GENERIC-FNO: Embedding Energy Conservation and Entropy Production into Fourier Neural Operators

Anonymous authors
Paper under double-blind review

Abstract

We introduce GENERIC-FNO, the first neural operator to embed the full GENERIC (metriplectic) structure of nonequilibrium thermodynamics—reversible, energy-conserving dynamics and irreversible, entropy-producing dynamics coupled through the degeneracy conditions—directly in function space. Existing structure-preserving neural operators enforce at most a single conservation law or a purely reversible (Hamiltonian) structure, while thermodynamically consistent learning has so far been confined to finite-dimensional, graph, or particle systems. GENERIC-FNO closes this gap: it learns the energy and entropy functionals as neural operators and parameterizes the Poisson and friction operators as diagonal Fourier multipliers sandwiched between rank-one projections that enforce the degeneracy conditions *exactly, by construction*—with no penalty term, no projection of the predicted update, and no free residual. The degeneracy identities therefore hold to machine precision (residuals $\sim 10^{-13}$) for any initialization, spatial dimension, or grid resolution, so the continuous-time dynamics conserve the learned energy and produce entropy exactly; the explicit time stepping we deploy adds only a small $\mathcal{O}(\Delta t^2)$ drift (per-step energy residual $\sim 10^{-6}$). We further observe that the (E, S, L, M) decomposition realizing a given flow is not unique, and introduce a gauge-invariant dissipation diagnostic that separates genuinely reversible from dissipative dynamics independently of the learned functionals. Across three operator backbones (1D and 2D Fourier neural operators and DeepONet) and four PDEs spanning reversible, dissipative, and mixed regimes, GENERIC-FNO preserves its exact structural guarantees zero-shot across a $4\times$ super-resolution range ($64 \to 256$), recovers the correct ground-truth ordering of physical dissipation, and remains competitive with strong unconstrained and energy-penalized baselines—outperforming them on several dissipative and mixed problems at comparable or fewer parameters.

1 Introduction

Neural operators such as the Fourier Neural Operator (FNO) (Li et al., 2021) and DeepONet (Lu et al., 2021) have become powerful surrogates for the solution operators of partial differential equations (PDEs), learning maps between function spaces that generalize across discretizations (Kovachki et al., 2023). Yet when such surrogates are rolled out autoregressively to predict long trajectories, they accumulate errors that are not merely numerical but *physical*: energy drifts, and quantities that should decay monotonically can spuriously increase. An unconstrained operator has no notion of the conservation and dissipation laws that govern the system it imitates, so nothing prevents it from leaving the physically admissible manifold.

A growing body of work injects physical structure as an inductive bias. In function space, however, this structure has so far been limited to a single conserved quantity or to purely *reversible*, Hamiltonian dynamics—energy is conserved, but there is no mechanism for the irreversible entropy production that defines every real dissipative process. Conversely, the machine-learning literature on full *thermodynamic* consistency—the GENERIC, or metriplectic, formalism (Grmela & Öttinger, 1997; Öttinger, 2005) that unifies reversible and irreversible dynamics—has been developed almost exclusively for finite-dimensional state vectors, graphs, or particle systems (Lee et al., 2021; Hernández et al., 2021; Zhang et al., 2022; Gruber et al., 2023). To our

knowledge, no neural operator embeds the full GENERIC structure in function space. This is the gap we close.

GENERIC writes the dynamics as $\partial_t u = L\,\delta E/\delta u + M\,\delta S/\delta u$, the sum of a reversible bracket generated by an energy functional E through a skew-adjoint operator L, and an irreversible bracket generated by an entropy functional S through a positive semi-definite operator M. Thermodynamic consistency—the first and second laws—follows from two *degeneracy conditions*, $L\,\delta S/\delta u = 0$ and $M\,\delta E/\delta u = 0$, stating that the reversible bracket produces no entropy and the irreversible bracket conserves energy. Carrying this to function space raises difficulties absent in the finite-dimensional case: the variational derivatives are densities rather than vectors, the operators must act consistently across resolutions and spatial dimensions, and—crucially—the degeneracy conditions couple the operators to the *state-dependent* gradient directions $\delta E/\delta u$ and $\delta S/\delta u$. Enforcing degeneracy with a soft penalty is known to be insufficient: the conditions hold only approximately, and the resulting violations compound over a rollout.

We therefore enforce the degeneracy conditions *exactly, by construction.* We learn E and S as scalar-valued neural operators and parameterize L and M as diagonal Fourier multipliers—skew-adjoint and positive semi-definite respectively for any parameter values—sandwiched between the rank-one projections that remove the conjugate gradient direction. The reversible operator is projected off $\delta S/\delta u$ and the irreversible operator off $\delta E/\delta u$, so $L\,\delta S/\delta u = 0$ and $M\,\delta E/\delta u = 0$ hold identically. The construction needs no penalty term, no projection of the predicted update, and no free residual; the degeneracy conditions hold identically, so in continuous time the learned energy is conserved and entropy produced to machine precision (residuals $\sim 10^{-13}$) regardless of initialization, dimension, or resolution, with the explicit update used for rollouts adding only the standard $\mathcal{O}(\Delta t^2)$ energy drift. Because the multipliers act on a fixed band of Fourier modes and the projections are scale-invariant ratios, the entire operator is resolution- and dimension-independent: a model trained at one grid evaluates zero-shot at finer grids with its guarantees intact.

A subtlety accompanies this expressiveness. The tuple (E, S, L, M) realizing a given flow is *not unique*—a well-known gauge freedom of GENERIC (Öttinger, 2005). The structural guarantees are gauge-invariant, but the relative attribution of the dynamics to the reversible and irreversible channels is not, so any per-channel interpretation of a learned model must be read with care. We make this explicit and introduce a gauge-invariant dissipation diagnostic, defined through a fixed quadratic energy rather than the learned functionals, that distinguishes genuinely reversible from dissipative dynamics and can be compared directly against ground truth.

Empirically, we study three operator backbones (1D and 2D FNOs and DeepONet) on four PDEs spanning reversible (advection), dissipative (heat), and mixed (Burgers, damped wave) dynamics. The structural guarantees hold to machine precision in every configuration and are preserved zero-shot across a 4× super-resolution range. The gauge-invariant diagnostic correctly identifies the reversible system and recovers the ground-truth ordering of physical dissipation. On accuracy, GENERIC-FNO is competitive with unconstrained and energy-penalized baselines and outperforms them on several dissipative and mixed problems at comparable or fewer parameters; we are equally explicit about where the structural constraint costs accuracy, most notably on pure linear transport in the FNO backbone.

Our contributions are:

- **The first neural operator with full GENERIC structure in function space.** GENERIC-FNO learns the energy and entropy functionals and the reversible and irreversible operators that together produce thermodynamically consistent dynamics in any spatial dimension (§4).
- **Degeneracy by construction.** A projection-sandwiched diagonal operator parameterization enforces both degeneracy conditions exactly—no penalty, no update projection, no residual—so the structural identities, and the continuous-time energy conservation and entropy production they imply, hold to machine precision for any initialization, dimension, or resolution; the explicit rollout adds only an $\mathcal{O}(\Delta t^2)$ drift (§4.2, §4.3, §6).
- **Gauge freedom and a gauge-invariant diagnostic.** We identify the non-uniqueness of the learned (E, S, L, M) decomposition, separate which claims are gauge-invariant from which are not, and provide a falsifiable dissipation diagnostic that references no learned quantity (§4.4).

- **Resolution- and dimension-independent guarantees.** We show the exact structural guarantees, not only the accuracy, transfer zero-shot across a 4× super-resolution range (§5).
- **An honest cross-backbone study.** Across three backbones and four PDEs we report where thermodynamic structure helps accuracy and where it does not, including a diagnosed and mitigated coarse-grid limitation for reversible transport (§5, §6).

2 Related Work

Neural operators. Neural operators learn mappings between function spaces and, unlike standard networks, are designed to be discretization-convergent. The Fourier Neural Operator (Li et al., 2021) parameterizes the integral kernel in the spectral domain and evaluates across resolutions; DeepONet (Lu et al., 2021) factorizes the operator into branch and trunk networks; and a broad subsequent literature (Kovachki et al., 2023) has extended these to new architectures and domains. These models are flexible black-box surrogates, but they encode no conservation or dissipation law, and we show that this is precisely what causes their autoregressive rollouts to leave the physically admissible manifold.

Structure preservation in finite dimensions. A large body of work builds physical structure into learned dynamics for finite-dimensional state vectors. Hamiltonian and Lagrangian neural networks and symplectic integrators learn reversible, energy-conserving dynamics (Greydanus et al., 2019; Cranmer et al., 2020; Jin et al., 2020), but model only the conservative half of the picture. The thermodynamically complete picture is given by the GENERIC (metriplectic) formalism (Grmela & Öttinger, 1997; Öttinger, 2005), and a line of work learns its ingredients directly: structure-preserving neural networks (Hernández et al., 2021), OnsagerNet (Yu et al., 2021), machine-learned brackets for irreversible processes (Lee et al., 2021), GFINNs (Zhang et al., 2022), and reversible– irreversible splittings (Gruber et al., 2023).

Recent work continues in this vein—efficiently parameterized neural metriplectic systems with by-construction guarantees and universal-approximation results (Gruber et al., 2025), and thermodynamics-informed graph networks with a node-local metriplectic bias (Tierz et al., 2025). Two concurrent directions are closest to ours and worth distinguishing precisely. Baheri & Lindemann (2025) learn dissipative dynamics by conditional flow matching with a metriplectic vector field $J\nabla H_\theta - G_\theta \nabla \Phi_\theta$ and a structure-preserving Strang–prox sampler, with degeneracy enforced hard or soft; the state, however, is a finite-dimensional vector (demonstrated on a damped pendulum), and the extension to PDEs is posed as future work. Oprisa & Toth (2026) repurpose the metriplectic equation as a general-purpose architecture primitive, evolving learned latent fields for image recognition, constraint satisfaction, and language modeling; there the dynamics operate on finite-dimensional latent grids with a canonical symplectic bracket, and the objective is computation itself rather than a surrogate for physical dynamics—no learned energy–entropy pair with mutual degeneracy in function space is involved. All of these methods remain tied to finite-dimensional state vectors, graphs, or latent grids; their parameterizations of L and M scale with the state dimension, and they do not confront the function-space issues—variational derivatives as densities, resolution and dimension independence—that a field theory requires.

A recurring and important lesson from this literature is that the degeneracy conditions must be enforced *by construction* rather than through a soft penalty (Lee et al., 2021; Zhang et al., 2022; Gruber et al., 2025): penalized models satisfy the conditions only approximately, and the residual violations accumulate over a rollout.

Structure-preserving neural operators. In function space, physical structure has been incorporated mainly as symmetry or as a single invariant. Physics-informed losses (Raissi et al., 2019) and conservation-constrained operators softly penalize residuals of a known equation or conserved quantity, and equivariant and Hamiltonian operator variants impose symmetry or reversible structure. To our knowledge, however, no neural operator embeds the *full* GENERIC structure—a learned energy and a learned entropy, with reversible and irreversible operators satisfying the mutual degeneracy conditions—in function space. GENERIC-FNO occupies exactly this gap: it carries the by-construction degeneracy of the finite-dimensional metriplectic networks into the resolution- and dimension-independent setting of neural operators, parameterizing L and

M as diagonal Fourier multipliers (rather than dimension-scaling matrices) so that the construction is intrinsically a field-space object. This also separates GENERIC-FNO from the recent thermodynamics-informed graph and mesh networks that achieve discretization flexibility (Tierz et al., 2025): those carry the metriplectic structure node-locally on a graph or mesh—a flexible but still discrete object—whereas our L and M are spectral multipliers on a fixed band of Fourier modes, making the model a genuine function-space map that is dimension-agnostic and zero-shot super-resolves the underlying continuous dynamics (§5.3) rather than merely re-meshing.

Gauge freedom in metriplectic models. That a given flow admits more than one (E, S, L, M) representation is classical in the GENERIC literature (Öttinger, 2005), but it has received little attention in the learning setting, where the learned functionals are often interpreted directly. We make the consequence explicit—structural guarantees are gauge-invariant while per-channel attribution is not—and propose a diagnostic that depends only on the predicted dynamics and a fixed quadratic energy, so that reversibility and the rate of physical dissipation can be measured without reference to the (gauge-dependent) learned functionals.

