem 1). Since the softplus network is nonnegative, φ_θ is bounded below by minus its own projected quadratic, and the second law gives a uniform bound on the mechanical energy whenever that quadratic has curvature $a < \frac{1}{2}$:

$$Q(u_t) \;\le\; \frac{E_\theta(z_0) - T_0\, S(z_0) + C_\theta}{1 - 2a}, \tag{3}$$

with C_θ fixed by the linear and constant parts of that quadratic (Theorem 2). The curvature is the gauge of Appendix A.1, and it is measurable: the learned reversible symbols of §4 match the exact ones only if L carries the whole linear flow, which bounds $|a| \le 3 \times 10^{-3}$ on every reversible cell (Remark 2). The bound and the conservation laws are exact properties of the model; whether the flow empirically matches the equations is what §4 seeks to answer.

Integration. The constant part of L with the quadratic Q generates a unitary map per Fourier mode, which we apply exactly in a Strang split around an explicit, sub-cycled step for the Lie–Poisson term, the potential and the friction; this is the integrator we train and evaluate with, and it is the map an unconstrained FNO has to learn on advection. The implicit discrete-gradient step of §A.5 conserves E_θ and produces S exactly in discrete time (Theorem 3), because $\ker L$ contains every entropy direction at every state; when its fixed-point solve converges (§A.5), it is the reference for the split step's energy error.

Diagnostics and ablations. All diagnostics are independent of learned targets: the Jacobi residual $\mathcal{J}$ on random test functionals; representability of the exact friction operators; learned reversible symbols against $-ck\Delta t$, $d = \Delta t$ and $\lambda = -\Delta t/3$; dissipation Π_{model} against $\Pi^\star$ in the same normalization; and the reversibility cosine r_{mech} of Sulskis & Ravi (2026), using a midpoint gradient so exact Q-conservation scores zero (§A.6). Two ablations isolate the main design choices: *learned* E, S, the closest constant-L analogue of GENERIC-FNO, and *constant* L, the model of record with $\lambda \equiv 0$ (§A.7).

4 Experiments

Setup. We test the four periodic equations of Sulskis & Ravi (2026) in 1D and 2D: heat ($\nu = 0.05$), advection ($c = 1$), viscous Burgers ($\nu = 0.02$), and damped wave ($c = 1, \gamma = 0.3$, observed as (q, v)). Data use band-limited random initial conditions, a de-aliased spectral solver, and $\Delta t = 0.1$; persistence errors are 0.18–1.03 (Table 2). FNO (Li et al., 2021), Transolver (Wu et al., 2024) (with a kernel-3 circular depthwise convolution on the grid before each block; Appendix B), and a periodic CNO-style operator serve both as unconstrained baselines, $u_{t+1} = u_t + G(u_t)$, and as backbones, isolating the effect of structure. Each cell compares the baseline, learned-E, S ablation, PG-phys, and PG-phys with $\lambda = 0$. Results are ten-frame held-out rollouts, mean $\pm$ std over three seeds; every model receives the same training protocol (Appendix B).

Structure and representability. Table 1 verifies the identities at random initialization in double precision. Skewness, symmetry, both degeneracies, and $\dot{E} = 0$ hold to 10^{-16} or better. Constant L satisfies Jacobi to 10^{-16}, versus 10^{-3} to 1 for the projected operator of GENERIC-FNO; the Lie–Poisson term gives 10^{-16} on band-limited states but 0.17–0.78 on generic states because of aliasing. The parameterization reproduces the exact heat and wave friction operators to 3–7 $\times 10^{-16}$ relative error in both dimensions, with $M\,\delta E/\delta z = 10^{-17}$–$10^{-18}$, without global projection.

Table 1: Structural identities at random initialization, double precision, 16^d grid, maximum over a batch; Lrn is the learned-E, S ablation, Phys the physical model of record. $\mathcal{J}$ is the normalized Jacobi residual of §A.6; "generic" and "band-limited" refer to the test states and functionals. The Lie–Poisson rows are empty for the wave, whose operator has no pencil; the projected-L row uses a free learned entropy in both columns, since with a Casimir entropy the projection is trivial. Discrete-gradient entries are at the tolerance of the fixed-point solve, not the floating-point floor. Entries at or below the double-precision floor are written $\leq 10^{-16}$.

	1D heat		2D Burgers		1D wave		2D wave	
	Lrn	Phys	Lrn	Phys	Lrn	Phys	Lrn	Phys
$\langle v, Lv\rangle$ (skewness)	$\leq 10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$
$\langle w, Mv\rangle - \langle v, Mw\rangle$ (symmetry)	$2{\cdot}10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$	$1{\cdot}10^{-16}$	$1{\cdot}10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$
$\|L\,\delta S/\delta z\|/\|\delta S/\delta z\|$	$4{\cdot}10^{-18}$	0	$8{\cdot}10^{-18}$	0	$2{\cdot}10^{-18}$	0	$8{\cdot}10^{-19}$	0
$\|M\,\delta E/\delta z\|/\|\delta E/\delta z\|$	0	$3{\cdot}10^{-18}$	0	$2{\cdot}10^{-18}$	0	$1{\cdot}10^{-19}$	0	$4{\cdot}10^{-20}$
$\langle \delta E/\delta z, \partial_t z\rangle$ ($\dot{E}$)	$1{\cdot}10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-16}$	$\leq 10^{-17}$	$\leq 10^{-16}$	$\leq 10^{-17}$
$\min \langle \delta S/\delta z, \partial_t z\rangle$ ($\dot{S}$)	0.58	0.014	0.68	0.060	0.34	0.31	0.78	0.0027
$\mathcal{J}$, our L, generic state	$7{\cdot}10^{-16}$	$1{\cdot}10^{-16}$	$1{\cdot}10^{-16}$	$3{\cdot}10^{-16}$	$3{\cdot}10^{-16}$	$4{\cdot}10^{-16}$	$2{\cdot}10^{-16}$	$2{\cdot}10^{-16}$
$\mathcal{J}$, Lie–Poisson ($\lambda = 0.3$), generic	–	0.78	–	0.17	–	–	–	–
$\mathcal{J}$, Lie–Poisson ($\lambda = 0.3$), band-limited	–	$3{\cdot}10^{-16}$	–	$4{\cdot}10^{-16}$	–	–	–	–
$\mathcal{J}$, projected L of GENERIC-FNO	1.0	0.38	$5.0{\cdot}10^{-2}$	$4.8{\cdot}10^{-2}$	$1.4{\cdot}10^{-2}$	$2.5{\cdot}10^{-2}$	$7.6{\cdot}10^{-4}$	$1.9{\cdot}10^{-3}$
$\lvert E_1 - E_0\rvert/\lvert E_0\rvert$, discrete-gradient step	$1{\cdot}10^{-16}$	$1{\cdot}10^{-11}$	$2{\cdot}10^{-16}$	$8{\cdot}10^{-11}$	$3{\cdot}10^{-16}$	$2{\cdot}10^{-14}$	$1{\cdot}10^{-16}$	$5{\cdot}10^{-13}$
$\lvert E_1 - E_0\rvert/\lvert E_0\rvert$, explicit step	$2{\cdot}10^{-11}$	$5{\cdot}10^{-2}$	$7{\cdot}10^{-11}$	$5{\cdot}10^{-2}$	$2{\cdot}10^{-11}$	$5{\cdot}10^{-4}$	$8{\cdot}10^{-12}$	$3{\cdot}10^{-3}$

Accuracy. PG-phys beats its baseline in 20 of 24 cells and 61 of 72 seed-level comparisons (13 of 16 and 39 of 48 on FNO and CNO; Table 2). Seventy of 72 rollouts remain finite and satisfy equation 3; one Transolver 1D-wave seed diverged in training and one Transolver 2D-heat seed was unstable under the split step ($\max_t Q/\text{bound} = 1.67$; marks in Table 2). Both are seed- and hardware-specific: retraining the same two seeds gave a finite 1D-wave rollout (0.040) and a stable 2D-heat rollout (0.033, $\max_t Q/\text{bound} = 0.15$). PG-phys reduces error by 2.7–33× on advection, 1.2–8.7× on the wave, and 2.2–3.5× on Burgers. Heat is the exception: PG-phys loses all three 1D-FNO seeds (0.017 vs. 0.0014) and the three 2D cells by mean, and beats the Transolver only in 1D, where that baseline is budget-limited (Appendix B). The learned reversible operators recover their known targets directly: the advection symbol matches $-ck\Delta t$ to three decimals in all nine 1D runs, $d(k) = 0.1004 \pm 0.0005$ on the wave, and $\lambda = -0.0331 \pm 0.0003$ on Burgers versus $-\Delta t/3 = -0.0333$. These symbol comparisons also expose structural errors that rollout accuracy can hide.

Thermodynamics. PG-phys recovers 0.87–1.00 of the true heat/Burgers dissipation rate in eleven of twelve cells and 1.12 in the twelfth (2D Transolver heat, two stable seeds), and 0.75–0.94 on the wave (Table 3). Advection is effectively nondissipative ($|\Pi_{\text{model}}| \leq 2 \times 10^{-4}$), with Q conserved to 4×10^{-3} and $r_{\text{mech}} = 10^{-6}$–$7 \times 10^{-4}$. The measured ratio $\max_t Q/\text{bound}$ is below one on 70 of 72 seeds under the split step, and below 0.9 on every cell of Table 3. By contrast, learned-E, S satisfies the same machine-precision identities while physical energy grows 2.3–5.8× on the wave and 1.6–2.6× on Burgers despite conserving E_θ to 10^{-8}. Figure 2 follows the physical energy for 40 frames, ten times the training horizon: PG-phys stays on the truth on all three equations, learned-E, S loses a quarter of the energy on advection and multiplies it six- and thirty-fold on Burgers

Table 2: Ten-frame rollout relative L^2 error, mean $\pm$ std over three seeds, with the persistence baseline and parameter counts (baseline / PG-phys, thousands). Bold: best of the four models (for the wave, whose operator has no pencil, the two PG columns are one model trained twice and are both marked; their differences are run-to-run noise). †: one of three seeds diverged; mean $\pm$ std of the other two. ‡: one seed finite but unstable under the split step ($\max_t Q/\text{bound} = 1.67$); three-seed statistics shown. Neither failure reproduced on retraining.

