Beyond Unit-Circle Eigenvalues: Invariant Bases for Stable State Space Dynamics

Jiaqian Zhu[1,†] **Yang Zhang**[2,*] **Junhua Ding**[3] **Xiaowei Yu**[1,*]
[1]Missouri University of Science and Technology [2]University of North Texas [3]University of Wyoming
jzbmn@mst.edu, xych8@mst.edu, Yang.Zhang@unt.edu, Junhua.Ding@uwyo.edu

Abstract

Diagonal state space models process long sequences at linear cost, yet the same transition is applied at every step of a rollout, so an unconstrained recurrence holds no quantity fixed and the state decays or grows without bound. Current designs constrain one quantity, and preserving one does not guarantee preservation of the others, so a model can hold its norm to machine precision while momentum drifts by hundreds of percent. We claim that the conserved quantities of a diagonal recurrence are layered, that each layer requires a strictly stronger condition on the operator, and that a unit-modulus spectrum reaches only the outermost. Acting on the claim, we enforce the conditions required for every layer by construction. The generator of the Kronecker–Cayley transition (KC) is a graph Laplacian over object subspaces, $\mathbf{H} = \mathbf{L} \otimes I_m$, advanced by a Cayley transform, and the Kronecker structure passes through the transform, $\mathrm{Cay}(\mathbf{L} \otimes I_m) = \mathrm{Cay}(\mathbf{L}) \otimes I_m$, so the step stays diagonal at $O(N^3+N^2m)$ cost. The conserved set is exactly a $2m^2$-dimensional family of quadratic forms, held at every parameter value, so the quantities and the dynamics the construction reaches are fixed before training. We evaluate KC on four conservation benchmarks and show that it holds total mass at or near the reference floor while every diagonal baseline loses it by orders of magnitude. We further compare against mass correction and find that two correctors fail once the underlying prediction diverges, which shows that assembly rather than correction is necessary for a guarantee that holds across the rollout. A free modal basis trained under a conservation penalty conserves only by freezing its spectrum, whereas the assembled operator conserves with every mode but one free to rotate. Beyond these benchmarks, KC holds the linear momentum of three N-body systems. At the operator level, the same layering holds the conserved quantity in molecular dynamics, chemical kinetics, pharmacokinetics and power-grid swing dynamics. Our code is available at `https://github.com/Jiaqian-Janelle/NeurIPS26-Kronecker-Cayley-Stable-State-Space-Dynamics`.

1 Introduction

Diagonal state space models have become a general parameterisation for long sequences, advancing a latent state by one diagonal transition per step at linear cost, which has carried them from language to audio, vision and physical dynamics [Gu et al., 2022a, Smith et al., 2023, Gu and Dao, 2024]. Yet the same transition is re-applied to every step of an autoregressive rollout, so that a deviation in one mode compounds rather than corrects and the latent norm decays towards zero or grows without bound. Two lines of work address this drift, one through a spectral or manifold constraint that places the eigenvalues on the unit circle [Orvieto et al., 2023, Arjovsky et al., 2016, Lezcano-Casado and Martínez-Rubio, 2019], and another through a structured generator that builds a prescribed quantity

*Corresponding authors. †Project leader.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

into a Hamiltonian [Greydanus et al., 2019, Chen et al., 2020] or into a constraint set solved at every step [Lee and Carlberg, 2021]. Both constrain the quantity they target, which leaves open the question we address: *which quantities can a diagonal recurrence preserve, which condition on the operator does each one require, and why does a conservation objective added to training not supply them?*

Three gaps separate that question from current practice. A transition with every eigenvalue on the unit circle holds the norm to machine precision, yet a summed quantity such as momentum drifts by 141 percent, because the norm is a property of the spectrum whereas a sum is preserved only when the uniform vector is an eigenvector, a property of the coordinates. *Stability and conservation are therefore distinct properties, and a constraint on the spectrum reaches only the first.* A conservation penalty is the usual response to the second, and it changes the objective without changing the structure, so it may not recover the physics even on modestly complex problems [Krishnapriyan et al., 2021]. An exact operator resolves that gap but leaves a third, since its range of applicability is unknown and a norm-preserving map degrades accuracy on dynamics that contract or expand [Vorontsov et al., 2017, Kerg et al., 2019]. The conditions that decide conservation must therefore be placed on the coordinates, encoded in the structure of the generator, and derived from a single algebraic condition that also yields the criterion for where they apply.

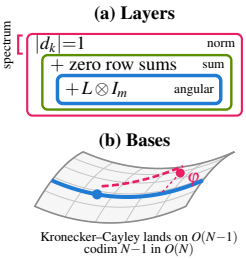


Figure 1: Conserved quantities are layered, and constraining the spectrum reaches only the outermost. **(a)** Each layer requires a strictly stronger condition. **(b)** The conserving pairs of basis and spectrum form a lower-dimensional set, and a basis trained under a penalty remains near its initialisation instead of moving to it.

Following this principle, we design the transition from the condition that decides what it conserves. For a transition obtained by the Cayley transform of a Hermitian generator, a quadratic form is conserved exactly when its generator commutes with $\mathbf{H}$, so the conserved set is fixed by the algebraic structure of $\mathbf{H}$ alone. Three requirements on that structure follow, each strictly stronger than the last: Hermiticity for the norm, vanishing row sums for the summed quantity, and the Kronecker form $\mathbf{L}\otimes I_m$ for the angular quantity, so the three are nested and a unit-modulus spectrum reaches only the outermost (Figure 1a). We therefore propose the Kronecker–Cayley transition (KC), which assembles $\mathbf{H}=\mathbf{L}\otimes I_m$ from one learned coupling per object pair and advances the latent by $\mathrm{Cay}(\mathbf{H})$, so that all three requirements are met at once. They are properties of the pattern rather than of the values placed in it, so they hold at every value of the coupling and no correction is required. The same equivalence supplies the missing criterion, since a quadratic physical quantity is held exactly when its generator lies in that family, and a unitary step cannot represent contraction. Holding the summed quantity requires every rotating mode to carry none of the uniform vector, which the assembly satisfies by placing that vector in the null mode and which a trained basis reaches only by freezing every mode (Figure 1b).