Thermodynamic consistency as an inductive bias. Interest in thermodynamically consistent learning is driven by applications where violating the second law is not a cosmetic error but a route to instability and unphysical predictions. Data-driven closure and subgrid models for turbulence and coarse-grained dynamics can inject or remove energy spuriously unless dissipation is constrained (Duraisamy et al., 2019); learned constitutive models for complex and viscoelastic fluids—the setting in which GENERIC was originally developed (Öttinger, 2005)—must respect energy balance and entropy production to remain stable under deployment (Hernández et al., 2021); and thermodynamics-aware reduced-order models target long-horizon rollout stability that unconstrained surrogates lack (Lee et al., 2021). In all of these, thermodynamic consistency is valued less as a property to be reported than as a strong, transferable inductive bias that keeps a learned model on the physical manifold far from its training data. Our contribution makes that bias available, exactly, to the neural-operator models increasingly used for these high-dimensional field problems.

3 Background: The GENERIC Formalism

We consider an evolving field $u(x,t)$ on a periodic spatial domain $\Omega \subset \mathbb{R}^d$, with the L^2 inner product $\langle f, g\rangle = \int_\Omega f(x)\, g(x)\, dx$ and variational (functional) derivatives $\delta F/\delta u$ defined by $\frac{d}{d\epsilon} F[u + \epsilon\, \eta]\big|_{\epsilon=0} = \langle \delta F/\delta u,\, \eta\rangle$ for all admissible η.

The GENERIC formalism (Grmela & Öttinger, 1997; Öttinger, 2005) (General Equation for the Non-Equilibrium Reversible–Irreversible Coupling) expresses the dynamics of u as the sum of a reversible and an irreversible generator,

$$\frac{\partial u}{\partial t} \;=\; \underbrace{L\,\frac{\delta E}{\delta u}}_{\text{reversible}} \;+\; \underbrace{M\,\frac{\delta S}{\delta u}}_{\text{irreversible}}\,, \tag{1}$$

where $E[u]$ is the total energy, $S[u]$ is the entropy, L is a skew-adjoint operator ($\langle f, Lg\rangle = -\langle Lf, g\rangle$) generating the reversible Hamiltonian-like dynamics, and M is a self-adjoint, positive semi-definite operator ($\langle f, Mf\rangle \geq 0$) generating the irreversible friction. Equivalently, in bracket form, $\dot{F} = \{F, E\} + [F, S]$ for any observable $F[u]$, with a skew (Poisson) bracket $\{F, G\} = \langle \delta F/\delta u, L\, \delta G/\delta u\rangle$ and a symmetric, positive semi-definite (dissipative) bracket $[F, G] = \langle \delta F/\delta u, M\, \delta G/\delta u\rangle$.

Degeneracy conditions. GENERIC requires the two generators to be mutually orthogonal in a precise sense:

$$L\,\frac{\delta S}{\delta u} = 0, \qquad M\,\frac{\delta E}{\delta u} = 0. \tag{2}$$

The first states that the reversible dynamics do not change the entropy; the second that the irreversible dynamics do not change the energy. In bracket form they read $\{\,\cdot\,, S\} = 0$ and $[\,\cdot\,, E] = 0$.

The first and second laws. Conditions equation 2, together with the skewness of L and the positivity of M, yield thermodynamic consistency directly. The energy is conserved,

$$\dot{E} = \left\langle \frac{\delta E}{\delta u}, \dot{u} \right\rangle = \underbrace{\left\langle \frac{\delta E}{\delta u}, L\,\frac{\delta E}{\delta u} \right\rangle}_{=0\ (L\ \text{skew})} + \underbrace{\left\langle \frac{\delta E}{\delta u}, M\,\frac{\delta S}{\delta u} \right\rangle}_{=0\ (\text{degeneracy})} = 0, \tag{3}$$

and the entropy is non-decreasing,

$$\dot{S} = \left\langle \frac{\delta S}{\delta u}, \dot{u} \right\rangle = \underbrace{\left\langle \frac{\delta S}{\delta u}, L\,\frac{\delta E}{\delta u} \right\rangle}_{=0\ (\text{degeneracy})} + \left\langle \frac{\delta S}{\delta u}, M\,\frac{\delta S}{\delta u} \right\rangle \ \geq\ 0, \tag{4}$$

the last inequality following from the positive semi-definiteness of M. Thus the first law is exact and the second law holds with the entropy production rate $[S, S] = \langle \delta S/\delta u, M\,\delta S/\delta u \rangle \geq 0$.

Limiting cases. The formalism interpolates between purely reversible and purely irreversible dynamics. When $M = 0$ the evolution is Hamiltonian and conserves energy (e.g. linear advection, which transports u while conserving $\frac{1}{2}\|u\|^2$); when $L = 0$ it is a pure gradient flow that monotonically dissipates (e.g. diffusion); and genuinely coupled systems use both generators at once. The PDEs we study in §5 are chosen to populate this spectrum.

Non-uniqueness. Finally, the representation is not unique: distinct tuples (E, S, L, M) can generate the same flow $\partial_t u$ (Öttinger, 2005). This gauge freedom is harmless for the guarantees above—which depend only on the flow and the algebraic conditions equation 2—but it bears directly on how a *learned* decomposition may be interpreted, a point we return to in §4.4.

4 Method: GENERIC-FNO

We parameterize the GENERIC dynamics of §3 (Eq. equation 1) with two learned scalar functionals, $E_\theta[u]$ and $S_\phi[u]$, and two operators L and M that satisfy the degeneracy conditions equation 2 *exactly, by construction.* Nothing in the model is supervised toward a known energy or entropy; E_θ and S_ϕ are shaped only by the data-fitting loss through the dynamics, so thermodynamic consistency is a property of the architecture rather than of an auxiliary penalty. Figure 1 summarizes the forward map.

4.1 Learned energy and entropy functionals

Each functional maps the input field to a single scalar, $E_\theta, S_\phi : u \mapsto \mathbb{R}$. We implement them as a neural-operator backbone—a Fourier neural operator (Li et al., 2021) in our main model, a DeepONet (Lu et al., 2021) in the appendix study—that maps u to a feature *density* field, followed by spatial averaging and a small MLP head:

$$E_\theta[u] = h_\theta\Big(\tfrac{1}{|\Omega|}\int_\Omega g_\theta(u)\,dx\Big), \qquad S_\phi[u] = h_\phi\Big(\tfrac{1}{|\Omega|}\int_\Omega g_\phi(u)\,dx\Big), \tag{5}$$

where $g_{(\cdot)}$ is the backbone and $h_{(\cdot)}$ a two-layer head. The spatial *average* (not a fixed flattening) is what keeps the functionals resolution-independent. The variational derivatives are obtained by reverse-mode automatic differentiation through the field,

$$\frac{\delta E}{\delta u} = \nabla_u E_\theta(u), \qquad \frac{\delta S}{\delta u} = \nabla_u S_\phi(u), \tag{6}$$

computed with `create_graph=True` so the dynamics remain differentiable end-to-end.

4.2 Structure-preserving operators by construction

We take L and M to be diagonal Fourier multipliers acting on the lowest m modes, a parameterization that is identical in any spatial dimension and on any grid. Writing $\widehat{(\cdot)}$ for the (real) FFT, the base multipliers are

$$\widehat{D_L}(k) = \mathrm{i}\,a(k), \qquad \widehat{D_M}(k) = |b(k)|^2 \geq 0, \tag{7}$$

with learnable real $a(k)$ and complex $b(k)$. Because the inverse real FFT enforces Hermitian symmetry, D_L is exactly skew-adjoint and D_M exactly self-adjoint positive semi-definite on real fields, for *any* parameter values.

Degeneracy by construction. A bare multiplier cannot satisfy the degeneracy conditions equation 2, which couple the operators to the *state-dependent* directions $\delta E/\delta u$ and $\delta S/\delta u$. Following the by-construction principle of finite-dimensional metriplectic networks (Lee et al., 2021; Zhang et al., 2022; Gruber et al., 2023) but lifting it to function space, we sandwich each multiplier between the rank-one L^2 projection that removes the conjugate direction. Let

$$P_E\, v = \frac{\langle \delta E/\delta u,\, v\rangle}{\langle \delta E/\delta u,\, \delta E/\delta u\rangle}\,\frac{\delta E}{\delta u}, \qquad P_S\, v = \frac{\langle \delta S/\delta u,\, v\rangle}{\langle \delta S/\delta u,\, \delta S/\delta u\rangle}\,\frac{\delta S}{\delta u}, \tag{8}$$

and define

$$L = (I - P_S)\, D_L\, (I - P_S), \qquad M = (I - P_E)\, D_M\, (I - P_E). \tag{9}$$

Since $(I - P_S)\,\delta S/\delta u = 0$ and $(I - P_E)\,\delta E/\delta u = 0$, the degeneracy conditions hold identically: $L\,\delta S/\delta u = 0$ and $M\,\delta E/\delta u = 0$. The projections are symmetric idempotents, so the sandwich preserves the structure of the base multipliers—L remains skew-adjoint and M positive semi-definite. The reversible operator is thus projected off the entropy gradient and the irreversible operator off the energy gradient, the directions supplied (dashed in Fig. 1) by the *opposite* functional. The resulting field evolution

$$u_{t+1} = u_t + \big(\, L\,\tfrac{\delta E}{\delta u} + M\,\tfrac{\delta S}{\delta u}\,\big) \tag{10}$$

satisfies the first and second laws by construction, with no penalty term, no projection of the update, and no free residual.

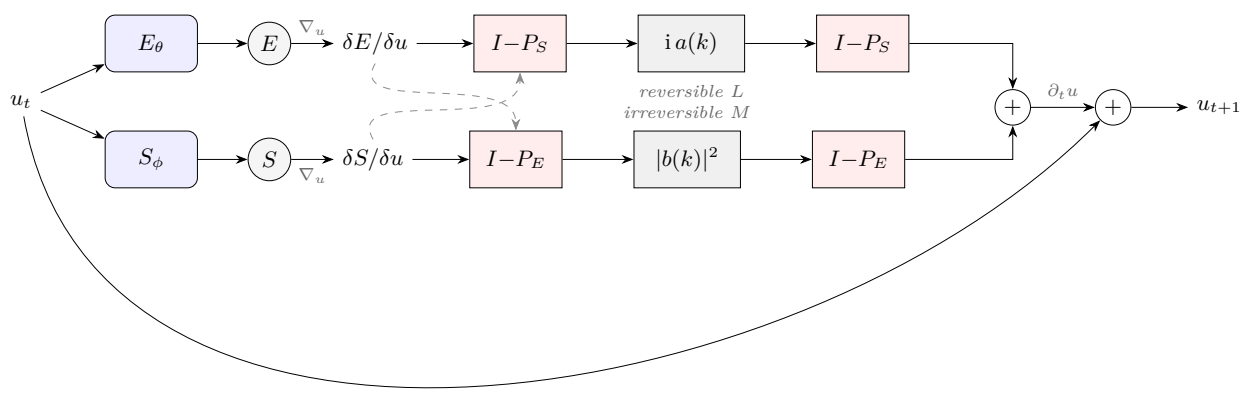

Figure 1: **GENERIC-FNO forward map.** Two neural operators produce scalar functionals $E_\theta[u]$ and $S_\phi[u]$, whose variational derivatives are taken by autodiff. Each diagonal Fourier multiplier—skew $\mathrm{i}\,a(k)$ for the reversible operator L, positive semi-definite $|b(k)|^2$ for the irreversible operator M—is sandwiched between the rank-one projection that removes the conjugate gradient (dashed: $\delta S/\delta u$ feeds the L branch, $\delta E/\delta u$ feeds the M branch). This makes $L\,\delta S/\delta u = 0$ and $M\,\delta E/\delta u = 0$ hold identically, so the metriplectic update $u_{t+1} = u_t + L\,\delta E/\delta u + M\,\delta S/\delta u$ conserves energy and produces entropy by construction.

4.3 Exact guarantees and machine-precision verification

The construction equation 9 makes the GENERIC structure hold to machine precision, independent of training, initialization, dimension, or resolution. At random initialization on a 16×16 grid we measure (maximum over a batch):

Property	Identity	Measured
Reversible skewness	$\langle \delta E/\delta u,\, L\,\delta E/\delta u\rangle = 0$	7×10^{-19}
Energy degeneracy	$\langle \delta E/\delta u,\, M\,\delta S/\delta u\rangle = 0$	3×10^{-13}
Entropy degeneracy	$\langle \delta S/\delta u,\, L\,\delta E/\delta u\rangle = 0$	1×10^{-15}
Energy conservation	$\langle \delta E/\delta u,\, \partial_t u\rangle = 0$	3×10^{-13}
Entropy production	$\langle \delta S/\delta u,\, \partial_t u\rangle \geq 0$	$\geq +7\times10^{-9}$
Operator positivity	$\langle v,\, Mv\rangle \geq 0$	$\geq +48.5$

The semi-discrete (continuous-time) dynamics conserve the learned energy exactly ($\langle \delta E/\delta u, \partial_t u\rangle = 0$ to machine precision); the explicit update equation 10 incurs only the standard $\mathcal{O}(\Delta t^2)$ integration error, which we report per step in §5 and revisit in §6. Three quantities should be kept distinct. (i) The *algebraic degeneracy residuals* (table above) are zero to machine precision by construction, at any state and any resolution. (ii) The *discrete energy drift* of a deployed rollout is not zero but $\mathcal{O}(\Delta t^2)$; we measure $r_E \sim 10^{-6}$ in practice (§5, §6). (iii) The conserved object is the *learned* energy E_θ, which—like S_ϕ—is defined only up to the gauge of §4.4; the physically meaningful, gauge-invariant statement is therefore the dissipation diagnostic r_{mech}, not the value of E_θ itself. The construction guarantees (i) exactly; (ii) and (iii) are what the experiments quantify.