		Persist.	Baseline	Learned E, S	PG-phys	PG-phys ($\lambda \equiv 0$)	Params
1D heat	FNO	0.184	$\mathbf{0.0014 \pm 0.0003}$	0.0096 ± 0.0066	0.017 ± 0.005	0.0056 ± 0.0011	44 / 18
	Transolver	0.184	0.088 ± 0.050	0.018 ± 0.006	0.015 ± 0.007	$\mathbf{0.0087 \pm 0.0049}$	34 / 41
	CNO	0.184	0.015 ± 0.011	0.0097 ± 0.0031	0.013 ± 0.006	$\mathbf{0.0046 \pm 0.0008}$	21 / 12
1D advection	FNO	0.863	0.0040 ± 0.0004	0.019 ± 0.006	0.0015 ± 0.0007	$\mathbf{0.0001 \pm 0.0000}$	44 / 18
	Transolver	0.863	0.029 ± 0.004	0.094 ± 0.096	0.0016 ± 0.0006	$\mathbf{0.0001 \pm 0.0001}$	34 / 41
	CNO	0.863	0.012 ± 0.002	0.035 ± 0.008	0.0017 ± 0.0010	$\mathbf{0.0001 \pm 0.0001}$	21 / 12
1D Burgers	FNO	0.645	0.090 ± 0.010	0.419 ± 0.068	$\mathbf{0.040 \pm 0.003}$	0.331 ± 0.012	44 / 18
	Transolver	0.645	0.151 ± 0.047	0.378 ± 0.024	$\mathbf{0.043 \pm 0.004}$	0.386 ± 0.066	34 / 41
	CNO	0.645	0.072 ± 0.005	0.377 ± 0.003	$\mathbf{0.033 \pm 0.011}$	0.279 ± 0.004	21 / 12
1D wave	FNO	0.938	0.036 ± 0.011	0.813 ± 0.645	$\mathbf{0.030 \pm 0.007}$	$\mathbf{0.029 \pm 0.008}$	44 / 22
	Transolver	0.938	0.221 ± 0.022	0.729 ± 0.267	$\mathbf{0.027 \pm 0.010}^{\dagger}$	$\mathbf{0.027 \pm 0.010}^{\dagger}$	34 / 45
	CNO	0.938	0.104 ± 0.024	0.511 ± 0.212	$\mathbf{0.035 \pm 0.003}$	$\mathbf{0.036 \pm 0.002}$	21 / 16
2D heat	FNO	0.275	0.038 ± 0.002	0.146 ± 0.063	0.042 ± 0.006	$\mathbf{0.033 \pm 0.005}$	445 / 128
	Transolver	0.275	0.033 ± 0.007	0.021 ± 0.006	$0.086 \pm 0.095^{\ddagger}$	$\mathbf{0.020 \pm 0.002}$	34 / 51
	CNO	0.275	0.036 ± 0.010	0.144 ± 0.029	0.044 ± 0.005	$\mathbf{0.033 \pm 0.009}$	62 / 32
2D advection	FNO	0.960	0.074 ± 0.007	0.506 ± 0.102	0.0038 ± 0.0022	$\mathbf{0.0002 \pm 0.0001}$	445 / 128
	Transolver	0.960	0.057 ± 0.007	0.095 ± 0.019	0.0024 ± 0.0021	$\mathbf{0.0001 \pm 0.0000}$	34 / 51
	CNO	0.960	0.054 ± 0.003	0.504 ± 0.023	0.0016 ± 0.0013	$\mathbf{0.0001 \pm 0.0001}$	62 / 32
2D Burgers	FNO	0.761	0.275 ± 0.007	0.568 ± 0.017	$\mathbf{0.091 \pm 0.013}$	0.367 ± 0.026	445 / 128
	Transolver	0.761	0.254 ± 0.002	0.556 ± 0.131	$\mathbf{0.094 \pm 0.011}$	0.365 ± 0.021	34 / 51
	CNO	0.761	0.222 ± 0.005	0.586 ± 0.055	$\mathbf{0.084 \pm 0.010}$	$0.417 \pm 0.017^{\dagger}$	62 / 32
2D wave	FNO	1.026	0.305 ± 0.025	0.922 ± 0.156	$\mathbf{0.035 \pm 0.002}$	$\mathbf{0.035 \pm 0.002}$	445 / 146
	Transolver	1.026	0.210 ± 0.040	0.876 ± 0.556	$\mathbf{0.035 \pm 0.002}$	$\mathbf{0.035 \pm 0.002}$	34 / 69
	CNO	1.026	0.301 ± 0.024	0.811 ± 0.165	$\mathbf{0.035 \pm 0.002}$	$\mathbf{0.035 \pm 0.002}$	62 / 50

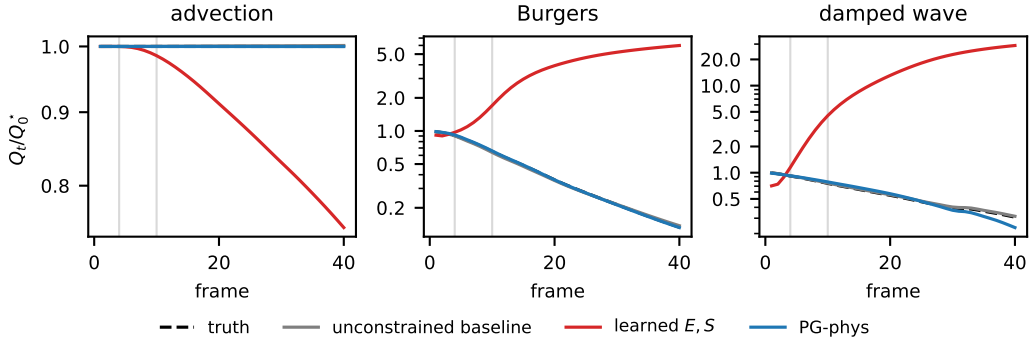


Figure 2: Physical energy $Q_t/Q_0^\star$ along a 40-frame rollout from held-out initial conditions (1D, FNO backbone, seed 0; grey lines mark the 4-frame training horizon and the 10-frame evaluation horizon). PG-phys stays on the truth on all three equations. The learned-energy ablation conserves its own E_θ while losing a quarter of the physical energy on advection and multiplying it six- and thirty-fold on Burgers and the wave. The unconstrained baseline matches the truth on advection and Burgers and drifts on the wave.

and the wave, and the unconstrained baseline drifts only on the wave ($+35\%$ at frame 40). Exact degeneracy constrains the model's functionals; physical energy constrains the physics.

Ablations. Setting $\lambda = 0$ gives exact Jacobi on every state and improves the linear equations: 10^{-4} advection error in every cell, better heat accuracy in all six cells, and unchanged wave accuracy. It cannot represent Burgers, however, where it is worse than baseline in all six cells and

Table 3: Dissipation calibration of PG-phys (three-seed means; FNO / Transolver / CNO): fractional one-step decay of the physical energy against the truth, physical energy after ten frames relative to the truth, and $\max_t Q$ relative to the bound equation 3, evaluated with $a = 0$ and $C_\theta = -\inf \varphi_\theta$ over the data range (Remark 2). †: two seeds (the third diverged in training). ‡: two seeds (the third was unstable under the split step, $\max_t Q/\text{bound} = 1.67$); neither failure reproduced on retraining (see Accuracy).

		$\Pi_{\text{model}}/\Pi^\star$	$Q_{10}/Q^\star_{10}$	$\max_t Q/\text{bound}$
1D	heat	0.94 / 0.99 / 0.96	1.02 / 1.00 / 1.01	0.16 / 0.17 / 0.14
	Burgers	1.00 / 1.00 / 1.00	1.00 / 1.00 / 1.00	0.13 / 0.14 / 0.10
	wave†	0.75 / 0.86 / 0.76	1.08 / 1.04 / 1.08	0.89 / 0.83 / 0.87
	advection	$\lvert\Pi_{\text{model}}\rvert \le 10^{-4}$	1.000 / 1.000 / 1.000	0.44 / 0.28 / 0.42
2D	heat‡	0.87 / 1.12 / 0.88	1.07 / 0.94 / 1.06	0.35 / 0.13 / 0.35
	Burgers	0.93 / 0.93 / 0.95	1.05 / 1.04 / 1.04	0.40 / 0.25 / 0.36
	wave	0.94 / 0.90 / 0.94	1.02 / 1.03 / 1.02	0.85 / 0.84 / 0.85
	advection	$\lvert\Pi_{\text{model}}\rvert \le 2 \cdot 10^{-4}$	0.998 / 0.998 / 0.999	0.44 / 0.41 / 0.40

one 2D-CNO seed diverges. Thus the Lie–Poisson pencil trades some linear accuracy for nonlinear transport within one model. The constant-L and learned-E, S ablations win 47 and 9 of 72 seed-level comparisons, respectively.

Super-resolution and integration. FNO PG-phys transfers directly across resolutions: for three 1D-heat seeds, errors at $64/128/256$ points are $0.01827/0.01826/0.01826$, $0.02162/0.02160/0.02159$, and $0.01117/0.01115/0.01113$, with thermodynamic diagnostics agreeing to four digits. CNO is excluded because its grid-defined kernels are not resolution-invariant; Transolver was not evaluated. Correct Riesz scaling of the variational derivative is essential: an earlier version with Euclidean scaling saw error grow from 0.011 to 0.078 to 0.125 across the same grids. Replacing the trained split step with $3/6/12$ discrete-gradient iterations reduces the per-step energy error from 1.2×10^{-5} to $3.5 \times 10^{-6}/1.9 \times 10^{-7}/2.6 \times 10^{-8}$, while rollout error changes only from 0.0183 to 0.0186 and the thermodynamic diagnostics remain unchanged to two digits; on a trained 2D-heat FNO model the same study gives 9×10^{-5} to $10^{-6}/10^{-7}/1.5 \times 10^{-7}$ with rollout error, $Q_{10}/Q^\star_{10}$ and $\max_t Q/\text{bound}$ unchanged to three digits. The fixed-point solve converges on the FNO models but not on the trained 2D-heat Transolver models (§A.5).