On four conservation benchmarks [Takamoto et al., 2022, Gupta and Brandstetter, 2023], KC holds total mass to between 4×10^{-6} and 9×10^{-3} in absolute units, at or near the reference floor, while four diagonal baselines under the same encoder lose between 4 and 4×10^{9}. Two mass correctors, one learned end-to-end [Liu et al., 2025] and one fixed and uniform, both fail once the underlying prediction has drifted too far, whereas KC reaches the final horizon on every seed. Removing one condition of the assembly costs exactly the quantity it holds, and a modal basis trained under a conservation penalty conserves only by freezing every mode. KC also holds the linear momentum of three N-body systems [Satorras et al., 2021, Brandstetter et al., 2022] to 2×10^{-6} against 0.5 to 8 for every baseline, and at the operator level the conserved quantity of molecular dynamics, chemical kinetics, pharmacokinetics and power-grid swing dynamics to within 10^{-2} percent, while on a dissipative system it fits less accurately than baselines that can contract, as the theory predicts.

Our contributions are as follows. (i) We prove that the conserved quantities of a diagonal recurrence are nested, each requiring a strictly stronger condition, and that a unit-modulus spectrum reaches only the outermost (§3.3). **(ii)** We propose the Kronecker–Cayley transition, which satisfies every condition at once, and prove that its conserved set is exactly a $2m^2$-dimensional family held at every parameter value (§3.2). **(iii)** We characterise the transitions that hold the summed quantity, show that a trained basis reaches them only by freezing its spectrum, and give a criterion, quadratic in the quantity and conservative in the dynamics, for where the construction applies (§3.4). **(iv)** We validate KC on four benchmarks, where it holds the conserved quantity at or near the floor while every baseline loses it and mass correction fails, and on dynamical systems across physics, chemistry and engineering (§4).

2 Related Work

Table 1: **What each design constrains, which layer it reaches, and what it leaves free.** The first three mark columns are the layers of Corollary 3, and a partial mark ✓ᵖ denotes a bounded norm, a quantity held only if named as a constraint, or one restored by a corrector while the underlying prediction stays within its reach.

Design	Designed object	Norm	Sum	Angular	Diag.	All θ	Leaves free
Spectrum: constrain the eigenvalues							
Diagonal SSM	eigenvalues	✗	✗	✗	✓	✗	eigenvectors
LRU	spectrum	✓ᵖ	✗	✗	✓	✓	eigenvectors
Unitary RNN	operator	✓	✗	✗	✗	✓	designated quantity
Generator: name the quantity							
Hamiltonian, SymODEN	scalar generator	✗	✓ᵖ	✓ᵖ	✗	✓	diagonal form
Deep conservation	constraint set	✓ᵖ	✓ᵖ	✓ᵖ	✗	✓	diagonal form
Operator learning: correct the output							
FNO	none	✗	✗	✗	✗	✗	conserved quantity
CPFNO	learned corrector	✗	✓ᵖ	✗	✗	✓ᵖ	conserved quantity
KC (ours)	generator, $\mathbf{L} \otimes I_m$	✓	✓	✓	✓	✓	coupling values

Stability and conservation in learned dynamics. Diagonal transitions have made long-sequence modelling practical [Gu et al., 2022b, Gupta et al., 2022, Gu et al., 2022a, Smith et al., 2023, Gu and Dao, 2024, Dao and Gu, 2024], and stability has been secured by restricting the operator to the unitary manifold [Arjovsky et al., 2016, Wisdom et al., 2016, Helfrich et al., 2018, Lezcano-Casado and Martínez-Rubio, 2019] or the eigenvalues to the unit disk [Orvieto et al., 2023]. A second line places the conserved quantity in the generator, as a constraint set [Lee and Carlberg, 2021] or a learned Hamiltonian [Greydanus et al., 2019, Cranmer et al., 2020, Chen et al., 2020, Zhong et al., 2020b]. What neither line constrains is the modal basis, which is inherited from the encoder in every design of Table 1, so a spectral constraint holds the norm alone, a named constraint holds only what is named, and the quantities that depend on the coordinates are left free. We fix the basis in which the spectrum acts, so that conservation follows from the assembly rather than from the values placed in it.

Graph Laplacians and conserved quantities. Zero row sums make the constant vector an eigenvector of a graph Laplacian at eigenvalue zero, so any sum over nodes is preserved by the associated flow. That identity is the object that carries the invariant in reaction network theory [Veerman et al., 2023] and in average consensus [Olfati-Saber et al., 2007]. Within machine learning, the same graph instead carries a learned message function [Battaglia et al., 2016, Sanchez-Gonzalez et al., 2020, Satorras et al., 2021, Kipf et al., 2018], which fixes the sparsity pattern of the dynamics but leaves the conserved quantity free, and the identity itself has not been carried into a recurrence. We assemble the Laplacian as the generator of a diagonal transition, so that the zero mode fixes which quantity the recurrence preserves at every parameter value.

Structure preservation under long integration. The loss of a conserved quantity under long integration is a recurring difficulty across scientific computing [Hairer et al., 2006], and each field has developed its own repair. In molecular dynamics, velocity rescaling drains kinetic energy into centre-of-mass motion until the internal degrees of freedom freeze [Harvey et al., 1998]. In stiff kinetics, a mass-conserving scheme leaves the positive orthant and is projected back onto the reaction simplex [Sandu, 2001]. In climate emulation, a water-budget corrector has been shown to hide a degrading model [Chapman et al., 2026], a failure mode we observe in §4. In each case, the violation is corrected after it occurs, because no condition on the operator excludes it in advance. We give that condition in §3, and extended related work is provided in Appendix B.