4.4 Gauge freedom and a gauge-invariant dissipation measure

We now make precise point (iii) of §4.3: the construction guarantees an *exact* thermodynamic structure, but not a *unique* energy and entropy. As noted in §3, the tuple (E, S, L, M) realizing a given flow is not unique (Öttinger, 2005), so the learned E_θ, S_ϕ—and any quantity read off them—are fixed only up to this gauge. The exact structural identities of §4.3 are gauge-invariant, depending only on the flow and the algebraic degeneracy conditions; the relative weight of the reversible and irreversible channels is not. We therefore separate two kinds of diagnostic. The *mechanism* is described by the (gauge-dependent) dissipative fraction $\rho_M = \|M\,\delta S/\delta u\|/(\|L\,\delta E/\delta u\| + \|M\,\delta S/\delta u\|)$ and the normalized learned-entropy production $r_S = \langle \delta S/\delta u, \partial_t u\rangle/(\|\delta S/\delta u\|\,\|\partial_t u\|)$, both of which we report only in the appendix. The *thermodynamic behavior* we report in the main text is gauge-invariant: using the fixed quadratic energy $Q[u] = \frac{1}{2}\|u\|^2$, which references no learned quantity, we measure the normalized dissipation rate

$$r_{\text{mech}} = \frac{-\langle u, \partial_t u\rangle}{\|u\|\,\|\partial_t u\|} \in [-1, 1], \tag{11}$$

which is > 0 for genuine dissipation of physical energy and ≈ 0 for reversible dynamics. As a ground-truth reference we compute the analogous relative rate $\Pi^\star = (Q[u] - Q[u^\star_{t+1}])/Q[u]$ from the reference trajectory, depending only on the data. Because r_{mech} is bounded and insensitive to single-step prediction error, it is stable where pointwise accuracy degrades. The decisive test of thermodynamic consistency is the reversible one: a faithful operator must add *no* spurious dissipation when $\Pi^\star \approx 0$, and this is gauge-invariant because reversible data admits no positive entropy-production representation. We show in §5 that r_{mech} vanishes on the reversible benchmark and recovers the correct ordering of physical dissipation across PDEs.

4.5 Discretization, resolution independence, and complexity

All operators act through the FFT and the L^2 inner product on the grid. The multipliers $a(k), b(k)$ are defined on a fixed band of m low-frequency modes, applied to the corresponding modes of whatever grid the input lives on; the projections are scale-invariant ratios. Consequently the entire map is resolution- and dimension-independent: a model trained at one grid evaluates zero-shot at finer grids, and the same code runs in 1D and 2D unchanged. Each projection costs two inner products and a scaled subtraction ($\mathcal{O}(N)$ for N grid points), negligible beside the FFTs, so the structural machinery adds essentially no cost over a plain spectral operator.

5 Experiments

5.1 Setup

PDEs. We evaluate on four PDEs that populate the reversible–irreversible spectrum of §3. The (dissipative) heat equation is a pure gradient flow ($L \to 0$),

$$\partial_t u = \nu \nabla^2 u, \tag{12}$$

which monotonically dissipates $\frac{1}{2}\|u\|^2$. Linear advection is reversible ($M \to 0$),

$$\partial_t u + c\,(\partial_x u + \partial_y u) = 0, \tag{13}$$

transporting u while conserving $\frac{1}{2}\|u\|^2$ exactly; it is the clean, fully observed reversible test. The viscous Burgers equation is mixed, combining nonlinear transport with diffusion,

$$\partial_t u + u\,(\partial_x u + \partial_y u) = \nu \nabla^2 u. \tag{14}$$

We report 2D forms; the 1D analogues used in the appendix are identical with a single spatial derivative. The fourth, a damped wave, is second-order in time and is treated separately in §5.2.

Data, backbones, and baselines. Initial conditions are band-limited random fields, integrated spectrally to generate trajectories. We study three operator backbones: a 2D FNO (the main model), a 1D FNO, and a 1D DeepONet. For the FNO backbones we compare GENERIC-FNO against an unconstrained FNO and an energy-penalized variant (EP-FNO) that adds a soft energy-conservation loss; for DeepONet we compare GENERIC-DeepONet against an unconstrained DeepONet. GENERIC models receive *no* supervision on E or S. All results are reported as mean $\pm$ std over three seeds, resampling both the data and the initialization.

Metrics. We report one-step and 10-step autoregressive *rollout* relative L^2 error. For GENERIC models we additionally report the structural residual $r_E = |\langle \delta E/\delta u, \partial_t u\rangle| / (\|\delta E/\delta u\|\,\|\partial_t u\|)$ (zero iff energy is conserved by the step) and the gauge-invariant dissipation rate r_{mech} of Eq. equation 11, compared against the ground-truth rate $\Pi^\star$.

5.2 Predictive accuracy

Table 1: **2D FNO, three-seed accuracy.** GENERIC-FNO uses fewer than half the parameters of the baselines. Bold = best per PDE. GENERIC wins heat (with $\sim$10$\times$ lower seed variance), Burgers, and the damped wave; it trades accuracy for structure on pure transport (advection).

PDE	Model	Params	L^2 (1-step)	L^2 (rollout)
Heat	FNO	2.10M	0.021 ± 0.006	0.113 ± 0.034
	EP-FNO	2.10M	0.020 ± 0.002	0.111 ± 0.008
	GENERIC-FNO	1.00M	$\mathbf{0.018 \pm 0.001}$	$\mathbf{0.091 \pm 0.003}$
Advection	FNO	2.10M	$\mathbf{0.022 \pm 0.001}$	$\mathbf{0.119 \pm 0.004}$
	EP-FNO	2.10M	0.025 ± 0.001	0.132 ± 0.006
	GENERIC-FNO	1.00M	0.047 ± 0.007	0.179 ± 0.020
Burgers	FNO	2.10M	0.034 ± 0.005	0.175 ± 0.025
	EP-FNO	2.10M	0.026 ± 0.002	0.138 ± 0.011
	GENERIC-FNO	1.00M	$\mathbf{0.004 \pm 0.001}$	$\mathbf{0.026 \pm 0.003}$

Table 2: **1D DeepONet, three-seed accuracy.** The GENERIC structure ports to a non-spectral backbone, with large gains on heat and advection. Here the constrained model uses *more* parameters than the baseline (the E, S branch); the FNO comparison in Table 1, where GENERIC wins at fewer parameters, shows the gains are structural rather than from capacity.

PDE	Model	Params	L^2 (1-step)	L^2 (rollout)
Heat	DeepONet	58K	0.010 ± 0.001	0.057 ± 0.004
	GENERIC-DeepONet	175K	$\mathbf{0.002} \pm \mathbf{0.000}$	$\mathbf{0.008} \pm \mathbf{0.001}$
Advection	DeepONet	58K	0.037 ± 0.002	0.201 ± 0.009
	GENERIC-DeepONet	175K	$\mathbf{0.003} \pm \mathbf{0.000}$	$\mathbf{0.018} \pm \mathbf{0.002}$
Burgers	DeepONet	58K	0.004 ± 0.000	0.020 ± 0.002
	GENERIC-DeepONet	175K	$\mathbf{0.003} \pm \mathbf{0.000}$	$\mathbf{0.016} \pm \mathbf{0.001}$

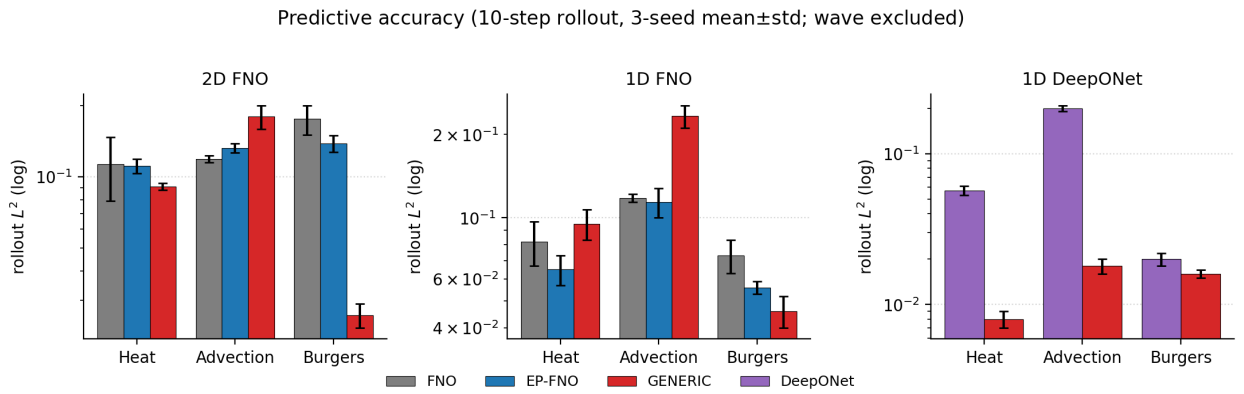

Figure 2: Rollout L^2 error with three-seed error bars, one panel per backbone (2D FNO, 1D FNO, 1D DeepONet), wave excluded. GENERIC (red) versus unconstrained and energy-penalized baselines. Visual companion to Tables 1–2; bars are log-scaled.

On the 2D FNO backbone (Table 1, Fig. 2), GENERIC-FNO improves on both baselines for the dissipative and mixed problems at less than half the parameters: Burgers rollout error falls by 85% relative to FNO (0.026 vs. 0.175), and on heat GENERIC is both more accurate on average and markedly more stable across seeds (0.091 ± 0.003 vs. 0.113 ± 0.034), the structural constraint acting as a regularizer. The one accuracy cost is advection, where rollout error rises to 0.179 vs. 0.119 for the free FNO. We read this as the principled price of the constraint rather than a defect: advection is exactly reversible, so GENERIC-FNO is *by construction* barred from using its irreversible channel—its dissipation diagnostic stays at $r_{\mathrm{mech}} \approx 0$ there (§5.4)—and it thereby forgoes the extra fitting flexibility a structurally unconstrained operator can spend on pure transport. The accuracy gap and the vanishing r_{mech} are two readings of the same fact: on a flow that produces no entropy, the model declines to manufacture any. The place GENERIC-FNO is least accurate is thus exactly where its thermodynamic guarantee binds hardest—a cost we make explicit rather than one we would erase by letting the model add spurious dissipation. The DeepONet study (Table 2) shows the construction ports to a non-spectral backbone, with GENERIC-DeepONet reducing heat and advection rollout error by 85–91%, and the advection cost seen on the FNO backbone does not appear here—suggesting that cost is a property of the spectral backbone rather than of the GENERIC constraint. These DeepONet gains are *not* parameter-controlled, however: GENERIC-DeepONet carries the additional E, S branch and is larger than the baseline (175K vs. 58K), so on their own they cannot separate structure from capacity. Our parameter-controlled evidence is therefore twofold: the 2D FNO comparison, where GENERIC-FNO wins heat and Burgers at *less than half* the baseline parameter count (1.00M vs. 2.10M), and a matched-budget DeepONet ablation (Appendix C)—trained at equal budget, both at $\approx$58K and at $\approx$175K, GENERIC-DeepONet has

lower rollout error at every PDE and both budgets, and growing the baseline to 175K does not close the gap. The headline DeepONet numbers in Table 2 thus reflect the construction, not its larger parameter count. On the smaller-capacity 1D FNO (full table in Appendix B), the picture is more mixed—GENERIC wins Burgers (-37%) but loses heat and advection—consistent with the gains growing with backbone capacity and dimension.

Damped wave. We also evaluated the damped wave equation $\partial_{tt}u = c^2\nabla^2 u - \gamma\,\partial_t u$, a coupled reversible (propagation) and irreversible (damping) system—exactly the regime GENERIC is designed for. In 2D, GENERIC-FNO attains rollout error 0.106 ± 0.009 versus 0.190 ± 0.031 for FNO (a 44% reduction) at half the parameters, its best relative margin among the reversible–irreversible problems. We omit wave from the tables because the equation is *second order in time* while we observe only the field u and not its velocity $\partial_t u$: the one-step map is therefore non-Markovian in the observed state. At the 2D resolution enough of the velocity is recoverable from spatial structure for all models to learn it, but in 1D at $n_x = 64$ this partial observation is severe enough that *every* model, including the unconstrained FNO, collapses to a rollout error of ≈ 0.48 (Appendix B). We thus report the favorable, well-posed 2D result in prose and exclude wave from the headline tables.