Cost. PG-phys uses 0.3–$0.8\times$ the parameters of the FNO/CNO baselines (1.2–$2\times$ for Transolver) but trains 25–$40\times$ slower (FNO, CNO; about $20\times$ for Transolver). The cost comes from eight automatic-differentiation evaluations of the energy gradient and friction coefficients per frame; the exact linear step requires only one FFT pair per step. A fused sub-cycled remainder should remove much of this overhead.

5 Limitations and Scope

Jacobi identity. For constant L, the Jacobi identity holds for every state. With the Lie–Poisson term, it holds to machine precision on band-limited states and functionals but incurs aliasing residuals of 0.2–0.8 on generic full-spectrum states. Thus, nonlinear transport is Poisson only on the resolved band. We restrict L to the compatible pencil and retain the constant-L variant for this reason. The dispersive D^3 member is excluded for stability, so dispersive equations remain outside the present scope.

Cost of the pencil. For linear equations, constant L performs better: by an order of magnitude on advection and a factor of 1.3–3 on heat. In the Lie–Poisson model, the transport coefficient and learned potential must converge to zero for a gradient flow while the friction operator is still being learned. Heat is the systematic loss: the only cell lost on all three seeds is heat with the 1D FNO, and the three 2D heat cells are lost by mean. This may reflect the added burden or a capacity mismatch; a targeted capacity and schedule sweep is the natural next test.

Baselines. All models share one protocol of roughly 350 optimizer steps. FNO and CNO converge within it; the Transolver baseline is still improving, and at five times the budget it reaches 0.005 on 1D heat and 0.007 on 1D advection (Appendix B). The Transolver column therefore compares equal budgets, not converged models, and against a fully trained Transolver PG-phys would lose heat and keep advection, Burgers and the wave.

Integration. The split step conserves mechanical energy exactly under the linear flow, but total energy only to $O(\Delta t^2)$, and one Transolver 2D-heat seed left the bound under it ($\max_t Q/\text{bound} = 1.67$; it did not reproduce on retraining). Exact discrete first and second laws require the implicit discrete-gradient step, which we use as a reference rather than during training, and whose fixed-point solve converges on the FNO models but not on the trained 2D-heat Transolver models (§A.5). The explicit remainder is stabilized by saturation and sub-cycling; at larger $\nu k^2 \Delta t$, an implicit or exponential friction step would be preferable.

Scope of the guarantees. The guarantees apply to the model's own functionals: E_θ is conserved and S is non-decreasing for any parameters, and the mechanical energy is bounded under the curvature hypothesis of Theorem 2, which the symbol diagnostic certifies for the trained reversible models (Remark 2). They do not guarantee that the learned flow matches the governing equation; that is assessed by comparing Π_{model} with $\Pi^\star$ under the same normalization. This study is limited to scalar and (q, v) fields on periodic domains and four canonical equations with known dissipation. Extending the framework to vector and multi-field systems, non-periodic boundaries, and complex fluids—where the latent entropy density can represent a physical thermodynamic variable—is left for future work.

6 Conclusion

GENERIC-FNO (Sulskis & Ravi, 2026) enforced the two GENERIC degeneracies by projection, yielding a state-dependent reversible operator with unverified Poisson structure. Here, degeneracy is intrinsic: entropy is the integral of a latent density and a Casimir of a Poisson operator acting only on the mechanical field. Energy combines physical quadratic and learned pointwise terms; state dependence enters through the friction operator, whose parity structure contains the exact heat and wave operators but excludes transport, while a compatible Lie–Poisson pencil represents nonlinear transport. Because the constant part of L is fixed and Q quadratic, the linear flow is exactly integrable. For arbitrary parameters, skewness, positivity, both degeneracies, and the Jacobi identity (on the resolved band for the Lie–Poisson term) hold; under Theorem 2's curvature hypothesis, certified by the trained symbols, physical energy is bounded; and a discrete-gradient step preserves both laws in discrete time. Across four equations, two dimensions, and three backbones, the model beats unconstrained baselines with the same backbone in 61/72 seed-level comparisons and 20/24 cells, recovers the exact transport and wave symbols and Burgers coefficient, reproduces physical heat and Burgers dissipation rates, and remains on the physical energy for $10\times$ the training horizon; the only systematic loss is heat, where the unconstrained operator is as good or better. Ablations expose the predicted tradeoffs between physical and learned energy and between exact Jacobi structure and nonlinear transport, while operator-level diagnostics reveal errors missed by accuracy metrics, including gradient normalization, transport splitting between L and M, the quadratic gauge, and an ablation trained with the wrong integrator. Thus, structure-preserving models should pair conventional accuracy metrics with direct comparison of learned and exact operator symbols.

Broader Impact

The model enforces exact conservation of its own energy and exact production of its own entropy. This is a structural property of the parameterization and not a certificate of physical correctness: applied to a forced, open or non-conservative system, which lies outside GENERIC, it will still report machine-precision residuals, and in this paper the learned-energy ablation conserves its energy to 10^{-8} while the physical energy of a wave grows six-fold within ten frames and thirty-fold within forty. Exact structure should not be read as validation, in particular in safety-critical simulation; the calibration against ground-truth dissipation is the check that matters, and it requires a target system that is in class. The intended benefit is a surrogate that stays on the physical manifold far from its training data, at a training cost we report to be 25–$40\times$ that of an unconstrained operator.

AI USE STATEMENT

In this work, we used generative AI tools to assist with verifying mathematical proofs and derivations, identifying and reviewing relevant literature, cross-checking citations, and comparing technical claims and research ideas against the existing literature. Generative AI was also used to refine and condense portions of the manuscript, particularly to improve clarity and meet the nine-page limit without changing the intended technical content.

We additionally used Claude Code to assist with implementing the Python experimental framework from author-specified mathematical formulations, architectural designs, algorithms, and experimental requirements. AI-generated and AI-assisted code was reviewed, tested, and validated against the intended formulations and expected numerical behavior. AI-assisted literature and citation suggestions were independently checked against the original sources, and AI-assisted mathematical reasoning was reviewed for correctness. The authors determined the research questions, methodology, model architecture, experimental design, initial skeleton code, proof sketches, the interpretation of the results, and the scope/limitations and conclusions.

We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every number in the tables is read from one fresh run of a single self-contained notebook (PyTorch only) that generates the data, trains all four models on all twenty-four cells for three seeds, evaluates the structural identities in double precision, runs the representability, super-resolution and integrator checks, and writes a single review file from which the tables were built. The learned λ, the energy rollout of Figure 2, the retraining and discrete-gradient checks of §4, and the training-time ratios come from separate single-cell notebooks and an earlier run of the same grid. All notebooks, the review file and the per-seed results are in the supplementary material `https://anonymous.4open.science/r/Poisson-GENERIC-Neural-Operators-F4C4`. Appendix B gives the architectures, data generation and training protocol, Appendix A the full construction, and Appendix C the proofs. Two implementation details that a reimplementation is likely to get wrong are called out in the text: the grid-measure inner product with Riesz-scaled gradients (Appendix A) and the separate learning rate for the operator symbols (Appendix B).

REFERENCES

Benedikt Alkin, Andreas Fürst, Simon Schmid, Lukas Gruber, Markus Holzleitner, and Johannes Brandstetter. Universal physics transformers: A framework for efficiently scaling neural operators. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Ali Baheri and Lars Lindemann. Metriplectic conditional flow matching for dissipative dynamics. In *NeurIPS Workshop on Dynamics at the Frontiers of Optimization, Sampling, and Games (DynaFront)*, 2025. arXiv:2509.19526.

Qianying Cao, Somdatta Goswami, and George Em Karniadakis. Laplace neural operator for solving differential equations. *Nature Machine Intelligence*, 6:631–640, 2024. doi: 10.1038/s42256-024-00844-4.

Elena Celledoni, Volker Grimm, Robert I. McLachlan, David I. McLaren, Dion O'Neale, Brynjulf Owren, and G. Reinout W. Quispel. Preserving energy resp. dissipation in numerical PDEs using the "average vector field" method. *Journal of Computational Physics*, 231(20):6770–6789, 2012. doi: 10.1016/j.jcp.2012.06.022.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

Sybren R. de Groot and Peter Mazur. *Non-Equilibrium Thermodynamics*. North-Holland, Amsterdam, 1962.

Oscar Gonzalez. Time integration and discrete Hamiltonian systems. *Journal of Nonlinear Science*, 6(5):449–467, 1996. doi: 10.1007/BF02440162.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Miroslav Grmela and Hans Christian Öttinger. Dynamics and thermodynamics of complex fluids. I. Development of a general formalism. *Physical Review E*, 56(6):6620–6632, 1997. doi: 10.1103/PhysRevE.56.6620.

Anthony Gruber, Kookjin Lee, and Nathaniel Trask. Reversible and irreversible bracket-based dynamics for deep graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Anthony Gruber, Kookjin Lee, Haksoo Lim, Noseong Park, and Nathaniel Trask. Efficiently parameterized neural metriplectic systems. In *International Conference on Learning Representations (ICLR)*, 2025.

Jacob Helwig, Xuan Zhang, Cong Fu, Jerry Kurtin, Stephan Wojtowytsch, and Shuiwang Ji. Group equivariant Fourier neural operators for partial differential equations. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12907–12930, 2023.

Quercus Hernández, Alberto Badías, David González, Francisco Chinesta, and Elías Cueto. Structure-preserving neural networks. *Journal of Computational Physics*, 426:109950, 2021. doi: 10.1016/j.jcp.2020.109950.

Marlis Hochbruck and Alexander Ostermann. Exponential integrators. *Acta Numerica*, 19:209–286, 2010. doi: 10.1017/S0962492910000048.

Pengzhan Jin, Zhen Zhang, Aiqing Zhu, Yifa Tang, and George Em Karniadakis. SympNets: Intrinsic structure-preserving symplectic networks for identifying Hamiltonian systems. *Neural Networks*, 132:166–179, 2020. doi: 10.1016/j.neunet.2020.08.017.