3 Method

3.1 Problem formulation

We study long-horizon rollout in a diagonal state space model, where the same transition is applied at every step and any quantity the transition does not preserve drifts with the horizon [Gu et al., 2022a, Smith et al., 2023]. By diagonal we mean that each step acts as a diagonal matrix in a modal basis, which for the construction below depends on the state and is recomputed at every step. Let an encoder map an observation x_t to a latent $z_t \in \mathbb{C}^d$, partitioned into N object subspaces $z = (z_1, \ldots, z_N)$ with $z_i \in \mathbb{C}^m$ and $d = Nm$. Given the latent and an action a_t, we seek a transition U that advances

the latent and holds a set $\mathcal{Q}$ of functions of z fixed,

$$z_{t+1} = U(z_t, a_t)\, z_t, \qquad Q\big(U(z,a)\, z\big) = Q(z) \quad \text{for all } Q \in \mathcal{Q},\ z,\ a, \text{ and all } \theta. \tag{1}$$

The transition should satisfy (i) exactness, so that Equation 1 holds at every parameter value θ rather than at a minimiser of a loss, and (ii) completeness, so that $\mathcal{Q}$ follows from the algebraic form of U rather than from a list named in advance. We follow three members of $\mathcal{Q}$ throughout,

$$\|z\|^2, \qquad \textstyle\sum_{i=1}^{N} z_i, \qquad \mathrm{Im}\big(z^\dagger \mathbf{J} z\big), \tag{2}$$

the latent norm, the sum over subspaces, and an angular pairing generated by $\mathbf{J} = \mathrm{diag}(\mathbf{J}_2, \ldots, \mathbf{J}_2)$, which arise from the unitary group, from translation $z_i \mapsto z_i + \epsilon$, and from rotation respectively. To meet (i) and (ii), we assemble U from a generator whose algebraic form fixes $\mathcal{Q}$ (§3.2), characterise $\mathcal{Q}$ exactly (§3.3), and state where the assembly applies (§3.4).

3.2 The Kronecker–Cayley assembly

Theorem 1 (Characterisation of the conserved set)**.** *Let* $\mathbf{H}$ *be Hermitian and* $U = \mathrm{Cay}(\mathbf{H})$*. For a Hermitian* G*, the quadratic form* $Q_G(z) = z^\dagger G z$ *satisfies*

$$Q_G(Uz) = Q_G(z) \ \textit{for every } z \quad \Longleftrightarrow \quad [\mathbf{H}, G] = 0. \tag{3}$$

By Theorem 1, a quantity is conserved exactly when its generator commutes with $\mathbf{H}$, so we make the transition conserve by assembling $\mathbf{H}$ rather than by fitting it, and the three quantities of Equation 2 translate into three requirements on $\mathbf{H}$, each stronger than the last. For the norm, it suffices that $\mathbf{H}$ be Hermitian, since I then commutes with $\mathbf{H}$. For the sum, $\mathbf{H}$ must in addition annihilate $\mathbf{1} \otimes e$ for every e, which puts the uniform direction at eigenvalue one under the transform. For the angular pairing, it suffices that every block of $\mathbf{H}$ be a scalar multiple of I_m, which makes every $I_N \otimes C$ commute with $\mathbf{H}$ and in particular $I_N \otimes \mathbf{J}$. Three stages meet these in turn, and only the first is learned.

Algorithm 1 summarises the rollout. Line 1 encodes and partitions the observation. Inside the loop, line 3 evaluates the couplings of Equation 4 and is the only line that depends on a parameter. Line 4 assembles them into $\mathbf{L}$ by Equation 5, line 5 applies the Cayley transform by Equation 6, and line 6 advances the latent, which Theorem 2 shows is diagonal in the modal coordinates of $\mathbf{L}$ and which §3.3 shows leaves every member of $\mathcal{Q}$ unchanged. The batched form is provided in Appendix C.10.

Algorithm 1 Kronecker–Cayley rollout

Require: x_0, $a_{0:T-1}$, encoder Φ, network MLP_θ

1: $z_0 \leftarrow \Phi(x_0)$, split as $(z_{0,1}, \ldots, z_{0,N})$
2: **for** $t = 0, \ldots, T-1$ **do**
3: $\quad w_{ij} \leftarrow \mathrm{MLP}_\theta(\|z_{t,i} - z_{t,j}\|^2, a_t)$ ▷ Eq. 4
4: $\quad \mathbf{L}_{ij} \leftarrow -w_{ij}, \mathbf{L}_{ii} \leftarrow \sum_{k \neq i} w_{ik}$ ▷ Eq. 5
5: $\quad \mathbf{V} \leftarrow \mathrm{Cay}(\mathbf{L})$ ▷ Eq. 6
6: $\quad z_{t+1} \leftarrow (\mathbf{V} \otimes I_m)\, z_t$ ▷ Thm. 2
7: **end for**
8: **return** $z_{1:T}$

Learned coupling for input dependence. A transition fixed in advance cannot represent dynamics that depend on the state and the action. We parameterise that dependence by one real scalar per pair of objects,

$$\underbrace{w_{ij} = \mathrm{MLP}_\theta\big(\|z_i - z_j\|^2,\ a_t\big)}_{\binom{N}{2} \text{ scalars, the only learned part}} \qquad \underbrace{w_{ij} = w_{ji}}_{\text{one value shared per pair}} \tag{4}$$

so the coupling varies with the input while its placement is fixed by the next two stages.