5.3 Exact structure and zero-shot super-resolution

The structural guarantees of §4.3 hold to machine precision in every configuration above ($r_E \sim 10^{-6}$–10^{-8}, degeneracy residuals $\sim 10^{-13}$). Because the operators act on a fixed band of Fourier modes and the projections are scale-invariant, these guarantees—and not only the accuracy—transfer across resolution. Table 3 and Figure 3 train at 64×64 and evaluate zero-shot up to 256×256 on band-limited data (fixed maximum frequency, so the continuous dynamics are identical across grids). GENERIC-FNO's rollout accuracy is flat-to-improving across the $4\times$ range and 4–8$\times$ better than FNO at every grid, while the structural residual stays $r_E \lesssim 10^{-5}$. The reversible diagnostic r_{mech} likewise remains ≈ 0 for advection at every resolution, so the thermodynamic structure is preserved zero-shot. Advection's *accuracy* degrades at the coarsest grid for the reversible operator; this is an explicit-integrator effect we isolate and mitigate in §6, not a breakdown of the structure.

Table 3: **Zero-shot super-resolution.** Trained at 64, evaluated zero-shot to 256 (rollout L^2). GENERIC-FNO accuracy is flat-to-improving and the structural residual r_E stays negligible; the model is genuinely resolution-invariant in both accuracy and guarantees.

		evaluation resolution				
PDE	Model	64	96	128	192	256
Heat	FNO	0.112	0.097	0.099	0.091	0.099
	GENERIC-FNO	**0.040**	**0.026**	**0.023**	**0.022**	**0.023**
Burgers	FNO	0.107	0.145	0.131	0.164	0.124
	GENERIC-FNO	**0.033**	**0.023**	**0.021**	**0.019**	**0.015**
GENERIC-FNO r_E (range)		$3\times10^{-7} \;\rightarrow\; 4\times10^{-5}$				

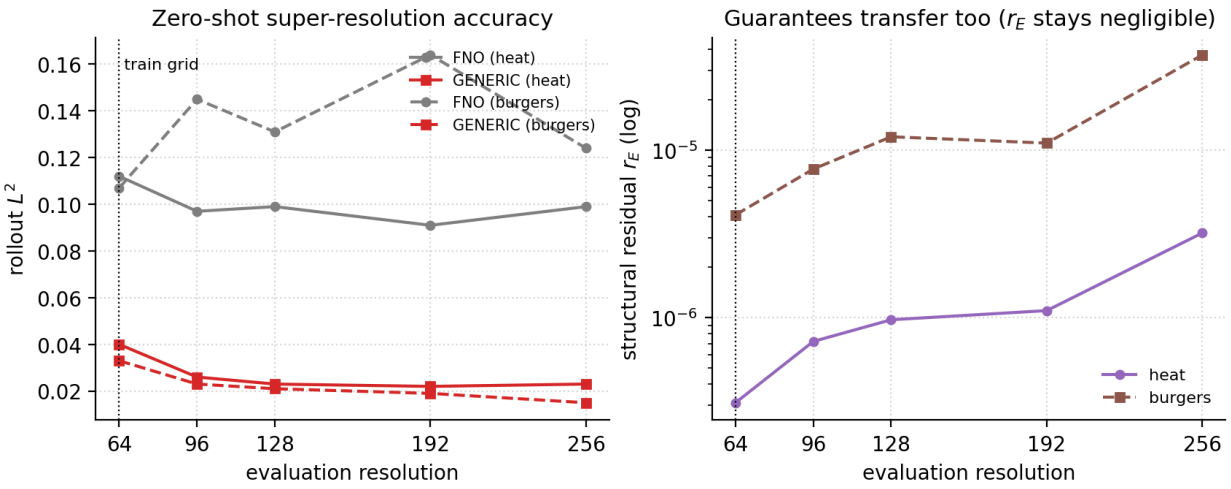

Figure 3: Zero-shot super-resolution (trained at 64). *Top:* rollout L^2 versus evaluation resolution for FNO and GENERIC-FNO on heat and Burgers (flat-to-improving across the 4× range). *Bottom:* the structural residual r_E and energy drift $|\mathrm{d}E/\mathrm{step}|$ of GENERIC-FNO versus resolution (log scale), confirming the guarantees, not only the accuracy, transfer zero-shot. Dotted line marks the training grid.

5.4 Gauge-Invariant Dissipation

Table 4 reports the gauge-invariant dissipation rate r_{mech} against the ground-truth rate $\Pi^\star$, across all three backbones, and Figure 4 visualizes it. Two points hold consistently. First, $r_{\mathrm{mech}} \approx 0$ for the reversible advection system—the model adds essentially no spurious dissipation, the decisive thermodynamic test—while $r_{\mathrm{mech}} > 0$ for the dissipative and mixed systems. Second, within every backbone the across-PDE ordering of r_{mech} reproduces that of $\Pi^\star$ exactly (heat > Burgers > advection; Spearman $\rho = 1$ in each backbone, and $\rho \approx 0.87$, $p = 0.002$ pooled over all nine points; Appendix F, Fig. 8 plots the nine points), so the model recovers *which* dynamics dissipate more physical energy without ever being told E or S.

We compare r_{mech} and $\Pi^\star$ *ordinally*, not by magnitude: $r_{\mathrm{mech}} \in [-1, 1]$ is a normalized alignment (a cosine between u and $\partial_t u$), whereas $\Pi^\star$ is a fractional one-step change of Q, so the two carry different normalizations and are not expected to agree in absolute value (e.g. heat on the 2D FNO: $r_{\mathrm{mech}} = 0.31$ vs. $\Pi^\star = 0.029$); only the sign and the across-PDE ranking are claimed. A related caveat: a positive r_{mech} records a decay of the fixed energy $Q = \frac{1}{2}\|u\|^2$—it would flag *any* norm-shrinking update. This is meaningful here because for the scalar PDEs we study Q *is* the physical mechanical energy that the true dynamics conserve (advection) or dissipate (heat, Burgers), so its decay is genuine physical dissipation rather than an artifact; on a system where Q is not the relevant invariant, r_{mech} would be redefined with the appropriate energy. Because both quantities reference only this fixed energy, the conclusion is independent of the gauge-dependent learned decomposition.

An illustrative single-rollout view of the learned E_θ and S_ϕ is in Appendix F (Fig. 10); being gauge-dependent it is shown only as interpretive color.

Table 4: **Gauge-invariant dissipation** (three-seed mean ± std). $r_{\text{mech}} \approx 0$ identifies the reversible system; within every backbone the across-PDE ordering of r_{mech} matches that of the ground-truth $\Pi^\star$ (Spearman $\rho = 1$ per backbone; $\rho \approx 0.87$ pooled). The two columns use different normalizations—$r_{\text{mech}} \in [-1, 1]$ is a cosine rate, $\Pi^\star$ a fractional one-step energy change—so they are compared ordinally, not by magnitude.

Backbone	PDE	r_{mech}	$\Pi^\star$ (truth)
2D FNO	Heat	$+0.31 \pm 0.11$	$+2.9 \times 10^{-2}$
	Advection	$+0.003 \pm 0.002$	$+1.4 \times 10^{-7}$
	Burgers	$+0.011 \pm 0.004$	$+1.2 \times 10^{-3}$
1D FNO	Heat	$+0.27 \pm 0.10$	$+1.7 \times 10^{-2}$
	Advection	$+0.029 \pm 0.028$	$+1.0 \times 10^{-7}$
	Burgers	$+0.051 \pm 0.026$	$+1.6 \times 10^{-3}$
1D DeepONet	Heat	$+0.83 \pm 0.02$	$+1.7 \times 10^{-2}$
	Advection	$+0.007 \pm 0.008$	$+7.4 \times 10^{-8}$
	Burgers	$+0.25 \pm 0.05$	$+1.4 \times 10^{-3}$

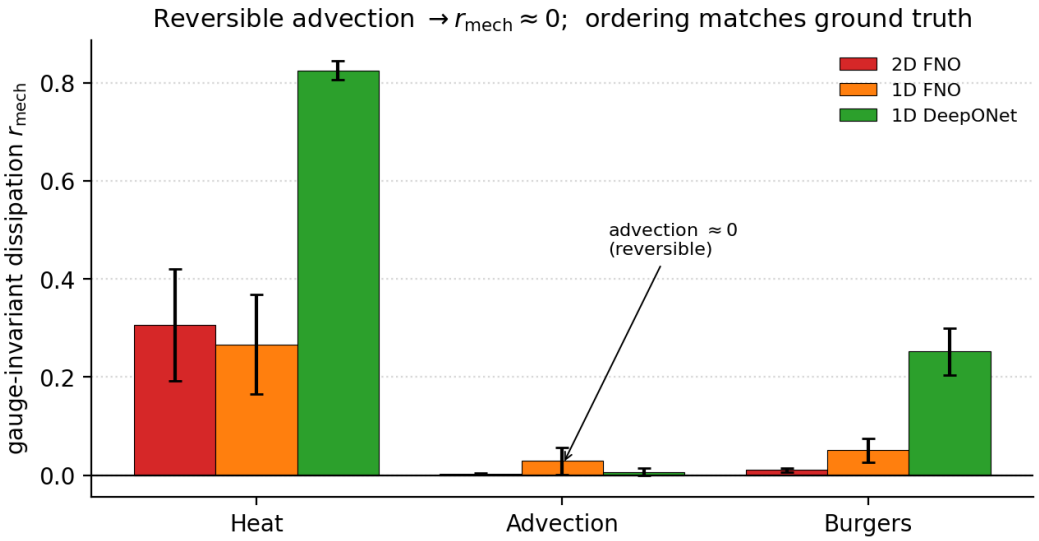

Figure 4: Gauge-invariant dissipation r_{mech} (three-seed mean ± std) per PDE, grouped by backbone. Advection sits at ≈ 0 (the model adds no spurious dissipation to reversible dynamics) while heat and Burgers are positive; the across-PDE ordering matches the ground-truth $\Pi^\star$ in every backbone. Computed from a fixed quadratic energy, so it is independent of the learned (gauge-dependent) E, S.

5.5 Long-Horizon Stability

The motivation for thermodynamic structure is stability far from the training data, so we roll each model out for 200 steps—more than 13× the 15-step horizon used in training—on held-out trajectories, training all three models (FNO, EP-FNO, GENERIC-FNO) under one protocol per PDE (a lighter protocol than §5.2, so the short-horizon errors sit above Table 1, but the across-model comparison is controlled). We report relative rollout error at the training horizon and at 200 steps (Table 5) and the physical energy $Q = \frac{1}{2}\|u\|^2$ along the rollout (Fig. 5).

Table 5: **Long-horizon rollout stability** (relative L^2; training horizon 15 steps). On the dissipative/mixed problems the unconstrained FNO diverges by 200 steps—its energy growing $\approx 9\times$ (heat) and $\approx 5\times$ (Burgers)—while GENERIC-FNO stays bounded; on reversible advection all models phase-drift and the physical energy (Fig. 5) is the discriminating quantity, where GENERIC-FNO conserves it best.

PDE	Model	L^2 @15 (train)	L^2 @200
Heat	FNO	0.158	4.914
	EP-FNO	0.201	0.927
	GENERIC-FNO	0.243	**0.750**
Advection	FNO	0.189	1.363
	EP-FNO	0.161	1.070
	GENERIC-FNO	0.338	1.400
Burgers	FNO	0.082	1.747
	EP-FNO	0.093	1.083
	GENERIC-FNO	0.055	**0.332**

On the dissipative and mixed problems the unconstrained baseline leaves the admissible set: by 200 steps the FNO rollout error reaches 4.91 (heat) and 1.75 (Burgers), and its physical energy has grown without bound—to $\approx 9\times$ its initial value on heat and $\approx 5\times$ on Burgers. GENERIC-FNO instead stays bounded (L^2 of 0.75 and 0.33; its energy remains within a factor of order one of its initial value throughout). The energy penalty helps but does not prevent the drift (EP-FNO energy reaches $\approx 2\times$ on Burgers). GENERIC-FNO is in fact *less* accurate than the baselines at the training horizon on heat (0.24 vs. 0.16)—the expressiveness it trades for structure—yet ends far ahead, precisely because it cannot manufacture the spurious energy growth that drives the baselines' blow-up. On reversible advection, where the true energy is exactly constant, no model blows up in amplitude; all three accumulate transport *phase* error and reach comparable rollout L^2 (≈ 1.1–1.4), so error is not the discriminating quantity. The energy is: GENERIC-FNO holds the constant true energy to within $\approx 9\%$ over 200 steps, while the baselines swing by ≈ 10–18% — confirming that what the structural prior controls is the energy, not the phase. The one honest exception is long-time heat, where GENERIC-FNO *under*-dissipates: its energy plateaus near 0.85 rather than relaxing to the true ≈ 0.1, so it is bounded and physically admissible but does not reproduce the fully-dissipated limit (the dominant contribution to its residual 0.75 error).