Pengzhan Jin, Zhen Zhang, Ioannis G. Kevrekidis, and George Em Karniadakis. Learning Poisson systems and trajectories of autonomous systems via Poisson neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. doi: 10.1109/TNNLS.2022.3148734.

Allan N. Kaufman. Dissipative Hamiltonian systems: A unifying principle. *Physics Letters A*, 100 (8):419–422, 1984.

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.

Kookjin Lee, Nathaniel Trask, and Panos Stinis. Machine learning structure preserving brackets for forecasting irreversible processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

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

Ning Liu, Yiming Fan, Xianyi Zeng, Milan Klöwer, Lu Zhang, and Yue Yu. Harnessing the power of neural operators with automatically encoded conservation laws. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30965–30997, 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(3):218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Huakun Luo, Haixu Wu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Transolver++: An accurate neural solver for PDEs on million-scale geometries. In *International Conference on Machine Learning (ICML)*, 2025.

Franco Magri. A simple model of the integrable Hamiltonian equation. *Journal of Mathematical Physics*, 19(5):1156–1162, 1978. doi: 10.1063/1.523777.

Jerrold E. Marsden and Tudor S. Ratiu. *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*, volume 17 of *Texts in Applied Mathematics*. Springer, 2 edition, 1999. doi: 10.1007/978-0-387-21792-5.

Robert I. McLachlan, G. Reinout W. Quispel, and Nicolas Robidoux. Geometric integration using discrete gradients. *Philosophical Transactions of the Royal Society A*, 357(1754):1021–1045, 1999. doi: 10.1098/rsta.1999.0363.

Philip J. Morrison. A paradigm for joined Hamiltonian and dissipative systems. *Physica D: Nonlinear Phenomena*, 18(1–3):410–419, 1986. doi: 10.1016/0167-2789(86)90209-5.

Philip J. Morrison. Hamiltonian description of the ideal fluid. *Reviews of Modern Physics*, 70(2): 467–521, 1998. doi: 10.1103/RevModPhys.70.467.

Peter J. Olver. *Applications of Lie Groups to Differential Equations*, volume 107 of *Graduate Texts in Mathematics*. Springer, 2 edition, 1993.

Lars Onsager. Reciprocal relations in irreversible processes. I. *Physical Review*, 37(4):405–426, 1931. doi: 10.1103/PhysRev.37.405.

Dan Oprisa and Peter Toth. Metriplector: From field theory to neural architecture, 2026. arXiv:2603.29496.

Hans Christian Öttinger. *Beyond Equilibrium Thermodynamics*. Wiley-Interscience, Hoboken, NJ, 2005. doi: 10.1002/0471727903.

Hans Christian Öttinger. GENERIC integrators: Structure preserving time integration for thermodynamic systems. *Journal of Non-Equilibrium Thermodynamics*, 43(2):89–100, 2018. doi: 10.1515/jnet-2017-0034.

Maziar Raissi, Paris Perdikaris, and George Em 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.

Bogdan Raonić, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of PDEs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Ignacio Romero. Thermodynamically consistent time-stepping algorithms for non-linear thermomechanical systems. *International Journal for Numerical Methods in Engineering*, 79(6):706–732, 2009. doi: 10.1002/nme.2588.

Gilbert Strang. On the construction and comparison of difference schemes. *SIAM Journal on Numerical Analysis*, 5(3):506–517, 1968. doi: 10.1137/0705041.

Jason Sulskis and Sathya Ravi. GENERIC-FNO: Embedding energy conservation and entropy production into Fourier neural operators. *arXiv preprint arXiv:2606.08343*, 2026.

Alicia Tierz, Icíar Alfaro, David González, Francisco Chinesta, and Elías Cueto. Graph neural networks informed locally by thermodynamics. *Engineering Applications of Artificial Intelligence*, 144:110032, 2025.

Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023. doi: 10.1016/j.cma.2022.115783.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In *International Conference on Machine Learning (ICML)*, 2024.

Haijun Yu, Xinyuan Tian, Weinan E, and Qianxiao Li. OnsagerNet: Learning stable and interpretable dynamics using a generalized Onsager principle. *Physical Review Fluids*, 6(11):114402, 2021. doi: 10.1103/PhysRevFluids.6.114402.

Zhen Zhang, Yeonjong Shin, and George Em Karniadakis. GFINNs: GENERIC formalism informed neural networks for deterministic and stochastic dynamical systems. *Philosophical Transactions of the Royal Society A*, 380(2229):20210207, 2022. doi: 10.1098/rsta.2021.0207.

A BUILDING ON GENERIC-FNO: THE FULL CONSTRUCTION

GENERIC-FNO (Sulskis & Ravi, 2026) enforced the degeneracy conditions in equation 1 by placing constant Fourier multipliers between rank-one projections,

$$L = (I - P_S)\, D_L\, (I - P_S), \qquad M = (I - P_E)\, D_M\, (I - P_E),$$

where E and S were learned neural operators. We remove the reversible sandwich. With it, L depends on $\delta S/\delta z$; a state-dependent skew operator is Poisson only if it also satisfies the Jacobi identity, and this one does not (Table 1).

The reconstruction follows from a single design choice: augment the state with a latent entropy density and set $S = \int_\Omega s\, dx$. Because L acts only on the mechanical field, S is a Casimir and $L\,\delta S/\delta z = 0$ holds identically. The remaining components follow. A freely learned energy would let the latent entropy act as an energy source, so the energy is the physical one (§A.7). Since $\delta S/\delta z = (0, 1)$, all dissipation of the mechanical field passes through its coupling to s, so the friction operator must carry the state dependence that physical friction operators carry (§A.3). And the reversible operator must be Poisson on its own: a constant multiplier is, and a compatible Lie–Poisson term is on the resolved band (§A.2).

Throughout, on a periodic domain $\Omega \subset \mathbb{R}^d$ we use the grid-measure inner product $\langle f, g\rangle = |\Omega|^{-1}\int_\Omega f \cdot g\, dx$; on an N-point grid this is a sum over channels and points divided by N, and the variational derivative of a mean-pooled functional is N times its automatic-differentiation gradient.[1]

A.1 STATE, ENERGY, ENTROPY

The state is $z = (u, s)$, where $u : \Omega \to \mathbb{R}^p$ is the observed mechanical field ($p = 1$ for heat, advection and Burgers; $u = (q, v)$ with $p = 2$ for the damped wave) and $s : \Omega \to \mathbb{R}$ is an unobserved entropy density, trained only through its effect on u. At the start of each rollout an encoder from the same backbone family initializes it, $s_0 = \psi_\eta(u_0)$, so a rollout can begin at any observed frame.

The energy is

$$E_\theta[z] = Q(u) + \frac{1}{|\Omega|}\int_\Omega \varphi_\theta(u)\, dx + \frac{1}{|\Omega|}\int_\Omega e_\theta(s)\, dx, \qquad Q(u) = \begin{cases} \frac{1}{2}\|u\|^2 & \text{(scalar fields)},\\ \frac{1}{2}\left(\|v\|^2 + c^2\|\nabla q\|^2\right) & \text{(wave)}, \end{cases} \tag{4}$$

where Q is the fixed quadratic functional used as a diagnostic in GENERIC-FNO: the mechanical energy for the wave, and for the scalar equations the quadratic Lyapunov functional that advection conserves and heat and Burgers dissipate. The remaining terms are pointwise learned functions. The internal energy is

$$e_\theta(s) = T_0 s + \sum_i w_i\, \mathrm{softplus}(a_i s + b_i), \qquad w_i, a_i > 0, \qquad T(s) := e_\theta'(s) \geq T_0 > 0. \tag{5}$$

[1] This normalization is not cosmetic. With a per-channel mean, the Riesz-scaled gradients of Q on the p mechanical channels and of E on the $p+1$ channels of z differ by a factor $(p+1)/p$, the exact linear step then integrates only part of the linear flow, and the learned reversible symbols settle at an integer fraction of their exact values—one half for scalar fields and one third for the wave—for every seed and backbone. We found this by comparing learned and exact symbols, which is why that comparison is one of the diagnostics.

The temperature floor is needed because the friction operators of §A.3 contain a factor $1/T$; without it, negative drift in s drives T toward zero and the explicit step with it. The learned potential is

$$\varphi_\theta = \psi_\theta - \Pi_{\leq 2}\psi_\theta, \qquad \psi_\theta = \mathrm{softplus}(\mathrm{MLP}), \tag{6}$$

where $\Pi_{\leq 2}$ is the weighted least-squares projection onto polynomials of degree at most two in the p components of u, with respect to a fixed Gaussian measure over the data range. This removes a quadratic gauge: a term proportional to $\frac{1}{2}u^2$ in φ_θ is indistinguishable from a rescaling of Q, and in practice the network learns one, so that L represents only part of the linear dynamics and the explicit step carries the rest. Subtracting the Taylor expansion at $u = 0$ is insufficient, because the learned quadratic need not be centered there. With equation 6 the linear dynamics stay in L, where they are integrated exactly. Write $\Pi_{\leq 2}\psi_\theta(u) = u^\top A u + b\cdot u + c$ and $a := \lambda_{\max}(A)$. The resulting φ_θ need not be nonnegative, but since $\psi_\theta \geq 0$ it satisfies $\varphi_\theta \geq -\Pi_{\leq 2}\psi_\theta$, which is bounded below by a quadratic of curvature $-a$; the mechanical energy is therefore bounded whenever $a < \frac{1}{2}$ (Theorem 2), and a is measurable from the learned symbols (Remark 2).

The entropy is fixed,

$$S[z] = \frac{1}{|\Omega|}\int_\Omega s\,dx, \qquad \frac{\delta S}{\delta z} = (0, 1), \tag{7}$$

the entropy of every GENERIC field theory (Öttinger, 2005). Since L acts only on u, the reversible degeneracy $L\,\delta S/\delta z = 0$ holds for every state and every admissible L.