Laplacian assembly for vanishing row sums. A free Hermitian generator built from the same scalars would yield a unitary step, but its conserved set would move with θ. We place the couplings in a graph Laplacian over the N subspaces and tensor it with the identity,

$$\mathbf{L}_{ij} = \begin{cases} \underbrace{-w_{ij}}_{\text{learned}}, & i \neq j, \\ \underbrace{\textstyle\sum_{k \neq i} w_{ik}}_{\text{set by the row}}, & i = j, \end{cases} \qquad \text{giving} \qquad \underbrace{\mathbf{H} = \mathbf{L} \otimes I_m}_{\text{block-scalar, zero row sums}} \tag{5}$$

for a symmetric Laplacian $\mathbf{L} \in \mathbb{R}^{N \times N}$. Real symmetric entries make $\mathbf{H}$ Hermitian, each diagonal entry cancels its own row so that $\mathbf{L}\mathbf{1} = 0$, and every block is a multiple of I_m. The three requirements are met by the pattern in Equation 5 rather than by its values, so they hold at every θ.

$$\underbrace{U = \mathrm{Cay}(\mathbf{H}) = \big(I + \tfrac{\mathrm{i}}{2}\mathbf{H}\big)^{-1}\big(I - \tfrac{\mathrm{i}}{2}\mathbf{H}\big)}_{\text{unitary for every } \mathbf{H} = \mathbf{H}^\dagger} \qquad \underbrace{U = \mathbf{V} \otimes I_m}_{\mathbf{V} = \mathrm{Cay}(\mathbf{L}) \in \mathcal{U}(N)} \tag{6}$$

Theorem 2 (Kronecker–Cayley decomposition)**.** *Let $\mathbf{H} = \mathbf{L} \otimes I_m$ with $\mathbf{L} = P\Lambda P^\top$ a symmetric graph Laplacian, and let $\tilde{z} = (P^\top \otimes I_m)z$. Then*

$$\tilde{z}^{(k)}_{t+1} = d_k\, \tilde{z}^{(k)}_t, \qquad d_k = \frac{1 - \mathrm{i}\lambda_k/2}{1 + \mathrm{i}\lambda_k/2}, \qquad |d_k| = 1, \qquad d_{k^*} = 1. \tag{7}$$

Here $\mathrm{i} = \sqrt{-1}$. Equation 6 rotates the N objects into one another and leaves every coordinate channel untouched, and by Theorem 2 the rotation is diagonal in the eigenbasis of $\mathbf{L}$. Since $\mathbf{L}$ depends on z_t through the couplings, that eigenbasis changes from one step to the next, and the two modal transforms in Algorithm 1 are the cost of the change. A diagonal state space model with input-dependent entries, as in Mamba, instead keeps the basis fixed, $P_t = P$. The step costs one $N \times N$ eigendecomposition and two modal transforms, $O(N^3 + N^2 m)$ against $O(N^3 m^3)$, and the coupling network carries 3,963 parameters against roughly 15M in the shared encoder and decoder, so the guarantee is not obtained through capacity on the transition. The three stages impose three structural properties on $\mathbf{H}$ at every parameter value, and §3.3 derives the quantities these properties conserve and identifies the boundary of the conserved set.

3.3 The invariant family

We now derive what the three properties of §3.2 conserve. Two results follow from Theorem 1, the first arranging the three quantities of Equation 2 into nested layers by the condition each one requires, the second identifying the full conserved set and showing that nothing lies outside it.

Corollary 3 (Layered conservation)**.** *Let U be a unitary diagonal transition on N subspaces. Then*

$$\underbrace{|d_k| = 1}_{\|z\|^2} \;\Leftarrow\; \underbrace{\mathbf{L}\mathbf{1} = 0}_{+\,\sum_i z_i} \;\Leftarrow\; \underbrace{\mathbf{H} = \mathbf{L} \otimes I_m}_{+\,\mathrm{Im}(z^\dagger \mathbf{J} z)} \tag{8}$$

where each condition implies the one to its left and no converse holds. A transition satisfying only the leftmost condition preserves $\|z\|^2$ and leaves $\sum_i z_i$ unconstrained.

Corollary 3 is falsifiable, since removing one condition should cost exactly the quantity that condition holds, leaving the others intact, and §4 confirms this by removing each condition in turn.

Corollary 4 (The invariant family)**.** *Let $U = \mathrm{Cay}(\mathbf{L} \otimes I_m)$ with $\mathbf{L}$ ranging over all symmetric $N \times N$ matrices with zero row sums, $N \geq 2$. The quadratic forms held at every such $\mathbf{L}$ are exactly*

$$Q_{C_1,C_2}(z) = z^\dagger\big(I_N \otimes C_1 + \mathbf{1}\mathbf{1}^\top \otimes C_2\big)z, \qquad C_1, C_2 \in \mathbb{C}^{m\times m} \text{ Hermitian}, \tag{9}$$

a real vector space of dimension $2m^2$. No quadratic form outside Equation 9 is held at every $\mathbf{L}$.

Equation 9 is the set $\mathcal{Q}$ of Equation 1 and meets requirement (ii) of §3.1. It contains $\|z\|^2$ as $(C_1, C_2) = (I_m, 0)$ and $\mathrm{Im}(z^\dagger \mathbf{J} z)$ as $(\mathbf{J}, 0)$, and the summed quantity is linear rather than quadratic and is held through the kernel of $\mathbf{H}$. The assembled operator holds this family and nothing else at every parameter value. Which physical quantities lie inside the family, and on which dynamics the assembly applies, we address in §3.4.

3.4 Where the construction applies

Corollary 5 (Applicability)**.** *Let the decoder be linear on the latent, $y = Dz$. A physical quantity $q(y)$ that is a quadratic form in y is held by the construction for every θ if and only if there exist Hermitian C_1, C_2 with*

$$q(Dz) = z^\dagger\big(I_N \otimes C_1 + \mathbf{1}\mathbf{1}^\top \otimes C_2\big)z \quad \text{for every } z. \tag{10}$$

The characterisation does not determine whether a nonquadratic physical quantity is conserved.

Proposition 6 (Lower bound for any unitary step)**.** *For every unitary U and every pair of consecutive latents (z_t, z_{t+1}),*

$$\|Uz_t - z_{t+1}\| \;\geq\; \big|\,\|z_{t+1}\| - \|z_t\|\,\big|. \tag{11}$$

Corollary 5 settles which physical quantities are held before any model is trained. A linear momentum $\sum_i m_i v_i$ is linear in the state and is held through the kernel. An angular momentum $\sum_i r_i \times m_i v_i$ is bilinear in the state and therefore a quadratic form. A potential energy of the form $\sum_{i<j} \|r_i - r_j\|^{-1}$ is not a quadratic form under any linear encoding, so Corollary 5 does not apply to it, and §4 reports energy as held only where the potential is quadratic. Proposition 6 is the reverse triangle inequality with $\|Uz_t\| = \|z_t\|$. The bound depends only on the encoded trajectory, so no choice of $\mathbf{L}$, θ or decoder reduces it. A unitary step thus applies to dynamics whose encoded norm is constant along the true trajectory and incurs Equation 11 as a per-step error on dissipative dynamics. §4 tests the quantity boundary on three N-body systems and the dynamics boundary on a dissipative system.