6 Limitations and Scope

We are explicit about where the method is constrained, both to guide use and because several limitations are instructive about the construction itself.

Reversible transport and the explicit integrator. The clearest cost of the constraint is on pure linear transport with the FNO backbone, where GENERIC-FNO is less accurate than an unconstrained FNO (Table 1). A sharper version of this appears in the super-resolution study: at the coarsest *training* grid, advection rollout error is anomalously high and worsens as the grid coarsens (Table 6). We traced this to the explicit integrator rather than to the structure. Explicit Euler applied to the skew (reversible) operator has a per-step amplitude growth $\propto \sqrt{1+(\omega\Delta t)^2}$; the pointwise nonlinearities in the E, S functional networks inject high-frequency harmonics that *alias* on a coarse grid, raising the effective ω, and because the reversible regime has $M \approx 0$ there is no dissipation to damp the aliased modes, so the 10-step rollout compounds them. The structural diagnostics stay clean throughout ($r_{\mathrm{mech}} \approx 0$, degeneracy $\sim 10^{-13}$), confirming this is an *integration* effect, not a breakdown of thermodynamic consistency—the heat and Burgers cases, which carry genuine dissipation, are unaffected and remain flat across resolution.

A fourth-order Runge–Kutta integrator with anti-aliased nonlinearities substantially mitigates the blow-up (Table 6), confirming the mechanism. We nonetheless retain explicit Euler as the model of record, because it makes the discrete update *inherit the degeneracy exactly*: the per-step energy residual is $r_E \sim 10^{-7}$, whereas

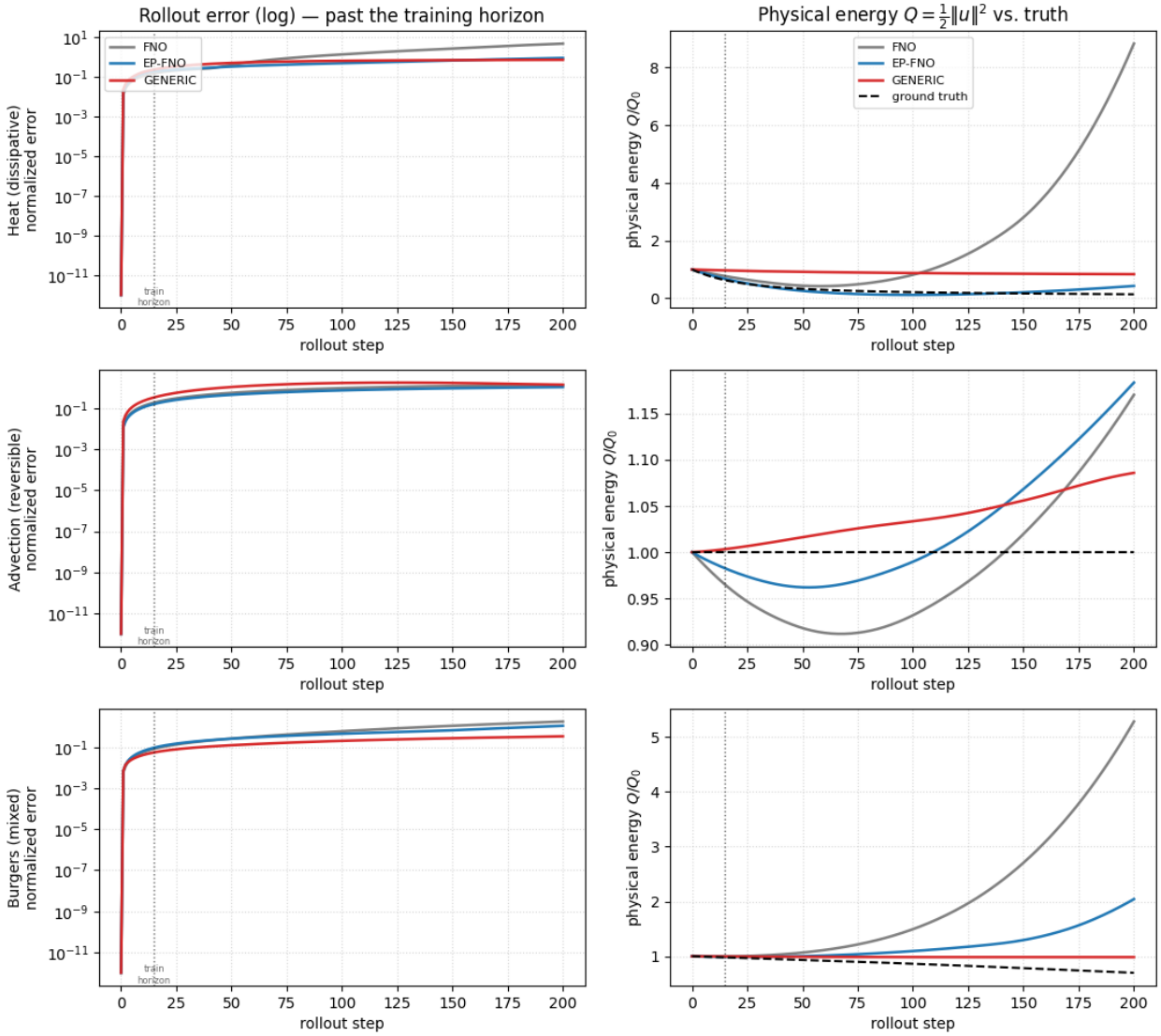

Figure 5: **Long-horizon rollouts** (200 steps; dotted line marks the 15-step training horizon). *Left:* relative error vs. step (log)—the unconstrained and penalized baselines diverge on heat and Burgers while GENERIC-FNO stays bounded. *Right:* physical energy $Q = \frac{1}{2}\|u\|^2$ relative to its initial value, against the ground truth (dashed). FNO's energy blows up on heat ($\approx 9\times$) and Burgers ($\approx 5\times$); GENERIC-FNO stays bounded on all three and holds the constant true energy most faithfully on advection. On heat GENERIC-FNO under-dissipates—its energy plateaus above the true decay—so it is bounded but does not reach the fully-relaxed limit.

RK4 mixes δE across four stages and raises it to $r_E \sim 10^{-3}$, trading away the machine-precision guarantee that is the point of the construction. We therefore treat coarse-grid reversible transport as a known limitation of the explicit scheme, with a documented—but structurally costlier—remedy. The degeneracy identities themselves hold at every stage of *any* explicit scheme (Proposition 2); what standard RK4 forfeits is only the discrete energy identity, so higher-order integrators that restore it—e.g. discrete-gradient or projection-based schemes—are a concrete route to both accuracy and the exact discrete guarantee.

Operator expressivity. Separately from the integrator, the operators are deliberately simple—diagonal Fourier multipliers on a low-frequency band, sandwiched by rank-one projections—which caps how expressive the reversible part can be on pure transport. Crucially, the exact degeneracy does not depend on this diagonal choice: the projection sandwich preserves skew-adjointness and positive semi-definiteness and yields $L\,\delta S/\delta u = M\,\delta E/\delta u = 0$ for *any* bounded D (Lemma 2, Proposition 2), so enriching the operators puts only expressivity, not exactness, at stake. Richer structure-preserving parameterizations of L and M (e.g. multi-rank or block-banded multipliers that remain skew-adjoint and PSD under the same projection sandwich) are a natural route to recovering transport expressiveness without giving up the exact degeneracy, and we leave them to future work.

Table 6: **Advection under explicit Euler vs. RK4 + anti-aliasing** (rollout L^2, trained at 64). RK4 roughly halves the coarse-grid blow-up but raises the per-step energy residual r_E from $\sim 10^{-7}$ to $\sim 10^{-3}$. We keep Euler so the discrete update conserves energy to machine precision.

	FNO	GENERIC (Euler)		GENERIC (RK4 + AA)	
res	L^2	L^2	r_E	L^2	r_E
64 (train)	0.174	0.606	2×10^{-8}	0.341	2.9×10^{-3}
128	0.134	0.154	9×10^{-8}	0.153	1.2×10^{-3}
256	0.117	0.161	1×10^{-7}	0.159	8×10^{-5}

Gains depend on backbone capacity and dimension. The benefit of the structural prior is largest for the higher-capacity 2D FNO and DeepONet backbones and smallest—occasionally negative—for the low-capacity 1D FNO, where GENERIC wins Burgers but loses heat and advection (Appendix B). The constraint is not a free lunch at every scale; it helps most when the backbone has enough capacity to fit the data and the prior can act as a regularizer rather than a bottleneck.

Partial observation. GENERIC-FNO models a first-order, Markovian map in the observed field. Systems that are not Markovian in the observed state—most directly, second-order dynamics such as the wave equation when only the displacement, not the velocity, is observed—are only approximately representable. This is benign in 2D but defeats every model, ours and the unconstrained baseline alike, at the coarse 1D grid (§5.2). Lifting to an augmented observed state (e.g. $(u, \partial_t u)$) is a natural extension we do not pursue here.

Parameterization cost on non-spectral backbones. For the FNO backbones the energy and entropy functionals add little overhead, and GENERIC-FNO in fact uses fewer parameters than the baselines. For DeepONet, however, the second functional branch makes GENERIC-DeepONet larger than the unconstrained model (175K vs. 58K). Its accuracy gains therefore cannot be attributed to structure on parameter count alone without a control; we supply that control in Appendix C, a parameter-matched ablation where GENERIC-DeepONet wins at every PDE and both budgets, and corroborate it with the FNO comparison, where GENERIC wins at fewer parameters.

Gauge-dependence of interpretation. As established in §4.4, the learned (E, S, L, M) decomposition is not unique. Consequently we make no claim about the absolute learned energy or entropy, nor about per-channel quantities such as the dissipative fraction ρ_M, which vary with the gauge. Our thermodynamic claims are restricted to gauge-invariant quantities: the exact structural guarantees and the dissipation diagnostic r_{mech}, for which we report only the sign (reversible vs. dissipative) and the across-PDE *ordering* relative to ground truth—not a calibrated dissipation rate.

Scope of the study. We study scalar fields on periodic domains for a small set of canonical PDEs spanning the reversible–irreversible spectrum. Vector and multi-field systems, non-periodic boundary conditions, and the complex-fluid and coarse-graining applications that motivate GENERIC in practice are left to future work; the construction is agnostic to these, but we have not validated it there. We emphasize that these

canonical PDEs are a deliberate controlled testbed rather than the target application: each has a known ground-truth dissipation rate, which is exactly what makes the falsifiable reversibility test of §5.4 and the exact super-resolution check of §5.3 possible. The complex-fluid and coarse-graining problems that motivate GENERIC are the intended use; demonstrating the construction there—at scale and on vector or multi-field states—is the main direction we leave open.

7 Conclusion

We introduced GENERIC-FNO, the first neural operator to embed the full GENERIC (metriplectic) structure of nonequilibrium thermodynamics in function space. By learning the energy and entropy functionals as neural operators and parameterizing the Poisson and friction operators as diagonal Fourier multipliers sandwiched between rank-one projections, the model enforces the degeneracy conditions—and hence the first and second laws—*exactly, by construction*, with no penalty, no projection of the predicted update, and no free residual. The structural (degeneracy) identities hold to machine precision for any initialization, spatial dimension, or resolution, and transfer zero-shot across a $4\times$ super-resolution range. We also identified the gauge freedom of the learned (E, S, L, M) decomposition, separated the gauge-invariant claims from the gauge-dependent ones, and gave a falsifiable, gauge-invariant dissipation diagnostic that recovers the correct ordering of physical dissipation and detects reversibility without reference to the learned functionals. Across three operator backbones and four PDEs, the structural prior is competitive with strong baselines and improves accuracy on several dissipative and mixed problems at comparable or fewer parameters.

The method also has clear boundaries, which we have tried to state plainly: the constrained reversible operator is less expressive than a free operator on pure linear transport in the FNO backbone; the explicit integrator we retain (for the machine-precision discrete guarantee) limits coarse-grid accuracy on reversible transport, a limitation we diagnose and mitigate but do not eliminate; and the first-order Markovian form does not natively represent second-order systems under partial observation. Each of these points to concrete future work—augmented observed states for second-order dynamics, structure-preserving higher-order integrators that retain exact degeneracy, and vector- and multi-field systems on non-periodic domains. The most compelling direction, however, is the one that motivated the formalism in the first place: applying exactly thermodynamically consistent neural operators to the high-dimensional closure and complex-fluid problems where staying on the physical manifold, far from the training data, is the difference between a stable surrogate and a useless one.