A.2 The Poisson operator

For the scalar equations,

$$L = \begin{pmatrix} L_u & 0 \\ 0 & 0 \end{pmatrix}, \qquad L_u = \alpha D + \lambda\,\mathcal{P}\,(uD + Du)\,\mathcal{P}, \qquad D = \sum_{a=1}^{d}\partial_a, \tag{8}$$

with learnable scalars α and λ. The constant term αD has symbol $i\alpha k_D$, $k_D = \sum_a k_a$, on the band $1 \leq |k|_\infty \leq m_L$; it is skew-adjoint and Poisson because its coefficients are constant. The second term is the Lie–Poisson operator of vector fields along D. It makes Burgers Hamiltonian with respect to the quadratic energy,

$$-\partial_x(u^2/2) = -\tfrac{1}{3}(u\partial_x + \partial_x u)\,\frac{\delta}{\delta u}\tfrac{1}{2}\|u\|^2,$$

and $\{D,\, uD + Du\}$ is a compatible pair of the KdV hierarchy, so L_u is Poisson for all (α, λ) (Magri, 1978; Olver, 1993). A generic learned multiplier $a(k)$ is not compatible with the Lie–Poisson term—the Jacobi residual of $a(k) + \lambda(uD + Du)$ is $\mathcal{O}(10^{-1})$—which is why we use the one-parameter pencil rather than the free multiplier of GENERIC-FNO. The operator $\mathcal{P}$ is the $2/3$-rule spectral projector. The sandwich preserves skewness on the grid (pointwise multiplication is symmetric, D is skew, $\mathcal{P}$ is symmetric) and removes the aliasing of quadratic products; it cannot remove the aliasing of the triple products that enter the Jacobi identity. The discrete Lie–Poisson operator therefore satisfies Jacobi to machine precision on band-limited states and functionals and has a residual of 0.2–0.8 on generic full-spectrum states (Table 1). We claim Poisson structure on the resolved band accordingly, and keep the $\lambda \equiv 0$ variant, Poisson for any state and unable to represent Burgers, as an ablation. The compatible dispersive term D^3 is omitted: its k^3 growth destabilizes both integrators of §A.5 at high wavenumbers, and none of the test equations needs it.

For the wave, $u = (q, v)$ and

$$L_u = J\ \mathrm{diag}(D_d, D_d), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix},$$

with D_d an even, nonnegative learned multiplier $d(k) \geq 0$ on the band. It is Poisson because its coefficients are constant, and skew because D_d is symmetric and commutes with J. The exact damped wave has $d \equiv 1$ up to the time scale, and the learned multiplier matches it to within 0.5% (§4). In both cases $\ker L$ contains the fixed subspace of s-directions, whatever the state.

A.3 THE FRICTION OPERATOR

Since $\delta S/\delta z = (0, 1)$, the irreversible update of u is the u-row of M applied to the entropy direction, and in the physics that row is state dependent. For heat conduction in these variables, $E = \frac{1}{2}\|u\|^2 + \int e(s)$, $S = \int s$, $T = e'(s)$, the positive semi-definite operator

$$M = CC^\top, \qquad C^\top(f, g) = \sqrt{\nu}\left(\sqrt{T}\,\nabla f - \frac{\nabla u}{\sqrt{T}}\,g\right) \tag{9}$$

satisfies $[M\,\delta S/\delta z]_u = \nu\Delta u$, $[M\,\delta S/\delta z]_s = \nu|\nabla u|^2/T$ (viscous heating), and $M\,\delta E/\delta z = 0$ identically. For the damped wave the analogue has no derivatives, $C^\top(f_v, g) = \sqrt{\gamma}\,(\sqrt{T}f_v - v\,g/\sqrt{T})$, giving $-\gamma v$ on v and $\gamma v^2/T$ on s. Three things are visible in equation 9: the operator depends on the state and on $\delta E/\delta z$; the derivative sits on the input side of the u channel and the divergence on the output side, with a pointwise matrix between them; and the row of $C^\top$ is orthogonal to $(\nabla u, T)$ pointwise, which is how degeneracy holds without a global projection. Our parameterization contains both operators exactly on the band $\nu k^2\Delta t < 1$, which contains the data band (Proposition 6, §4).

Slots and symbols. Let $c = p + 1$ be the number of channels of z and introduce $r = c(d + 1)$ slots: d flux slots (i, a) and one local slot i per channel. The operator $D_M : \mathbb{R}^r \to \mathbb{R}^c$ has symbols

$$\widehat{D_M}_{\,i,(j,a)}(k) = \frac{-ik_a b_{ija}(k)}{\sqrt{1 + k_a^2 b_{ija}(k)^2}}, \qquad \widehat{D_M}_{\,i,j}(k) = \frac{b'_{ij}(k)}{\sqrt{1 + b'_{ij}(k)^2}}, \tag{10}$$

where b and b' are learned on the band $|k|_\infty \le m_M$ and symmetrized to be even. The saturation bounds every symbol by one, which limits the gain of M under explicit integration for any parameters; it does not restrict the physical operators on the data band, where $\nu k^2\Delta t \lesssim 0.3$ (0.32 at $k = 8$). Two masks encode what the equations conserve: for the scalar conservation laws no local slot feeds u, so $\int u$ is conserved and there is no Rayleigh-type damping of a density; for the wave no slot feeds q, since friction acts on momentum and entropy only.

Pointwise matrix field with Onsager parity. $\Gamma_\vartheta(z) : \Omega \to \mathbb{R}^{r\times r}$ is a pointwise matrix field. Its flux–flux and local–local blocks are I plus bounded functions of $z(x)$; its flux–local blocks are linear in the gradient features,

$$\Gamma_{(i,a),\,j}(x) = \sum_{l,b} h_{(i,a),j,(l,b)}\big(z(x)\big)\,\partial_b z_l(x), \tag{11}$$

and the local–flux blocks likewise. The coefficient functions come from one bounded, zero-initialized pointwise MLP, so $\Gamma_\vartheta = I$ at initialization and $\|\Gamma_\vartheta - I\| \le 1$ throughout. Every path from a channel back to a channel then has an even number of derivatives—one through a flux slot of D_M, one through a flux–local entry of equation 11, none through local slots—so M can diffuse and damp but cannot transport (Proposition 5). Without the restriction the friction operator can carry part of the transport and take it from L; in a preliminary version the learned advection symbol converged to exactly half its true value for every seed and backbone, with M providing the other half. With it, the reversible operator must carry the transport, and the learned symbols converge to the exact ones (§4).

Pointwise degeneracy. Let $w(x) = (D_M^{\mathsf{H}}\,\delta E/\delta z)(x) \in \mathbb{R}^r$ be the flux of the energy gradient at x and $P(x) = I - \hat{w}\hat{w}^\top$ the projector off it in slot space. Then

$$A = D_M\,P\,\Gamma_\vartheta(z), \qquad M = AA^\top \tag{12}$$

is symmetric positive semi-definite, and $A^\top\delta E/\delta z = \Gamma_\vartheta^\top P w = 0$, so $M\,\delta E/\delta z = 0$ exactly, pointwise, with no global projection. The heat-conduction row of equation 9 is of this form, since $(\sqrt{T}, -\nabla u/\sqrt{T})$ is orthogonal to $(\nabla u, T)$ at each point. The global projection of GENERIC-FNO is unnecessary here, and with $S = \int s$ it is harmful when T is nearly uniform, because $\delta S/\delta z$ is then nearly parallel to the entropy component of $\delta E/\delta z$ and the projection removes most of the dissipative direction.

A.4 Dynamics and what is exact

The model evolves according to

$$\partial_t z = L\,\frac{\delta E_\theta}{\delta z} + M(z)\,\frac{\delta S}{\delta z}, \tag{13}$$

so $\partial_t u = L_u\big(u + \varphi'_\theta(u)\big) + [M(0,1)]_u$ and $\partial_t s = [M(0,1)]_s$: the reversible channel moves only u, and everything that enters s is dissipative. For all parameters, L is Poisson (on the resolved band when $\lambda \neq 0$), M is symmetric positive semi-definite, and $L\,\delta S/\delta z = 0$, $M\,\delta E_\theta/\delta z = 0$; hence $\dot{E}_\theta = 0$ and $\dot{S} = \langle (0,1), M(0,1)\rangle \geq 0$ (Theorem 1). Since E_θ is conserved and the potential is bounded below by a quadratic of curvature $-a$ (§A.1), the mechanical energy satisfies, whenever $a < \frac{1}{2}$,

$$Q(u_t) \;\leq\; \frac{E_\theta(z_0) - T_0 S(z_0) + C_\theta}{1-2a}, \qquad C_\theta = c + \frac{|b|^2}{2(1-2a)}, \tag{14}$$

where $T_0 S(z_t) \geq T_0 S(z_0)$ is the second law (Theorem 2). A freely learned energy provides no such bound. The guarantees concern E_θ, S, L and M; agreement with the governing equations is assessed separately by the accuracy results and the dissipation calibration.

A.5 Time integration

Exact linear flow and split integration. The constant part of L_u with the quadratic Q generates a unitary linear flow $u_t = L_u^{\text{const}}\,\delta Q/\delta u$: multiplication by $e^{i\alpha k_D \Delta t}$ per mode for scalar fields, a 2×2 rotation with frequency $\omega = d(k)\,c|k|$ for the wave. We apply it exactly in a Strang split (Strang, 1968): an exact half-step, an explicit step of the remainder—the Lie–Poisson term, the learned potential and the friction—sub-cycled $n_{\text{sub}} = 8$ times, and an exact half-step. The linear substep conserves Q to machine precision; the bounded symbols of equation 10 together with the sub-cycling keep the explicit remainder stable. This is the integrator used for training and evaluation. It may also be why the unconstrained FNO won on advection in GENERIC-FNO: the exact phase shift is a map the FNO has to learn, whereas here it is applied. The split conserves E_θ only to $\mathcal{O}(\Delta t^2)$ when $\varphi_\theta \neq 0$; exact discrete conservation is the property of the next step.