Proposition 7 (Conservation of the summed quantity)**.** *A unitary diagonal transition on N subspaces is determined by a modal basis and a spectrum, $U = \big(Q\,\mathrm{diag}(d)Q^\top\big) \otimes I_m$ with $(Q, d) \in O(N) \times (S^1)^N$. Let $c = Q^\top \mathbf{1}/\sqrt{N}$. The transition preserves $\sum_i z_i$ for every z if and only if*

$$(d_k - 1)\, c_k = 0 \quad \textit{for every } k. \tag{12}$$

Every mode that rotates carries none of the uniform vector, and every mode that carries it is stationary.

We note that Equation 12 has two kinds of solutions. The assembly satisfies it with $c = e_{k^*}$ and $d_{k^*} = 1$, since $\mathbf{L}\mathbf{1} = 0$ puts $\mathbf{1}/\sqrt{N}$ in the null mode before any parameter is set, and leaves the remaining $N - 1$ modes free to rotate. A transition may instead satisfy it with every $d_k = 1$, at which point it no longer evolves. A basis initialised away from Equation 12 and trained under a conservation penalty must move c or d to one of these solutions, and we confirm in §4 that the seeds which conserve reach the second solution, freezing every mode, while the basis remains near its initialisation on every seed under both initialisations tested.

4 Experiments

We conduct experiments to answer five core research questions: **Q1.** How well does the Kronecker–Cayley transition hold a conserved quantity over a long rollout, against diagonal transitions that constrain only the spectrum? **Q2.** Can a corrector applied to the output restore what the operator itself does not hold? **Q3.** Is each condition of the assembly necessary, and does removing one cost exactly the quantity it holds? **Q4.** Can a conservation penalty bring a trained basis to the condition the assembly satisfies by construction? **Q5.** Does the same layering appear in dynamical systems from physics, chemistry and engineering, and where does the construction cease to apply?

4.1 Q1: Mass Is Held Where Every Diagonal Baseline Loses It

Setup. We test Corollary 3, which predicts that reaching only the outermost layer holds the norm and loses the summed quantity, on the dam-break and advection equations of PDEBench [Takamoto et al., 2022], the spherical shallow-water system of PDEArena [Gupta and Brandstetter, 2023] and a conservative Allen–Cahn flow. To isolate the transition, we fix the encoder and decoder of KC across every baseline, so that the levels reproduce the layers of the corollary. Level A leaves the spectrum free, level B imposes unit modulus, and level C fixes one mode per object at unity in the coordinate basis. To separate the operator from the read-out, we add an LRU-style recurrence [Orvieto et al., 2023] and KC with an unconstrained decoder, which breaks the tie required by Corollary 5, and the external references FNO [Li et al., 2021] and CPFNO [Liu et al., 2025]. We therefore measure conservation as the drift $|Q(T) - Q(1)|$ from the first decoded frame rather than from the input, which removes the encoder–decoder round trip, and prediction error as $\|\hat{y}_T - y_T\|/\|y_T\|$. Every run is repeated at seeds 42, 43 and 44, and datasets are given in Appendix A.

Results. Total mass is held by KC at the round-off level of the float32 rollout on the three flat benchmarks and within a discretisation residual of the reference on the sphere, whereas every diagonal baseline loses it by two to nine orders of magnitude: the drift of KC is 2.0×10^{-5} on Allen–Cahn, 2.6×10^{-3} on advection and 8.3×10^{-3} on the dam break, against 3 to 2×10^{11} for every other diagonal transition (Table 2). On the sphere the residual of 483 against 30.8 is the variation of $\cos(\mathrm{lat})$ within a patch (Appendix D.1). What remains is numerical rather than learned, since it accumulates smoothly with the rollout step and is largest on advection, whose horizon is twice that of the other benchmarks (Appendix D.1). The guarantee reaches the output only through the decoder: an unconstrained decoder holds the latent invariant just as tightly, between 10^{-6} and 10^{-2} under both read-outs, while the physical mass it emits drifts by 7 to 115 on the three flat benchmarks,

Table 2: **Conservation and prediction on four grid benchmarks, seeds 42 to 44.** Mass drift $|\Delta M|\downarrow$ from the first decoded frame to the final horizon in absolute units, as a geometric mean over seeds, and nRMSE$\downarrow$ at that horizon as an arithmetic mean. Best learned value in bold, second underlined, dagger at seed 42.

	Dam-break		Advection		Allen–Cahn		Sphere	
Model	$\lvert\Delta M\rvert$	nRMSE	$\lvert\Delta M\rvert$	nRMSE	$\lvert\Delta M\rvert$	nRMSE	$\lvert\Delta M\rvert$	nRMSE
reference solver	0	—	10^{-6}	—	10^{-16}	—	30.8	—
constant field	0	0.103	0	0.702	0	0.999	0	1.00
Diagonal recurrences, shared encoder and decoder								
level B	<u>27.7</u>	0.226	477	1.10	4.70	0.813	2.7e4	1.68
level C	41.9	0.227	409	1.02	3.28	<u>0.811</u>	2.4e4	1.62
free	286	<u>0.156</u>	421	<u>1.00</u>	53.9	**0.707**	1.9e11	div.
LRU	2.3e3	0.992	1.8e3	1.74	18.6	1.07	6.7e4	3.01
Neural operators, with and without a mass corrector								
FNO	3.6e6†	1.8e3†	9.3e3	63.2	61.0	2.20	2.4e7†	div.†
CPFNO, uncorrected	4.4e18†	∞†	396	1.85	171	1.83	8.3e4†	1.29†
CPFNO, corrected	1.3e6†	∞†	**2.8e−4**	1.85	<u>2.7e−4</u>	1.83	1.8e4†	1.29†
KC, unconstrained decoder	115	0.221	7.25	**0.727**	33.4	2.07	<u>785</u>	<u>1.12</u>
KC	**8.3e−3**	**0.146**	<u>2.6e−3</u>	1.79	**2.0e−5**	0.961	**483**	**1.05**