Broader Impact

GENERIC-FNO enforces exact energy conservation and non-negative entropy production for a *learned* energy and entropy. This is a structural guarantee about the model's own functionals, not a certificate of physical correctness: under the gauge non-identifiability of §4.4, and outside the scalar, periodic regimes tested here, matching the true governing physics is not implied by the guarantee. The intended positive impact is improved stability and physical plausibility of learned surrogates for scientific and engineering simulation. The corresponding risk is that exact conservation of a learned quantity is mistaken for validated physical accuracy in safety-critical simulation or engineering decision-making. We therefore stress that deployment on real systems requires additional validation against the target governing equations, boundary conditions, and operating regimes; the guarantees we prove are necessary structural conditions, not a substitute for that validation.

References

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.

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. arXiv:2003.04630.

Karthik Duraisamy, Gianluca Iaccarino, and Heng Xiao. Turbulence modeling in the age of data. *Annual Review of Fluid Mechanics*, 51:357–377, 2019. doi: 10.1146/annurev-fluid-010518-040547.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. arXiv:1906.01563.

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. arXiv:2305.15616.

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

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.

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.

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. arXiv:2106.12619.

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. arXiv:2010.08895.

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.

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.

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.

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.

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 Mathematical Background and Proofs

This appendix makes the claims of §3–§4.4 precise and self-contained. Throughout, all operators are understood pointwise in the state u: $L = L(u)$ and $M = M(u)$ depend on u through the gradients $\delta E/\delta u$ and $\delta S/\delta u$, and every statement below is asserted at a fixed but arbitrary u.

A.1 Function-space setting

Let $\Omega = \mathbb{T}^d$ be the periodic torus and $\mathcal{H} = L^2(\Omega;\mathbb{R})$ the real Hilbert space with inner product $\langle f, g\rangle = \int_\Omega f(x)g(x)\,dx$ and norm $\|f\| = \langle f, f\rangle^{1/2}$. For a Fréchet-differentiable functional $F : \mathcal{H} \to \mathbb{R}$, the variational derivative $\delta F/\delta u \in \mathcal{H}$ is the Riesz representative of the differential, $\langle \delta F/\delta u,\, \eta\rangle = \lim_{\epsilon\to 0} \epsilon^{-1}\big(F[u+\epsilon\eta]-F[u]\big)$ for all $\eta \in \mathcal{H}$.

On an N-point grid, $\mathcal{H}$ is replaced by $\mathbb{R}^N$ with $\langle f, g\rangle = \Delta x \sum_i f_i g_i$. Our functionals are mean-pooled, $F[u] = h(\frac{1}{N}\sum_i g(u)_i)$, so the Euclidean gradient returned by automatic differentiation, $\nabla_u F$, equals $\frac{\Delta x}{N}$ times the discrete Riesz representative of $\delta F/\delta u$. Every property used below (orthogonal projection, skew-adjointness, positive semi-definiteness, the degeneracy identities) is invariant to a positive scalar multiple of $\delta E/\delta u$ and $\delta S/\delta u$, so this constant is immaterial and we write $\delta F/\delta u$ for the autodiff gradient without further comment.

A.2 Rank-one projections

For $w \in \mathcal{H} \setminus \{0\}$ define $P_w v = \frac{\langle w,v\rangle}{\langle w,w\rangle}\, w$.

Lemma 1. *P_w is the orthogonal projection onto* $\mathrm{span}\{w\}$*: it is self-adjoint ($P_w^* = P_w$) and idempotent ($P_w^2 = P_w$), $P_w w = w$, and $(I - P_w)w = 0$. Consequently $I - P_w$ is the self-adjoint, idempotent orthogonal projection onto $w^\perp$.*

Proof. For any u, v, $\langle P_w u, v\rangle = \frac{\langle w,u\rangle}{\langle w,w\rangle}\langle w, v\rangle = \langle u, P_w v\rangle$, so $P_w^* = P_w$. $P_w^2 v = \frac{\langle w,v\rangle}{\langle w,w\rangle} P_w w = \frac{\langle w,v\rangle}{\langle w,w\rangle} w = P_w v$ since $P_w w = w$. Finally $(I - P_w)w = w - w = 0$. Self-adjointness and idempotence of $I - P_w$ are immediate, and its range is $w^\perp$ because P_w has range $\mathrm{span}\{w\}$. □

A.3 The base Fourier multipliers

Let $\widehat{(\cdot)}$ denote the Fourier transform on $\mathbb{T}^d$, so that $\langle f, g\rangle = \sum_k \overline{\widehat{f}_k}\, \widehat{g}_k$ (Parseval), and recall that f is real iff $\widehat{f}_{-k} = \overline{\widehat{f}_k}$. Let D_L and D_M be the diagonal multipliers with symbols $\widehat{D_L}(k) = \mathrm{i}\, a(k)$ and $\widehat{D_M}(k) = |b(k)|^2$, where $a(k) \in \mathbb{R}$ is odd ($a(-k) = -a(k)$) and $|b(k)|^2 \geq 0$ is even.

Proposition 1. *D_L maps real fields to real fields and is skew-adjoint, $\langle f, D_L g\rangle = -\langle D_L f, g\rangle$. D_M maps real fields to real fields and is self-adjoint and positive semi-definite, $\langle f, D_M f\rangle \geq 0$.*

Proof. Reality of $D_L f$ follows from $\widehat{D_L f}_{-k} = \mathrm{i}\, a(-k)\widehat{f}_{-k} = -\mathrm{i}\, a(k)\overline{\widehat{f}_k} = \overline{\mathrm{i}\, a(k)\widehat{f}_k} = \overline{\widehat{D_L f}_k}$, using that a is odd. By Parseval,

$$\langle f, D_L g\rangle = \sum_k \overline{\widehat{f}_k}\, \mathrm{i}\, a(k)\widehat{g}_k, \qquad \langle D_L f, g\rangle = \sum_k \overline{\mathrm{i}\, a(k)\widehat{f}_k}\, \widehat{g}_k = -\sum_k \mathrm{i}\, a(k)\overline{\widehat{f}_k}\, \widehat{g}_k,$$

so $\langle f, D_L g\rangle = -\langle D_L f, g\rangle$. For D_M, evenness of $|b(k)|^2$ gives reality by the same computation, self-adjointness because the symbol is real, and $\langle f, D_M f\rangle = \sum_k |b(k)|^2\, |\widehat{f}_k|^2 \geq 0$. □

In the implementation the inverse real FFT enforces the Hermitian symmetry that makes $a(k)$ effectively odd and $|b(k)|^2$ even, so these properties hold for *any* values of the learnable parameters.

A.4 Structure preservation under the projection sandwich

Lemma 2. *Let P be a self-adjoint projection and D a bounded operator. Then: (i) if D is skew-adjoint, $(I-P)D(I-P)$ is skew-adjoint; (ii) if D is self-adjoint and positive semi-definite, $(I-P)D(I-P)$ is self-adjoint and positive semi-definite.*

Proof. Write $Q = I - P$, which is self-adjoint ($Q^* = Q$) by Lemma 1. Then $(QDQ)^* = Q^*D^*Q^* = QD^*Q$. (i) If $D^* = -D$ then $(QDQ)^* = -QDQ$. (ii) If $D^* = D$ then $(QDQ)^* = QDQ$, and for any v, $\langle v, QDQv\rangle = \langle Qv, D\,Qv\rangle \geq 0$ since $D \succeq 0$. □

Corollary 1. *With $L = (I-P_S)D_L(I-P_S)$ and $M = (I-P_E)D_M(I-P_E)$, the operator L is skew-adjoint and M is self-adjoint positive semi-definite.*

Proof. Combine Proposition 1 with Lemma 2, taking $P = P_S$ for L and $P = P_E$ for M. □

A.5 Exact degeneracy

Proposition 2. $L\,\dfrac{\delta S}{\delta u} = 0$ *and* $M\,\dfrac{\delta E}{\delta u} = 0$ *hold identically, for any parameters and any state u.*

Proof. By Lemma 1, $(I-P_S)\,\delta S/\delta u = 0$. Hence $L\,\delta S/\delta u = (I-P_S)D_L\big[(I-P_S)\,\delta S/\delta u\big] = (I-P_S)D_L\,0 = 0$. Symmetrically $(I-P_E)\,\delta E/\delta u = 0$ gives $M\,\delta E/\delta u = 0$. □

A.6 The first and second laws

Theorem 1. *Let $\partial_t u = L\,\delta E/\delta u + M\,\delta S/\delta u$ with L, M as above. Then the semi-discrete (continuous-time) dynamics conserve energy and produce entropy:*

$$\frac{dE}{dt} = 0, \qquad \frac{dS}{dt} = \left\langle \tfrac{\delta S}{\delta u},\, M\,\tfrac{\delta S}{\delta u}\right\rangle \geq 0.$$

Proof. By the chain rule $dE/dt = \langle \delta E/\delta u,\, \partial_t u\rangle$. Expanding,

$$\frac{dE}{dt} = \left\langle \tfrac{\delta E}{\delta u}, L\,\tfrac{\delta E}{\delta u}\right\rangle + \left\langle \tfrac{\delta E}{\delta u}, M\,\tfrac{\delta S}{\delta u}\right\rangle.$$

The first term vanishes because L is skew-adjoint (Corollary 1): $\langle v, Lv\rangle = -\langle Lv, v\rangle = -\langle v, Lv\rangle$ forces $\langle v, Lv\rangle = 0$. The second vanishes because $M\,\delta E/\delta u = 0$ (Proposition 2) and M is self-adjoint, so $\langle \delta E/\delta u, M\,\delta S/\delta u\rangle = \langle M\,\delta E/\delta u,\, \delta S/\delta u\rangle = 0$. Likewise

$$\frac{dS}{dt} = \left\langle \tfrac{\delta S}{\delta u}, L\,\tfrac{\delta E}{\delta u}\right\rangle + \left\langle \tfrac{\delta S}{\delta u}, M\,\tfrac{\delta S}{\delta u}\right\rangle;$$

the first term equals $-\langle L\,\delta S/\delta u,\, \delta E/\delta u\rangle = 0$ by skew-adjointness and Proposition 2, and the second is ≥ 0 by positive semi-definiteness of M (Corollary 1). □

Proposition 3 (Discrete energy drift)**.** *Let $u_{t+1} = u_t + \Delta t\, f(u_t)$ with $f = L\,\delta E/\delta u + M\,\delta S/\delta u$ the explicit Euler update, and suppose E is C^2 with Hessian bounded by β on the trajectory. Then $\big|E[u_{t+1}] - E[u_t]\big| \leq \frac{1}{2}\beta\,\Delta t^2\,\|f(u_t)\|^2 = \mathcal{O}(\Delta t^2)$.*

Proof. By Taylor's theorem, $E[u_t + \Delta t f] = E[u_t] + \Delta t\langle \delta E/\delta u, f\rangle + R$ with $|R| \leq \frac{1}{2}\beta\Delta t^2\|f\|^2$. The first-order term $\langle \delta E/\delta u, f\rangle = dE/dt$ vanishes by Theorem 1, leaving $|E[u_{t+1}] - E[u_t]| = |R|$. □

Thus energy conservation is *exact in continuous time* (the first-order term is identically zero, verified numerically to $\sim 10^{-13}$ in §4.3); the explicit update incurs only the standard $\mathcal{O}(\Delta t^2)$ integrator error, which a higher-order integrator reduces at the cost discussed in §6.

A.7 Gauge freedom and gauge-invariance of the diagnostics

Definition 1. Two tuples (E, S, L, M) and (E', S', L', M') are *gauge-equivalent* on a domain $\mathcal{U} \subseteq \mathcal{H}$ if they generate the same flow,

$$L(u)\,\frac{\delta E}{\delta u} + M(u)\,\frac{\delta S}{\delta u} = L'(u)\,\frac{\delta E'}{\delta u} + M'(u)\,\frac{\delta S'}{\delta u} =: f(u) \qquad \text{for all } u \in \mathcal{U}.$$

Proposition 4. *The dissipation diagnostic $r_{\mathrm{mech}}(u) = -\langle u, f(u)\rangle/(\|u\|\,\|f(u)\|)$ and the ground-truth rate $\Pi^\star$ are gauge-invariant: gauge-equivalent tuples yield identical values. Moreover, if the flow conserves the quadratic energy $Q[u] = \frac{1}{2}\|u\|^2$ (a reversible flow), then $r_{\mathrm{mech}}(u) = 0$ for all u.*

Proof. r_{mech} is a function of the pair $(u, f(u))$ alone; by Definition 1 gauge-equivalent tuples share the same f, hence the same r_{mech}. $\Pi^\star$ is computed from the reference trajectory and does not reference (E, S, L, M) at all. For the reversible case, $\frac{d}{dt}Q[u] = \langle u, \partial_t u\rangle = \langle u, f(u)\rangle$; if Q is conserved this is zero, so the numerator of r_{mech} vanishes. □

Remark 1. By contrast, the dissipative fraction $\rho_M = \|M\,\delta S/\delta u\|/(\|L\,\delta E/\delta u\| + \|M\,\delta S/\delta u\|)$ is *not* gauge-invariant: it depends on the individual generators $L\,\delta E/\delta u$ and $M\,\delta S/\delta u$, which gauge-equivalent tuples may apportion differently while preserving their sum f. This is why we report ρ_M only as an appendix mechanism diagnostic and base all thermodynamic claims on r_{mech}.