Discrete-gradient integration. Let $\bar{\nabla} F(z, z')$ be a discrete gradient, $\langle \bar{\nabla} F(z,z'), z' - z\rangle = F[z'] - F[z]$ (Gonzalez, 1996; McLachlan et al., 1999). We use the Gonzalez form for E_θ; since S is linear, $\bar{\nabla} S = (0,1)$ exactly. The implicit update

$$z_{t+1} = z_t + \Delta t\big[L(\bar{z})\,\bar{\nabla} E(z_t, z_{t+1}) + \bar{M}\,\bar{\nabla} S\big], \qquad \bar{z} = \tfrac{1}{2}(z_t + z_{t+1}), \tag{15}$$

with $\bar{M}$ built from $\bar{\nabla} E$ in place of $\delta E/\delta z$, conserves E_θ and realizes the entropy production law exactly for any Δt (Theorem 3). The proof uses $\bar{\nabla} S \in \ker L$, which holds because $\ker L$ contains every s-direction at every state, including with the Lie–Poisson term evaluated at $\bar{z}$; the scheme is the discrete analogue of the thermodynamically consistent integrators of Romero (2009) and Öttinger (2018), and it is unavailable to the projected operator of GENERIC-FNO, whose kernel moves with the state. We solve equation 15 by relaxed fixed-point iteration from the explicit predictor. It converges on every case we report: the random-initialization checks of Table 1 and the trained 1D and 2D heat FNO models of §4. On the trained 2D heat Transolver models it diverges at $\Delta t = 0.1$ from either the Euler or the split predictor. The equation is the same; what differs is the map $z_{t+1} \mapsto M(\bar{z})$ inside the iteration: the Transolver's coefficient field Γ_ϑ is a global-attention function of the whole field, so its Lipschitz constant, and with it that of the fixed-point map, is far larger than for the FNO's coefficients, and the relaxed iteration stops being a contraction. A Newton or sub-stepped solve would remove the restriction; we leave it to future work. In the experiments the discrete-gradient step is the reference against which the split step's energy error is measured; training with it is exact but several times more expensive.

A.6 Diagnostics

All diagnostics reference no learned functional. The ground-truth rate $\Pi^\star$ is the fractional one-step decay of Q along the reference trajectory. The reversibility rate is

$$r_{\text{mech}} = -\frac{\langle \nabla Q(\bar{u}), \Delta u\rangle}{\|\nabla Q(\bar{u})\|\,\|\Delta u\|}, \qquad \bar{u} = \tfrac{1}{2}(u + u'), \tag{16}$$

with the midpoint gradient, for which the numerator is exactly $Q(u) - Q(u')$ when Q is quadratic, so a Q-conserving step scores zero; the endpoint gradient used in GENERIC-FNO carries an $\mathcal{O}(\|\Delta u\|/\|u\|)$ bias, about 0.1 at our frame spacing, even for the exact advection solution. Because r_{mech} is a cosine it under-weights dissipation on transport-dominated flows, so magnitudes are reported through

$$\Pi_{\mathrm{model}} = \frac{Q(u) - Q(u')}{Q(u)}, \tag{17}$$

the normalization of $\Pi^\star$; a calibrated model has $\Pi_{\mathrm{model}}/\Pi^\star = 1$. Three diagnostics are new.

- *Jacobi residual.* For random smooth test functionals F, G, H with nontrivial second variations, $\mathcal{J} = |\{\{F, G\}, H\} + \mathrm{cyc}|/\sum_{\mathrm{cyc}} |\{\{F, G\}, H\}|$, computed by automatic differentiation on generic full-spectrum states and on band-limited states with band-limited functionals, for our L, the constant-L ablation, and the projected operator of GENERIC-FNO built from an unrestricted learned entropy of the same size in both columns of Table 1 (with a Casimir entropy the projection is trivial and the comparison empty).
- *Representability.* We set D_M and Γ to the closed-form operators of equation 9 and its wave analogue inside the model and compare $M\,\delta S/\delta z$ and $M\,\delta E/\delta z$ with the spectral truth on band-limited states.
- *Learned versus exact symbols.* For advection and the wave the exact reversible symbols are $-c\,k\,\Delta t$ and $d \equiv \Delta t$, and for Burgers the exact pencil coefficient is $\lambda = -\Delta t/3$; we compare them with the learned ones. This diagnostic found both the gradient-scaling error of the footnote above and the transport split of §A.3, and it certifies the curvature hypothesis of Theorem 2 (Remark 2).

A.7 Ablation variants

Two variants isolate the two design decisions. *Learned* E, S is the closest analogue of GENERIC-FNO under a constant L: E_θ and S_ϕ are both mean-pooled neural operators, S_ϕ sees u only through the modes L does not touch (so it is a Casimir), and M is a constant matrix-symbol multiplier with the global sandwich. It has exact Jacobi and exact degeneracy and nothing that ties E_θ to Q. *Constant* L is the model of record with $\lambda \equiv 0$: Poisson for any state and unable to represent nonlinear transport. Together they bracket the model of record between exact structure without physics and exact Jacobi without Burgers.

A.8 Learned Poisson structures by pullback

The constant L is a Darboux normal form, and any change of coordinates φ gives another Poisson operator $L_\varphi = (D\varphi)^{-1} L\,(D\varphi)^{-\mathsf{T}}$ whose Casimirs are $C \circ \varphi$ (Proposition 7); this is the function-space analogue of Poisson neural networks (Jin et al., 2022), and §A.2 is the case $\varphi = \mathrm{id}$. The Lie–Poisson term is a different route to state dependence, through a compatible pencil rather than a coordinate change. An invertible, resolution-independent φ and the cost of its Jacobian products are what a pullback model would need, and we leave it to future work.

B Architecture, data and training protocol

Data. The equations are, on the periodic box $[0, 2\pi)^d$ with $D = \sum_a \partial_a$: heat $\partial_t u = \nu\Delta u$; advection $\partial_t u = -c\,Du$ (transport along $(1, \ldots, 1)$ in 2D); viscous Burgers $\partial_t u = -u\,Du + \nu\Delta u$; damped wave $\partial_t q = v$, $\partial_t v = c^2\Delta q - \gamma v$. Initial conditions are band-limited random fields ($|k|_\infty \leq 8$ in 1D, ≤ 6 in 2D) with unit RMS, integrated by a de-aliased spectral RK4 solver with resolution-dependent sub-stepping; frames are $\Delta t = 0.1$ apart. Grids are 64 points in 1D and 32^2 in 2D, with 120 (1D) and 80 (2D) trajectories of 20 (15) frames, split 80/20. For the damped wave both q and v are observed.

Backbones. One backbone family is used throughout a model: an FNO (width 24, 12 modes in 1D and 8 in 2D, three layers; InstanceNorm on in 2D and off in 1D, following the diagnosis of Sulskis & Ravi (2026)), a Transolver (width 32, three physics-attention blocks, four heads, 32

slices, with a kernel-3 circular depthwise convolution on the grid before each block), or a CNO-style periodic convolutional operator (two-level U-shape, 3-point circular kernels). The convolution in the Transolver is the structured-mesh form of its slicing: without it every projection is a linear map on flattened tokens, the only path between grid points is the 32 global slice tokens, and the residual baseline cannot represent a one-cell shift; its training loss then stays at the persistence level on advection, wave and Burgers, and what it learns is a pointwise damping. The backbone appears in the baseline as the residual map G, and in the PG models as the entropy encoder ψ_η and the coefficient network of Γ_ϑ; the pointwise networks φ_θ and e_θ are small MLPs. The operator symbols $\alpha, \lambda, d(k), b, b'$ are explicit parameters on the resolved band, with $m_L = 16$ (1D) and 10 (2D) for the reversible operator and $m_M = N_x/3$ for the friction operator. Parameter counts are in Table 2.

Training. All models minimize the relative L^2 error of the observed field over a short rollout from a random start ($H = 4$ frames in 1D, 3 in 2D) with AdamW (weight decay 10^{-4}), a one-cycle schedule, gradient clipping at one, batch size 16 (1D) and 8 (2D), for 60 epochs in 1D and 40 in 2D—about 350 optimizer steps for every model. Networks use learning rate 10^{-3}; the operator symbols use 3×10^{-2}. The larger rate is needed because a symbol is a single scalar with a small gradient—the exact advection symbol is -0.8 at $k = 8$ and cannot be reached from a 0.1 initialization at the network rate in the steps available—and it is safe because the friction symbols are saturated by construction. PG-phys trains and evaluates with the split integrator ($n_{\mathrm{sub}} = 8$); the learned-E, S ablation, whose energy is not quadratic and so has no exact linear substep, uses explicit Euler, whose energy error is $\mathcal{O}(\Delta t^2)$ per step under exact degeneracy and whose structural residuals ($r_E \approx 10^{-8}$) confirm that the integrator is not what separates it from PG-phys. Evaluation is a ten-frame rollout from $t = 0$ on held-out trajectories; every number is a mean over three seeds that resample data and initialization, and the full grid is one fresh run of the released notebook on a single GPU. The learned λ of §4 and the energy rollout of Figure 2 come from separate runs of single cells of the same notebook.

Training budget. FNO and CNO converge within the shared protocol; the Transolver baseline does not. At five times the epochs it reaches 0.0051 ± 0.0027 on 1D heat, 0.0070 ± 0.0019 on 1D advection, 0.069 ± 0.001 on 1D Burgers and 0.038 ± 0.010 on the 1D wave (three seeds), so the Transolver column of Table 2 compares equal budgets rather than converged models; against a fully trained Transolver, PG-phys would lose heat and keep advection, Burgers and the wave.

C Mathematical Statements and Proofs

Throughout, $H = L^2(\Omega; \mathbb{R}^{p+1})$ with the grid-measure inner product $\langle f, g\rangle = |\Omega|^{-1} \int_\Omega f \cdot g \, dx$, the state is $z = (u, s)$, brackets are $\{F, G\} = \langle \delta F/\delta z, L\, \delta G/\delta z\rangle$ and $[F, G] = \langle \delta F/\delta z, M\, \delta G/\delta z\rangle$, and every operator is asserted at a fixed but arbitrary state unless stated otherwise. On an N-point grid all statements hold verbatim with integrals replaced by grid sums; where the discrete case differs from the continuous one we say so.