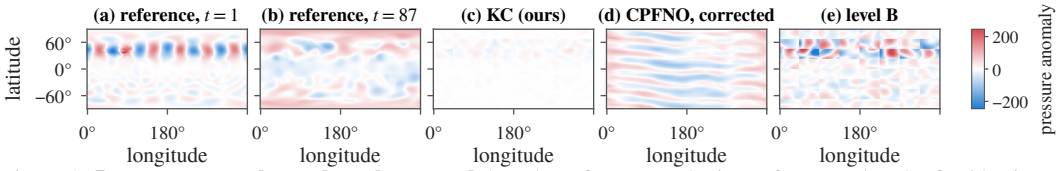


Figure 2: **Pressure anomaly on the sphere, seed 42.** The reference at the input frame and at the final horizon, beside the decoded fields of KC, of CPFNO after its corrector and of level B, on one colour scale. The field of KC stays near the mass-matched constant, whereas the corrected output keeps a zonal structure and drifts.

three to six orders above the tied decoder. Mass correction reaches a lower drift than KC on advection, 2.8×10^{-4} against 2.6×10^{-3}, on a field whose nRMSE of 1.85 is worse than the constant field at 0.702, which is the failure the next question examines. Conservation and accuracy are separate, as Figure 2 shows: the field of KC lies near the mass-matched constant, and no model predicts the evolving field better than that constant under the shared per-patch encoder (Appendix D.6).

4.2 Q2: Correction at the Output Does Not Replace the Condition

Setup. We compare two correctors against the assembly on the dam break over five seeds. The first is the corrector of CPFNO [Liu et al., 2025], $\hat{u} + (m_0 - m(\hat{u}))A$, applied inside the forward pass with a learned weight A and trained end-to-end on the corrected output, and we log the prediction before and after it at every step so that the two can be read separately. The second adds the missing mass uniformly over the grid at inference on top of FNO, without training, which separates the correction from the fit that precedes it. We count a seed as diverged at the first step where nRMSE exceeds 100, after which it enters the median as infinite, so that divergence is not hidden by the seeds that survive.

Results. Correction restores the scalar and not the field, and it does neither once the prediction it corrects diverges. What it does achieve is large: the corrected CPFNO output holds mass near 10^{-3} for fifty steps while the prediction before correction drifts by 10^4 from the first step (Figure 3a), so seven orders of magnitude are closed at every step. What the drift curve does not show is visible in the cross-sections, where the uncorrected prediction lies at $h \approx 0.06$ across the domain and the corrected output, with its mass added back, retains a bump at the centre while the true solution has spread into a ring. The same split appears wherever the prediction stays bounded, and on advection and Allen–Cahn correction drives the drift to 3×10^{-4} on fields whose error of 1.8 exceeds that of a constant field (Table 2). Once the prediction diverges the corrector goes with it, with four seeds out of five diverging between steps 48 and 83 and the fixed corrector diverging on every seed between steps 58 and 97 (Figure 3b), and KC is the only model that reaches the final horizon with bounded error under every seed. On the one seed where FNO reaches that horizon it is twice as accurate as KC and drifts its mass by 230, which is accuracy without the guarantee. The same failure mode has been reported for a climate emulator [Chapman et al., 2026]. The condition must therefore be encoded in the operator, which leaves open which condition carries which quantity.

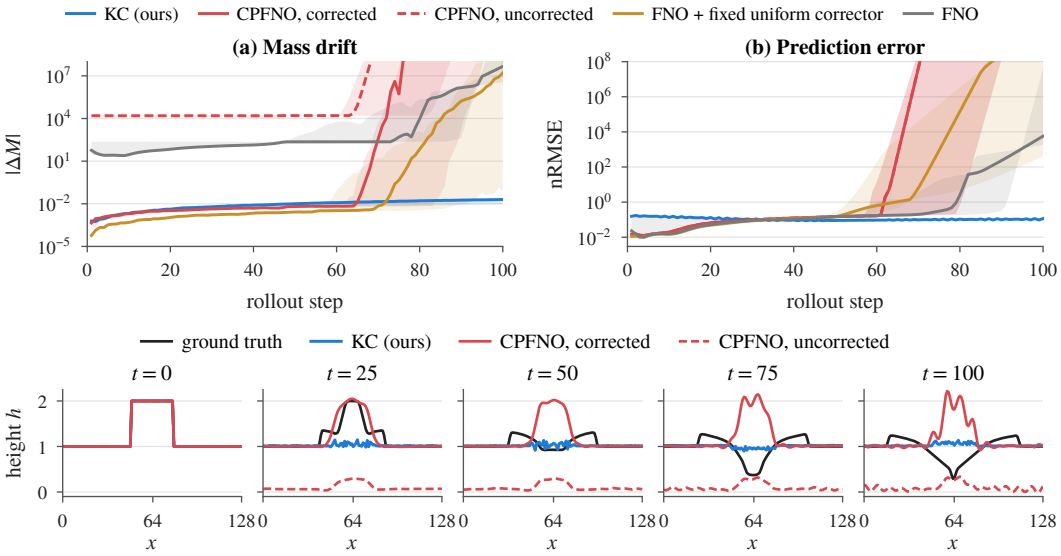


Figure 3: **Mass correction on the dam-break benchmark.** Top: **(a)** mass drift and **(b)** prediction error over a 100-step rollout, median over five seeds with the interquartile range shaded. Bottom: water height along the central row at five steps, seed 42. Per-seed values are in Tables 9 and 10.

Table 3: **Which condition holds which quantity.** Two controls, each removing one condition of Corollary 3 while keeping the others. Drift in percent, where a value near 10^{-6} is the floor of the float32 measurement, with the per-seed values of both controls in Appendix D.3.