A.8 Resolution and dimension independence

Let $\Pi_N : \mathcal{H} \to V_N$ be the orthogonal projection onto the span of the first N Fourier modes (the grid of resolution N), and let the operator band m satisfy $2m \le N$. Write Φ for one GENERIC-FNO step.

Proposition 5. *For any band-limited field u with $\widehat{u}_k = 0$ for $|k| > K$, and any two resolutions N, N' with $2m \le N \le N'$ and $K \le m$, the step commutes with prolongation: $\Phi_{N'}(\iota_{N\to N'}u) = \iota_{N\to N'}\Phi_N(u)$, where $\iota_{N\to N'}$ is zero-padding in Fourier (band-limited interpolation).*

Proof sketch. The multipliers D_L, D_M act diagonally on modes $|k| \le m$ and as zero outside, identically at any $N \ge 2m$; zero-padding adds only modes $|k| > m$, which both Φ_N and $\Phi_{N'}$ leave untouched. The spectral-convolution and mean-pooling layers of the functional networks are likewise defined per-mode and by the resolution-independent spatial average, so $\delta E/\delta u$ and $\delta S/\delta u$ agree (after prolongation) across N, N'. Finally the projections P_E, P_S are ratios of inner products and are invariant to the common grid measure Δx. Hence every component of Φ commutes with $\iota_{N\to N'}$ on band-limited inputs. □

Proposition 5 is the formal content of the zero-shot super-resolution result (§5.3): a model trained at one resolution applies unchanged at any finer one, and the structural guarantees of Theorem 1 and Proposition 2, holding pointwise at every u, transfer with it. Pointwise nonlinearities in the functional networks are the only components that are not exactly band-limited; their aliasing is the source of the coarse-grid effect analyzed in §6.

B 1D Fourier Neural Operator Results

Table 7 reports the full 1D FNO study referenced in §5.2. The 1D FNO is the smallest-capacity backbone we study (GENERIC-FNO uses 47K parameters versus 71K for the baselines), and it is where the structural prior helps least: GENERIC wins Burgers (-37% rollout error) but loses heat and advection, consistent with the trend that the prior's benefit grows with backbone capacity and dimension (§6). The damped wave row substantiates the partial-observation claim of §5.2: at $n_x = 64$, with only the displacement u observed and not its velocity, every model—including the unconstrained FNO—collapses to a rollout error of ≈ 0.48, so wave is excluded from the main-text tables.

Table 7: **1D FNO, three-seed accuracy** (relative L^2, mean ± std). GENERIC-FNO: 47K parameters; FNO and EP-FNO: 71K. Bold = best per PDE on rollout. Wave entries are the partial-observation collapse (all models ≈ 0.48).

PDE	Model	L^2 (1-step)	L^2 (rollout)
Heat	FNO	0.016 ± 0.004	0.082 ± 0.015
	EP-FNO	0.013 ± 0.002	$\mathbf{0.065 \pm 0.008}$
	GENERIC-FNO	0.029 ± 0.016	0.095 ± 0.012
Advection	FNO	0.026 ± 0.002	0.118 ± 0.004
	EP-FNO	0.022 ± 0.003	$\mathbf{0.114 \pm 0.014}$
	GENERIC-FNO	0.042 ± 0.003	0.233 ± 0.022
Burgers	FNO	0.015 ± 0.004	0.073 ± 0.010
	EP-FNO	0.010 ± 0.000	0.056 ± 0.003
	GENERIC-FNO	0.007 ± 0.001	$\mathbf{0.046 \pm 0.006}$
Wave	FNO	0.085 ± 0.008	0.484 ± 0.029
	EP-FNO	0.085 ± 0.007	0.480 ± 0.038
	GENERIC-FNO	0.078 ± 0.013	0.480 ± 0.034

C Parameter-Matched DeepONet Ablation

The DeepONet gains in §5.2 (Table 2) were obtained with GENERIC-DeepONet carrying its extra E, S branch, so it had more parameters than the baseline (175K vs. 58K). To rule out capacity as the explanation we retrain both models at *two matched budgets*, ≈ 58K and ≈ 175K, under one protocol—matching from both directions: GENERIC-DeepONet *shrunk down* to the baseline's budget, and the baseline *grown up* to GENERIC-DeepONet's budget. Counts are matched to within ∼2% (baseline 58,113 / 173,409; GENERIC 56,790 / 174,582). As Table 8 shows, GENERIC-DeepONet has lower 10-step rollout error at *every* PDE and *both* budgets (6/6 matched comparisons); shrinking it to 58K barely changes its error, while growing the baseline to 175K does not close the gap (and slightly worsens it on Burgers). The advantage is therefore structural, not a capacity artifact.

Table 8: **Parameter-matched DeepONet ablation** (10-step rollout L^2, single matched run per cell; lower is better, best per budget in bold). Both models are trained under the same protocol at two matched parameter budgets. GENERIC-DeepONet wins at every PDE and both budgets, and the baseline does not catch up when grown to 175K, so its advantage reflects the structure rather than the parameter count. The gaps far exceed the three-seed variation in Table 2.

	≈58K params		≈175K params	
PDE	DeepONet	GENERIC	DeepONet	GENERIC
Heat	0.055	**0.004**	0.052	**0.004**
Advection	0.209	**0.010**	0.190	**0.009**
Burgers	0.024	**0.009**	0.037	**0.015**

D Gauge-Dependent Mechanism Diagnostics

The main text bases all thermodynamic claims on the gauge-invariant rate r_{mech} (§4.4, Table 4). For completeness we describe here the two *mechanism* diagnostics that inspect the learned (E, S, L, M) decomposition directly, and we make explicit why they are reported only as supporting color: both are gauge-dependent.

The *dissipative fraction* measures how much of the raw update is routed through the irreversible channel,

$$\rho_M = \frac{\|M\,\delta S/\delta u\|}{\|L\,\delta E/\delta u\| + \|M\,\delta S/\delta u\|} \in [0,1], \tag{15}$$

and the *normalized learned-entropy production* is the cosine alignment of the entropy gradient with the dynamics,

$$r_S = \frac{\langle \delta S/\delta u,\, \partial_t u\rangle}{\|\delta S/\delta u\|\,\|\partial_t u\|}. \tag{16}$$

Both are scale-invariant (rescaling S leaves them fixed), which makes them tempting to report as physical quantities. They are not. As proved in Appendix A.7 (Remark following Proposition 4), ρ_M depends on the individual generators $L\,\delta E/\delta u$ and $M\,\delta S/\delta u$, which gauge-equivalent tuples apportion differently while preserving their sum; and r_S depends on the direction of $\delta S/\delta u$, which the gauge also moves. We observed this directly. For Burgers on the 2D FNO backbone, the dissipative fraction ρ_M swings from 0.002 in one seed to 1.000 in another—the model represents the *same* flow once as almost purely reversible and once as almost purely irreversible—while the rollout accuracy (≈ 0.026) and the gauge-invariant r_{mech} (≈ 0.01) are essentially unchanged across those seeds. The mechanism attribution is gauge; the dissipation of the physical energy is not. This is exactly the behavior Proposition 4 predicts, and it is why the falsifiable thermodynamic claims live in the main-text r_{mech} table rather than here.

Table 9: **Gauge-*dependent* mechanism diagnostics** ρ_M and r_S per PDE and backbone (three-seed mean $\pm$ std). These values shift under the entropy gauge and are reported only to illustrate the learned channel split; see the gauge-invariant r_{mech} in Table 4 for the load-bearing result. The large seed-to-seed variation is itself a symptom of the gauge freedom: for 2D-FNO Burgers, ρ_M ranges from 0.002 to 1.000 across seeds at essentially unchanged accuracy and r_{mech}.

Backbone	PDE	ρ_M	r_S
2D FNO	Heat	0.999 ± 0.000	0.51 ± 0.12
	Advection	0.000 ± 0.000	0.000 ± 0.000
	Burgers	0.67 ± 0.47	0.53 ± 0.38
1D FNO	Heat	0.54 ± 0.19	0.29 ± 0.09
	Advection	0.008 ± 0.004	0.006 ± 0.003
	Burgers	0.46 ± 0.33	0.35 ± 0.27
1D DeepONet	Heat	0.88 ± 0.06	0.70 ± 0.12
	Advection	0.031 ± 0.006	0.022 ± 0.009
	Burgers	0.53 ± 0.14	0.31 ± 0.05

E Training and Reproducibility Details

Data. Initial conditions are band-limited random fields with a fixed maximum wavenumber per PDE, integrated with a spectral solver to produce reference trajectories. The 2D experiments use 150 training samples with 15 time steps; the 1D experiments use 200 samples with 20 time steps. The headline 2D benchmark is trained and evaluated at 128×128; the super-resolution study (§5.3) trains at 64×64 and evaluates zero-shot up to 256×256 on data band-limited below the coarse-grid Nyquist frequency, so the continuous dynamics are identical across resolutions.

Models and training. All models are trained with Adam to minimize a relative L^2 objective combining the one-step and a short two-step rollout error; the GENERIC models receive *no* supervision on E or S and no degeneracy penalty (the conditions hold by construction). The 2D FNO and EP-FNO baselines have 2.10M parameters and GENERIC-FNO 1.00M; the 1D FNO baselines have 71K and GENERIC 47K; the DeepONet baseline has 58K and GENERIC-DeepONet 175K (the second functional branch). The energy and entropy functionals share the backbone family of their baseline and add a spatial-mean pooling followed

by a two-layer MLP head; L and M act on a fixed low-frequency band. We train 120 epochs (2D) and 150 epochs (1D).

Seeds and evaluation. Every reported number is the mean ± standard deviation over three seeds, with both the sampled data and the model initialization resampled per seed. Checkpoints are saved on the first seed. Accuracy is the relative L^2 error of a 10-step autoregressive rollout on held-out trajectories; structural residuals (r_E, degeneracy) and the gauge-invariant r_{mech} are computed on held-out initial conditions as defined in §4.4. With three seeds the reported standard deviations are indicative rather than precise variance estimates; we lean on them only where the effect is large relative to the spread (e.g. the order-of-magnitude difference in seed variance on 2D heat, Table 1), and the headline accuracy and long-horizon gaps far exceed it.

Computational cost. The overhead beyond a plain spectral operator is the reverse-mode autodiff through the learned functionals: each step computes $\delta E/\delta u$ and $\delta S/\delta u$ as two backward passes, and this is intrinsic to the method rather than an implementation detail. At the 128^2 config (Table 10; batch 8, single GPU) the deployed forward costs $\approx 4.4\times$ a plain FNO forward, a training step $\approx 10\times$ (the loss backpropagates through that graph, making the step a double-backward, which dominates and does not parallelize away on GPU where a plain FNO step is only $\approx 10\,\text{ms}$), and peak training memory $\approx 2.3\times$—all at half the parameters (1.00M vs. 2.10M). We report this plainly as the price of exact variational derivatives; disabling `create_graph` at inference saves only $\approx 7\%$, since the cost is the functional gradients themselves, not graph retention. Two factors bound the impact in practice: training is one-time and the absolute step cost is small ($\approx 0.1\,\text{s}$), and the projection machinery itself is $O(N)$ and negligible beside the FFTs (§4.5). The functional-gradient cost scales with the number of fields; cheaper E, S gradients (shared backbones, a fused double-backward) are natural future work.

Table 10: **Computational cost** at the 2D 128×128 config (batch 8, single GPU; latencies averaged over 30 iterations after warmup). "Forward" is the deployed inference latency (`create_graph=False`); the training step is a double-backward through the learned functionals. GENERIC-FNO's overhead is real (forward $\approx 4.4\times$, step $\approx 10\times$, peak memory $\approx 2.3\times$ a plain FNO), though it uses half the parameters and the absolute step cost is small ($\approx 0.1\,\text{s}$).