C.1 The reversible operator

Proposition 1 (Constant multipliers are Poisson)**.** *Let L be a bounded skew-adjoint operator on H with constant coefficients. Then $\{\cdot,\cdot\}$ is bilinear, skew, satisfies the Leibniz rule, and satisfies the Jacobi identity $\{\{F, G\}, H\} + \{\{G, H\}, F\} + \{\{H, F\}, G\} = 0$ for all C^2 functionals. In particular the block αD of equation 8 on the band, and the wave block $J\,\mathrm{diag}(D_d, D_d)$ with D_d symmetric, are Poisson.*

Proof. Write $F_z = \delta F/\delta z$ and F_{zz} for the second variation. By the chain rule, $\{\{F, G\}, H\} = \langle F_{zz}\, LG_z, LH_z\rangle + \langle F_z, L\, G_{zz}\, LH_z\rangle$, the derivative of L itself being zero. In the cyclic sum each functional's second variation appears twice, once as $\langle F_{zz} LG_z, LH_z\rangle$ and once as $\langle F_{zz} LH_z, LG_z\rangle$ with the opposite sign of the skew operator between them; symmetry of F_{zz} and skewness of L make the two cancel. The remaining terms cancel the same way. This is the standard argument for constant-coefficient Hamiltonian operators (Olver, 1993, Cor. 7.5). Skewness of the two blocks: αD has symbol $i\alpha k_D$, purely imaginary and odd in k; $J\,\mathrm{diag}(D_d, D_d)$ satisfies $(JD)^\top = D^\top J^\top = -DJ = -JD$ because $D = \mathrm{diag}(D_d, D_d)$ is symmetric and commutes with J. □

Proposition 2 (The pencil). *Let $D = \sum_a \partial_a$ on the torus and $\mathcal{P}$ the $2/3$-rule projector. (i) $uD+Du$ is skew-adjoint, and so is $\mathcal{P}(uD + Du)\mathcal{P}$ on the grid. (ii) In the continuum, $\alpha D + \lambda(uD + Du)$ is a Poisson operator for all (α, λ). (iii) On the grid, the Jacobi identity for $\alpha D + \lambda\mathcal{P}(uD + Du)\mathcal{P}$ holds exactly on states and test functionals whose Fourier support is contained in the band on which cubic products are alias-free, and not in general otherwise.*

Proof. (i) $\langle f, uDg+D(ug)\rangle = \langle uf, Dg\rangle - \langle Df, ug\rangle = -\langle D(uf)+uDf, g\rangle$ by skewness of D and symmetry of pointwise multiplication; both properties survive discretization, and $\mathcal{P}$ is symmetric, so the sandwich is skew on the grid. (ii) D and $uD + Du$ are the Hamiltonian pair of the KdV hierarchy in the direction D; each is Hamiltonian and their sum is, which is the compatibility verified by Magri (1978), see also Olver (1993, §7.3); compatibility of two Hamiltonian operators is the Jacobi identity for every linear combination. (iii) For band-limited z and test functionals with band-limited gradients, every quantity in the cyclic sum is a polynomial of degree at most three in band-limited fields; if the grid resolves those products without aliasing, the discrete computation coincides with the continuum one and Jacobi holds by (ii). If the products alias, the discrete Leibniz rule $D(fg) = fDg + gDf$ fails on the aliased modes and the cancellation in (ii) does not go through; the residuals in Table 1 (0.17–0.78 on generic states, 10^{-16} on band-limited ones) are this effect. □

Remark 1. *A free learned multiplier $a(k)$ is not compatible with $uD + Du$: the Jacobi residual of $a(k) + \lambda(uD + Du)$ is $O(10^{-1})$ even on band-limited states. Only members of the pencil spanned by D and D^3 are, which is why the constant part of L is αD.*

C.2 Casimir entropy and the friction operator

Proposition 3 (The entropy is a Casimir). *For $S[z] = |\Omega|^{-1}\int_\Omega s\,dx$ and any L of the block form $\mathrm{diag}(L_u, 0)$, $L\,\delta S/\delta z = 0$ for every state.*

Proof. $\delta S/\delta z = (0, 1)$ and the s-block of L is zero. □

Proposition 4 (Symmetry, positivity, pointwise degeneracy). *Let $A = D_M\,P\,\Gamma_\vartheta(z)$ with D_M as in equation 10, Γ_ϑ any pointwise matrix field, and $P(x) = I - \hat{w}\hat{w}^\top$ with $w = D_M^{\mathsf{H}}\,\delta E/\delta z$. Then $M = AA^\top$ is self-adjoint and positive semi-definite on real fields, and $M\,\delta E/\delta z = 0$, for any parameters and any state.*

Proof. $AA^\top$ is self-adjoint and $\langle v, AA^\top v\rangle = \|A^\top v\|^2 \geq 0$ for any A; reality follows from the even, real symmetrization of the symbols in equation 10. For degeneracy, $A^\top\,\delta E/\delta z = \Gamma_\vartheta^\top P\,D_M^{\mathsf{H}}\,\delta E/\delta z = \Gamma_\vartheta^\top P w$, and $Pw = w - \hat{w}\langle\hat{w}, w\rangle_{\mathrm{slot}} = 0$ pointwise. Hence $M\,\delta E/\delta z = A(A^\top\delta E/\delta z) = 0$. □

Proposition 5 (Onsager parity: M cannot transport). *Let the flux slots of D_M carry one derivative and the local slots none, as in equation 10, and let the flux–local and local–flux blocks of Γ_ϑ be linear in first derivatives of z with pointwise coefficients, the flux–flux and local–local blocks pointwise, as in equation 11. Then, for the linearization of $M(z)\,\delta S/\delta z$ about any state, every term contributing to the mechanical update $[M(0, 1)]_u$ contains an even number of spatial derivatives. In particular no term of the form $c(x)\,\partial_a u$ (transport) can appear, whereas terms of the form $\partial_a\big(c(x)\,\partial_b u\big)$ (diffusion) and $c(x)\,u$ (damping) can.*

Proof. Follow one slot path from the entropy direction $(0, 1)$ to a channel of u through $AA^\top = D_M P\Gamma\Gamma^\top P D_M^{\mathsf{H}}$. The input $D_M^{\mathsf{H}}(0, 1)$ places one derivative on a flux slot and none on a local slot. The projector P mixes slots with coefficients $\hat{w}_i\hat{w}_j$; since w itself has one derivative on flux slots and none on local slots, a flux–local entry of P carries one derivative and the diagonal blocks carry zero or two. The blocks of $\Gamma\Gamma^\top$ carry, by hypothesis, an even number of derivatives on the diagonal blocks and an odd number on the off-diagonal ones. The output D_M again adds one derivative for a flux slot and none for a local slot. Counting, a path that enters and leaves through slots of the same type accumulates an even number of off-diagonal crossings, hence an even total; a path that enters through a flux slot and leaves through a local slot (or vice versa) crosses an odd number of times, each crossing contributing one derivative, and the two ends contribute one derivative in total, again even. Every path is therefore of even order. A first-order term is odd and cannot arise. □

Proposition 6 (Representability). *Within the parameterization equation 10–equation 12 there are parameter values for which, on any band with $\nu k^2 \Delta t < 1$, $M\,\delta S/\delta z$ equals, exactly, $(\nu\Delta u,\ \nu|\nabla u|^2/T)$ for heat conduction and $(0, -\gamma v,\ \gamma v^2/T)$ for the damped wave, with $M\,\delta E/\delta z = 0$ holding without the projector P.*

Proof. Heat, one channel u and one channel s: take the flux symbol $-ik_a\kappa$ on the u row of D_M with a constant κ, identity on the local slot of s, and Γ with $\Gamma_{(u,a),(u,a)} = \sqrt{\nu T}/\kappa$ and $\Gamma_{s,(u,a)} = -\sqrt{\nu}\,\partial_a u/(\kappa\sqrt{T})$, all other entries zero. Then $A^\top(f,g) = \Gamma^\top D_M^{\mathsf{H}}(f,g) = \sqrt{\nu}(\sqrt{T}\,\nabla f - g\,\nabla u/\sqrt{T})$, which is $C^\top$ of equation 9, and $AA^\top(0,1) = (\nabla\cdot(\nu\nabla u),\ \nu|\nabla u|^2/T)$ by direct expansion. Degeneracy: $A^\top(u,T) = \sqrt{\nu}(\sqrt{T}\nabla u - T\nabla u/\sqrt{T}) = 0$ pointwise, so P acts as the identity on the relevant directions and is not needed. Inverting the saturation in equation 10 gives $b_a(k)^2 = \kappa^2/(1-\kappa^2 k_a^2)$, finite exactly when $\kappa|k_a| < 1$; choosing $\kappa^2 = \nu\Delta t$ makes this the band $\nu\Delta t\,k^2 < 1$, which contains the data band ($\nu\Delta t\,k^2 \le 0.36$ there), and keeps Γ within its bound $\|\Gamma - I\| \le 1$ for $T \le 4/(\nu\Delta t\,k_{\max}^2) \approx 10$. Wave, channels (q,v,s): take D_M as the identity on the local slots of v and s (no flux slots used, q row zero), $\Gamma_{v,v} = \sqrt{\gamma T}$, $\Gamma_{s,v} = -\sqrt{\gamma}\,v/\sqrt{T}$; then $A^\top(f_q, f_v, g) = \sqrt{\gamma}(\sqrt{T}f_v - g\,v/\sqrt{T})$, $AA^\top(0,0,1) = (0, -\gamma v, \gamma v^2/T)$, and $A^\top(-c^2\Delta q, v, T) = 0$; no band condition arises because no flux slot is used. Both instantiations are checked numerically to 10^{-16} on band-limited states in §4. □

C.3 The laws

Theorem 1 (Metriplectic structure and the two laws). *For any parameters, L of equation 8 is skew-adjoint and Poisson (on every state if $\lambda = 0$; on the resolved band otherwise), M of equation 12 is self-adjoint and positive semi-definite, and $L\,\delta S/\delta z = 0$, $M\,\delta E_\theta/\delta z = 0$. Along $\partial_t z = L\,\delta E_\theta/\delta z + M\,\delta S/\delta z$,*

$$\frac{dE_\theta}{dt} = 0, \qquad \frac{dS}{dt} = \langle \delta S/\delta z, M\,\delta S/\delta z\rangle \ge 0,$$

and every Casimir of L—in particular S and, for the scalar equations, $\int u$—is invariant under the reversible flow $\partial_t z = L\,\delta E_\theta/\delta z$.