	Condition present			Drift (%)	
Transition	$\lvert d_k\rvert{=}1$	$\mathbf{L1}{=}0$	$\mathbf{L}{\otimes}I_m$	energy / norm	momentum
Matched perturbation of the trained Laplacian, $\varepsilon = 0.5$, step 25, mean over three seeds					
row sums kept	✓	✓	✓	**1.8e−6**	**3.5e−5**
row sums broken	✓	✗	✓	**2.0e−6**	110
Norm-preserving operators, operator level, 2000 steps, mean over three draws					
Haar unitary	✓	✗	✗	3.5e−3	169.7
Cayley, dense	✓	✗	✗	1.2e−3	170.7
diagonal unitary	✓	✗	✗	2.5e−4	172.5
KC (ours)	✓	✓	✓	**3.7e−5**	**4.0e−3**

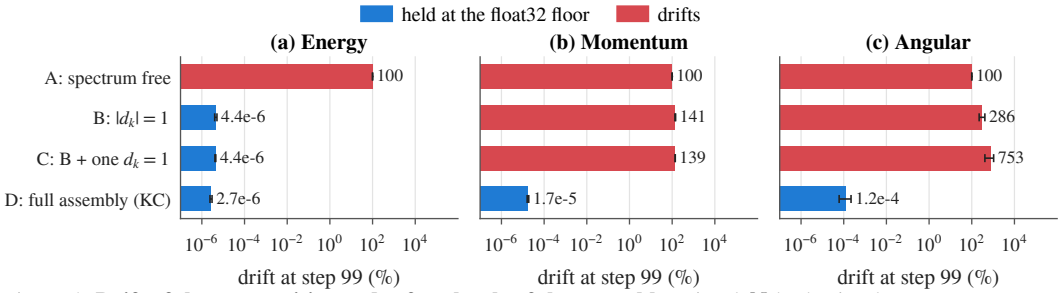


Figure 4: **Drift of three quantities under four levels of the assembly.** Visual N-body simulator, step 99, mean over three seeds with min–max bars, one panel per quantity. Blue marks a drift at the float32 floor and red one above it. Unit modulus closes the energy panel at level B, whereas the summed and the angular quantity stay open until the full assembly closes them at level D. Per-seed values are in Table 12.

4.3 Q3: Each Condition Holds Exactly One Quantity

Setup. We test Corollary 3 by removing one condition at a time, since it holds only if each removal costs exactly the quantity that condition carries and leaves the others at the floor. We first train the four levels of §3.2 on a visual N-body simulator, which carries all three quantities at once. We then perturb the trained Laplacian by a symmetric draw of scale ε, once with the diagonal corrected so that $\mathbf{L1} = 0$ still holds after the change and once with the same draw left uncorrected, which separates the row-sum condition from the coupling it travels with. We finally compare four norm-preserving

operators at the operator level, where no encoder or decoder is involved. Each control is repeated three times, over seeds where a model is trained and over draws at the operator level.

Results. The three controls separate the three conditions, and each removal costs exactly the quantity that condition holds (Table 3 and Figure 4). Energy is held at the float32 floor from level B onward, whereas momentum and the angular pairing are held only at level D. Unit modulus alone holds the norm to 4×10^{-6} and leaves momentum at 141 percent and the angular pairing at 286, and fixing one mode per object at unity in the coordinate basis holds neither, since Equation 12 requires the uniform vector to lie in a stationary mode, which the coordinate modes do not provide. The perturbation isolates the row sums: a change of up to a quarter of the operator's Frobenius norm costs nothing when they are restored and 110 percent of the summed quantity when they are not, with energy at 2×10^{-6} in both cases at every ε (Table 13). At the operator level four norm-preserving transitions are equally stable below 10^{-2}, yet three lose the summed quantity at 170 percent, including the diagonal unitary that a spectral parameterisation produces. Stability and conservation are therefore distinct, and the condition separating them lies in the basis rather than the spectrum, which raises the question of whether training can supply it.

4.4 Q4: Training Reaches the Condition Only by Freezing the Spectrum

Setup. We test which of the two solutions admitted by Proposition 7 training reaches, since one places the uniform vector in a stationary mode while the remaining modes rotate and the other freezes every mode, so that the latent no longer evolves. We replace the assembled basis with a free orthogonal matrix, initialised near the identity or Haar-randomly, and add a penalty $\lambda \, \mathrm{drift}_{\mathrm{momentum}}$ to the reconstruction loss. We train sixteen seeds at $\lambda = 10$ and sweep the weight over $\lambda \in \{0, 0.1, 1, 10\}$, so that the outcome does not rest on one choice of penalty, and we train a Hamiltonian and a graph network to convergence at six training-set sizes, which fixes the reference for a conservation that is learned rather than assembled. We read every run by the angle of its basis to the uniform vector, by the mean $|\lambda_k|$ over the modes that do not carry it and by the momentum drift at step 25.

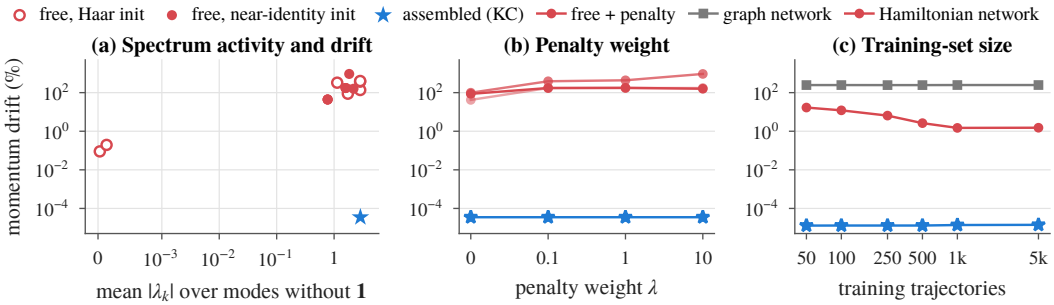


Figure 5: **Conservation and spectral activity of a learned modal basis, visual N-body simulator. (a)** Momentum drift at step 25 against the mean $|\lambda_k|$ over the modes that do not carry the uniform vector, for sixteen free-basis seeds and for the assembled operator. **(b)** Drift as the penalty weight varies, where every nonzero weight lands above the unpenalised run. **(c)** Drift as the training-set size varies, where the learned references flatten and the assembled operator does not move. Values in Tables 17–19.