Model	Params	Forward (ms)	Train step (ms)	Peak mem (MB)
FNO	2,102,529	3.0	9.6	358
EP-FNO	2,102,529	2.8	9.8	392
GENERIC-FNO	1,001,958	13.0	98.6	809

Code. Exact architectures, wavenumber bands, learning rates, and batch sizes are fixed in the accompanying code, available as an anonymized repository at `https://anonymous.4open.science/r/GENERIC-FNO-F414/`, which reproduces every table and figure—including the three-seed wrappers and the super-resolution, long-horizon, and Euler-versus-RK4 ablations. A de-anonymized release will follow upon acceptance.

F Supplementary Figures

This appendix collects five figures that visualize, rather than re-tabulate, the exactness, gauge, ordering, integrator, and learned-functional results. All are generated by the released script from the same three-seed runs.

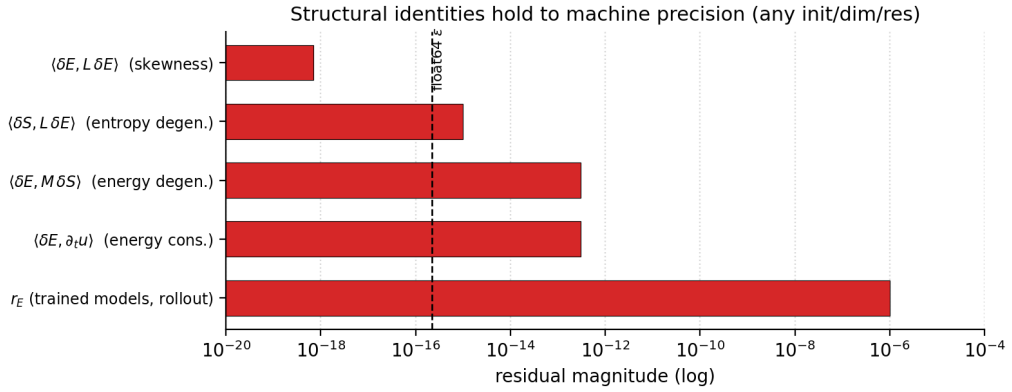

Figure 6: **The structural identities hold to machine precision.** Magnitude (log scale) of the GENERIC residuals at random initialization on a 16×16 grid: reversible skewness $\langle \delta E, L\,\delta E \rangle$, the two degeneracy conditions $\langle \delta S, L\,\delta E \rangle$ and $\langle \delta E, M\,\delta S \rangle$, energy conservation $\langle \delta E, \partial_t u \rangle$, and the trained-model energy residual r_E. All sit at or near the float64 machine epsilon (dashed line), confirming that degeneracy is enforced by construction rather than approximately (§4.3, Appendix A). Only r_E, which also reflects the explicit-Euler step, rises to $\sim 10^{-6}$.

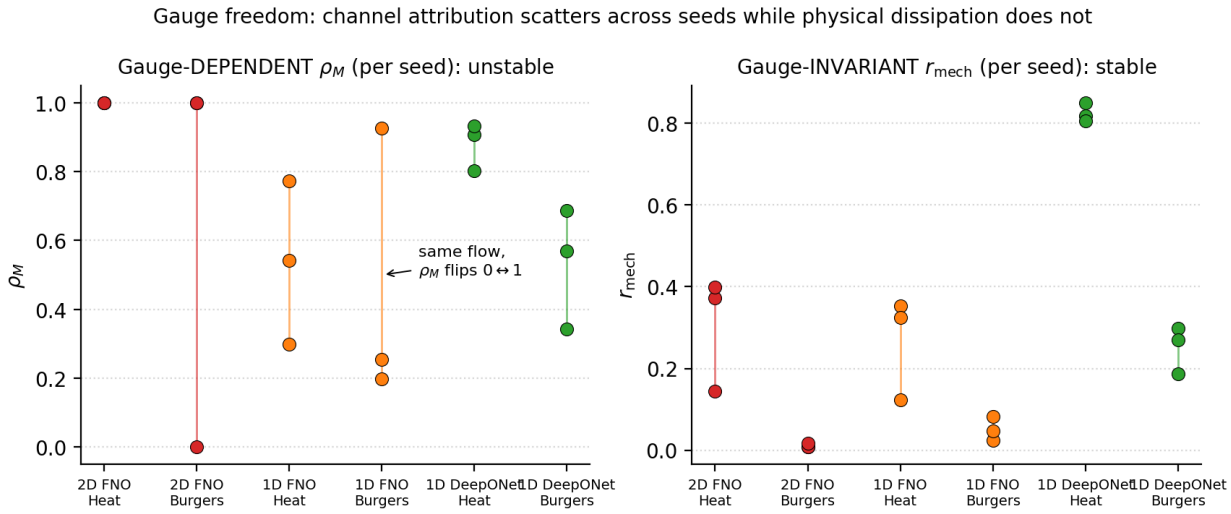

Figure 7: **Gauge freedom, made visual.** Per-seed values of the gauge-*dependent* dissipative fraction ρ_M (left) and the gauge-*invariant* dissipation rate r_{mech} (right), for the non-reversible PDEs across all three backbones. The channel attribution ρ_M scatters wildly across seeds—for 2D-FNO Burgers it spans the full range from 0.002 to 1.000, i.e. the model represents the *same* flow once as almost purely reversible and once as almost purely irreversible—whereas r_{mech}, computed from the fixed physical energy, stays stable for that same flow. This is the empirical signature of the non-uniqueness proved in Appendix A.7, and the reason thermodynamic claims rest on r_{mech} (§4.4, Appendix D).

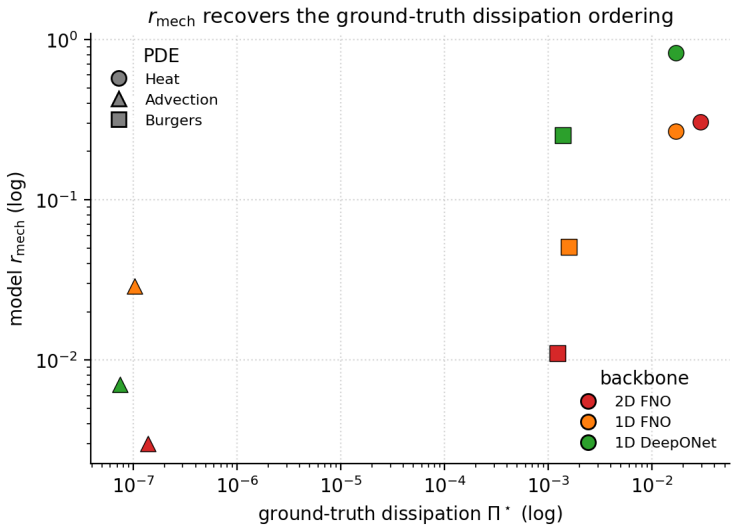

Figure 8: r_{mech} **recovers the ground-truth dissipation ordering.** Each point is one (backbone, PDE) pair: the model's gauge-invariant rate r_{mech} against the ground-truth rate $\Pi^\star$ (both log scale; markers denote PDE, colors denote backbone). The reversible advection points cluster at the bottom-left (≈ 0) and the dissipative/mixed points lie up-and-to-the-right. Within each backbone the three PDEs are ranked exactly as $\Pi^\star$ ranks them (Spearman $\rho = 1$ per backbone; $\rho \approx 0.87$ pooled over all nine points), so the model orders physical dissipation correctly without ever being supervised on energy or entropy (§5.4). The comparison is ordinal: r_{mech} and $\Pi^\star$ use different normalizations (§5.4), so only the ranking—not the absolute offset between the axes—is meaningful.

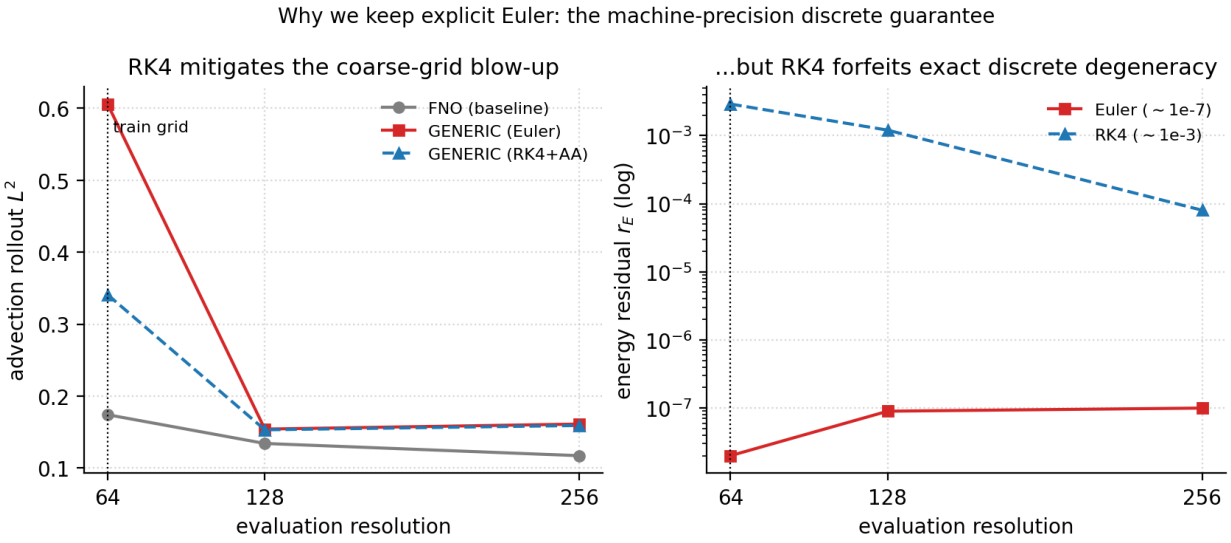

Figure 9: **Why we keep explicit Euler.** Advection trained at 64 and evaluated to 256. *Left:* rollout L^2 for the FNO baseline, GENERIC with explicit Euler, and GENERIC with RK4 + anti-aliasing. RK4 roughly halves the coarse-grid blow-up (Euler 0.61 $\rightarrow$ RK4 0.34 at the train grid) and the two integrators agree at finer grids. *Right:* the per-step energy residual r_E, which RK4 raises from $\sim 10^{-7}$ (Euler, exact discrete degeneracy) to $\sim 10^{-3}$. We retain Euler so the discrete update conserves energy to machine precision (§6).

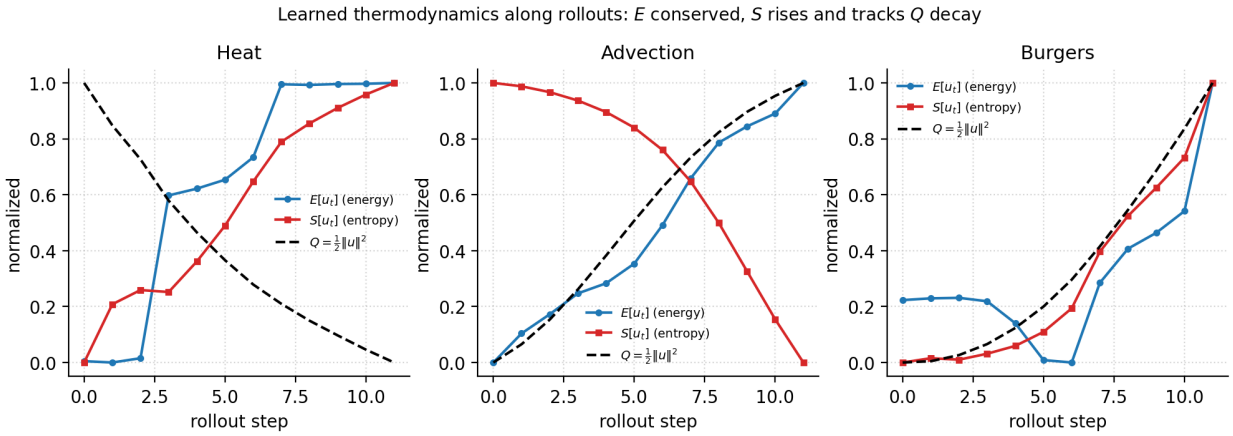

Figure 10: **Learned thermodynamics along a rollout (illustrative, single realization).** For one trained model per PDE we plot the learned energy $E_\theta[u_t]$, the learned entropy $S_\phi[u_t]$, and the fixed mechanical energy $Q = \frac{1}{2}\|u_t\|^2$ along a trajectory, each affinely rescaled to $[0, 1]$ within its panel for display. In this realization E_θ stays flat (conserved) while S_ϕ rises monotonically and tracks the decay of Q for the dissipative and mixed PDEs, and both are flat for reversible advection—so the unsupervised functionals behave like a physical energy and entropy. We show this only as an interpretive aid: the absolute values, scale, and per-channel attribution of E_θ, S_ϕ are gauge-dependent (§4.4), so the figure is *not* a gauge-invariant claim. The falsifiable, gauge-invariant statements are the dissipation diagnostics of Table 4 (§5.4).