Proof. The structural statements are Propositions 1–4. Then $\dot E = \langle E_z, LE_z\rangle + \langle E_z, MS_z\rangle$; the first term is zero by skewness ($\langle v, Lv\rangle = -\langle Lv, v\rangle$), the second by $ME_z = 0$ and self-adjointness of M. Likewise $\dot S = \langle S_z, LE_z\rangle + \langle S_z, MS_z\rangle = -\langle LS_z, E_z\rangle + \langle S_z, MS_z\rangle = 0 + \|A^\top S_z\|^2$. For a Casimir C, $\dot C = \langle C_z, LE_z\rangle = -\langle LC_z, E_z\rangle = 0$ under the reversible flow; $\int u$ is a Casimir because L vanishes at $k = 0$. □

Theorem 2 (Bounded mechanical energy). *Write $\Pi_{\le 2}\psi_\theta(u) = u^\top A u + b\cdot u + c$ for the projected quadratic in equation 6, let $a := \lambda_{\max}(A)$, and let $e_\theta(s) = T_0 s + \tilde e_\theta(s)$ with $\tilde e_\theta \ge 0$. If $a < \frac{1}{2}$, then along the exact flow, and along the discrete-gradient step of Theorem 3, for all $t \ge 0$,*

$$Q(u_t) \ \le\ \frac{E_\theta(z_0) - T_0 S(z_0) + C_\theta}{1-2a}, \qquad C_\theta := c + \frac{|b|^2}{2(1-2a)}.$$

Proof. Since $\psi_\theta \ge 0$, $\varphi_\theta = \psi_\theta - \Pi_{\le 2}\psi_\theta \ge -\Pi_{\le 2}\psi_\theta$ pointwise, so, with the normalized measure, $Q(u) + \int\varphi_\theta(u) \ge \int\big(\frac{1}{2}|u|^2 - u^\top A u - b\cdot u - c\big) \ge (\frac{1}{2} - a)\|u\|^2 - |b|\,\|u\| - c \ge \frac{1-2a}{2}\,\|u\|^2 - C_\theta$ by completing the square. Hence $(1-2a)\,Q(u_t) \le E_\theta(z_t) - T_0 S(z_t) - \int\tilde e_\theta(s_t) + C_\theta \le E_\theta(z_0) - T_0 S(z_0) + C_\theta$, using $\tilde e_\theta \ge 0$, $E_\theta(z_t) = E_\theta(z_0)$ and $S(z_t) \ge S(z_0)$ (Theorem 1; Theorem 3 for the discrete step). (Under the split step the two facts hold only to the step's energy error; the ratio $\max_t Q$/bound of Table 3, evaluated with $a = 0$ and $C_\theta = -\inf\varphi_\theta$ over the data range, is measured under the split step and is below one on 70 of 72 seeds, the exception being the seed discussed in §4.) □

Remark 2. *The hypothesis is the quadratic gauge of §A.1: a term $a|u|^2$ surviving in φ_θ is indistinguishable from rescaling Q, and $a \ge \frac{1}{2}$ would flip the sign of the effective quadratic energy. It is also measurable. The exact linear step integrates the flow of Q with the learned symbol, while the explicit remainder carries $2a$ of the same flow, so a trained model matches the data only if the*

learned symbol equals $(1+2a)^{-1}$ *times the exact one. The symbol errors of §4 (*$\leq 0.07\%$ *on advection,* 0.5% *on the wave,* 0.6% *for the Burgers* λ*) therefore bound* $|a| \leq 3 \times 10^{-3}$ *on every reversible cell, three orders of magnitude inside the hypothesis. For heat, which has no reversible symbol to compare, the bound column of Table 3 uses* $\inf \varphi_\theta$ *over the data range and is loose (*≤ 0.36*).*

Lemma 1 (Discrete gradients)**.** *Let* $\bar{\nabla} E(z, z')$ *be the Gonzalez discrete gradient of* E_θ*,*

$$\bar{\nabla} E(z, z') = E_z(\bar{z}) + \frac{E_\theta[z'] - E_\theta[z] - \langle E_z(\bar{z}), z' - z\rangle}{\|z' - z\|^2}\,(z' - z), \qquad \bar{z} = \tfrac{1}{2}(z + z'),$$

and $\bar{\nabla} S = (0, 1)$*. Then* $\langle \bar{\nabla} E, z' - z\rangle = E_\theta[z'] - E_\theta[z]$*,* $\langle \bar{\nabla} S, z' - z\rangle = S[z'] - S[z]$*, and* $L(\bar{z})\,\bar{\nabla} S = 0$*.*

Proof. The first identity is immediate from the definition; the second because S is linear; the third because $\bar{\nabla} S = (0, 1)$ lies in the s-block, which every L of the form $\mathrm{diag}(L_u, 0)$ annihilates, whatever the state at which L_u is evaluated. □

Theorem 3 (Exact discrete first and second laws)**.** *Let* $z_{t+1} = z_t + \Delta t\big[L(\bar{z})\,\bar{\nabla} E(z_t, z_{t+1}) + \bar{M}\,\bar{\nabla} S\big]$ *with* $\bar{M} = \bar{A}\bar{A}^\top$*,* $\bar{A} = D_M \bar{P}\,\Gamma_\vartheta(\bar{z})$*, and* $\bar{P}$ *the projector off* $D_M^{\mathsf{H}} \bar{\nabla} E$*. Then, exactly and for any* Δt*,*

$$E_\theta[z_{t+1}] = E_\theta[z_t], \qquad S[z_{t+1}] - S[z_t] = \Delta t\,\|\bar{A}^\top \bar{\nabla} S\|^2 \;\geq 0.$$

Proof. By Lemma 1, $E_\theta[z_{t+1}] - E_\theta[z_t] = \langle \bar{\nabla} E, z_{t+1} - z_t\rangle = \Delta t \langle \bar{\nabla} E, L(\bar{z})\bar{\nabla} E\rangle + \Delta t \langle \bar{\nabla} E, \bar{M}\bar{\nabla} S\rangle$. The first term vanishes by skewness of $L(\bar{z})$ (Propositions 1, 2(i)); the second because $\bar{A}^\top \bar{\nabla} E = \Gamma^\top \bar{P}\, D_M^{\mathsf{H}} \bar{\nabla} E = 0$ by construction of $\bar{P}$, so $\bar{M}\bar{\nabla} E = 0$, and $\bar{M}$ is self-adjoint. Likewise $S[z_{t+1}] - S[z_t] = \Delta t \langle \bar{\nabla} S, L(\bar{z})\bar{\nabla} E\rangle + \Delta t \langle \bar{\nabla} S, \bar{M}\bar{\nabla} S\rangle$; the first term equals $-\Delta t \langle L(\bar{z})\bar{\nabla} S, \bar{\nabla} E\rangle = 0$ by Lemma 1, the second is $\Delta t \|\bar{A}^\top \bar{\nabla} S\|^2$. □

Remark 3. *The proof uses only that* $\ker L(z)$ *contains* $\bar{\nabla} S$ *for every* z*, which is the case because* S *is a Casimir of every* L *of block form. For the projected operator* $(I - P_S(z))D_L(I - P_S(z))$ *of GENERIC-FNO the kernel is spanned by* $\delta S/\delta z|_z$ *and moves with* z*; no single vector lies in* $\ker L(z_t)$*,* $\ker L(\bar{z})$ *and* $\ker L(z_{t+1})$ *at once, and the first-order term in the entropy balance does not cancel. This is the concrete payoff of degeneracy through a Casimir rather than through a projection. The theorem concerns the solution of equation 15; whether the fixed-point iteration of §A.5 reaches it is a separate question, answered there.*

C.4 Two further statements

Proposition 7 (Pullback structures)**.** *For any* C^2 *diffeomorphism* φ *of* H *and any Poisson operator* L*,* $L_\varphi = (D\varphi)^{-1} L\,(D\varphi)^{-\mathsf{T}}$ *is Poisson, and* C *is a Casimir of* L *if and only if* $C \circ \varphi$ *is a Casimir of* L_φ*.*

Proof. A Poisson operator is a contravariant 2-tensor and the Jacobi identity is coordinate-free, so the pullback of a Poisson tensor is Poisson (Marsden & Ratiu, 1999, Ch. 10). By the chain rule $\delta(C \circ \varphi)/\delta z = (D\varphi)^\top\, \delta C/\delta w|_{\varphi(z)}$, hence $L_\varphi\, \delta(C \circ \varphi)/\delta z = (D\varphi)^{-1} L\, \delta C/\delta w$, which vanishes iff $L\, \delta C/\delta w = 0$. □

Proposition 8 (Resolution and dimension independence)**.** *Let* $\iota_{N \to N'}$ *be band-limited prolongation (Fourier zero-padding) and* Φ_N *one split or discrete-gradient step of the model on an* N*-point grid. For states whose Fourier support lies within the bands of* L *and* M *and below the de-aliasing limit of the coarser grid,* $\Phi_{N'} \circ \iota_{N \to N'} = \iota_{N \to N'} \circ \Phi_N$*.*

Proof sketch. Every component of the step is per-mode (the multipliers of L and D_M, the exact unitary flow, the projector $\mathcal{P}$), per-point (the constitutive functions, Γ_ϑ, the slot projector P), or a mean-pooled functional whose Riesz-scaled gradient is grid-independent by the inner-product convention of Appendix A. Per-mode maps commute with zero-padding on band-limited inputs; per-point maps commute with band-limited interpolation exactly when their outputs are resolved, which the band hypothesis guarantees; the encoder and the coefficient network are neural operators that are discretization-convergent by construction for the FNO backbone, and are excluded from the claim for the convolutional backbone and for the Transolver, whose local-mixing kernels are defined in grid units. The measured invariance to four digits in §4 is the empirical content. □