Results. The condition is reached by training only through the second solution, in which the dynamics are given up. Fourteen of sixteen seeds keep a rotating spectrum, with mean $|\lambda_k|$ between 0.77 and 2.85 on the modes that do not carry the uniform vector, and drift by 44 to 932 percent (Figure 5a). The remaining two reach a mean $|\lambda_k|$ below 10^{-3} and drift by 0.1 to 0.2 percent, having driven every mode to $d_k = 1$ so that the latent no longer evolves (Table 17). The assembled operator alone conserves with an active spectrum, at mean $|\lambda_k| = 2.87$ and drift 3.5×10^{-5} percent, because $c = e_{k^*}$ holds by construction. The basis is not moved by the penalty either, rotating by under ten degrees from its initialisation on every seed under both initialisations (Tables 15 and 16). Any nonzero weight raises drift above the unpenalised run on every seed (Figure 5b), and a Hamiltonian network improves from 16.9 percent at $n = 50$ to 1.49 at $n = 1000$ and no further (Figure 5c). Assembly is thus the only solution that holds conservation and the dynamics together, which leaves the question of how far it reaches.

4.5 Q5: The Layering Transfers as Far as the Algebra Allows

Setup. The two boundaries of §3.4 are tested separately. The quantity boundary of Corollary 5 is tested on the charged, gravitational and spring N-body systems of the EGNN benchmark [Satorras

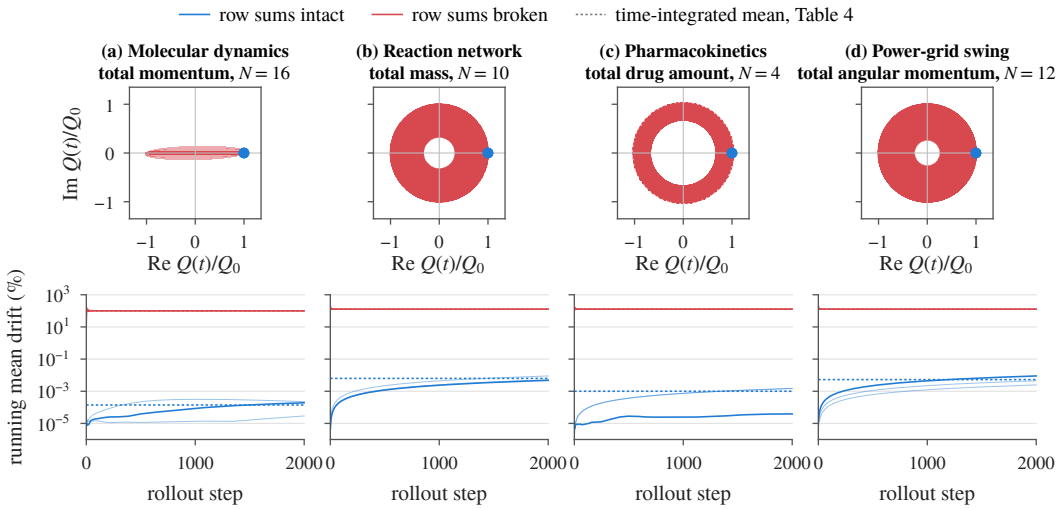


Figure 6: **Four dynamical systems from other fields, operator level,** 2000 **steps.** Top: the conserved quantity $Q(t)$ relative to its initial value as a point in the complex plane, for molecular dynamics the momentum vector in the plane of Q_0, seeds 42/43/44. With the row sums intact it stays at 1, and with them broken its magnitude is nearly kept while its direction rotates, which is the summed quantity lost while the norm is held. Bottom: running time average of $\|Q(t) - Q_0\|/\|Q_0\|$, whose endpoints are the values of Table 21.

et al., 2021, Brandstetter et al., 2022] at $T = 1000$, where linear momentum is linear in the state, angular momentum is bilinear and energy is quadratic only on the spring system, together with five Hermitian forms outside the family of Corollary 4. The dynamics boundary of Proposition 6 is tested on a stiff reaction network relaxing to equilibrium, and transfer at the operator level on four systems from other fields, with the row sums intact and broken by a matched perturbation.

Results. Exactly the quantities admitted by Corollary 5 are held, and the boundary is visible in the same run. Linear momentum is held at 10^{-6} while every other baseline drifts by 1 to 8, angular momentum improves by 3 to 12 times without reaching the floor, since a bilinear quantity is not placed on a held mode by the linear tie, and the lowest energy drift is reached on the spring system, whose potential is quadratic (Table 20). The same boundary appears inside the latent, where the zero mode of the family drifts by 10^{-5} while five Hermitian forms outside it drift by 10 to 117 under the same operator (Table 24), and the decoded total momentum accordingly stays at a point over 1000 steps under KC while every diagonal transition carries it away (Figure 13). On the dynamics boundary the latent zero mode is still held to 5.6×10^{-5} percent on the stiff system, while baselines that can contract fit the state more accurately, as Proposition 6 predicts (Table 25). The layering further carries to molecular dynamics, chemical kinetics, pharmacokinetics and power-grid swing dynamics, where the conserved quantity drifts by 10^{-4} to 10^{-2} percent with the row sums intact and by 100 to 128 with them broken (Table 21 and Figure 6). What is held, and where the construction ceases to apply, is therefore fixed before any parameter is set.

5 Conclusion

Conservation in a diagonal recurrence is layered: the norm, the summed quantity and an angular one each rest on a separate condition on the operator, and a unit-modulus spectrum conserves only the first. We therefore propose the Kronecker–Cayley transition (KC), which assembles the generator as a graph Laplacian and advances it by a Cayley transform, so that every quantity the algebra admits is held at every parameter value without correction. We prove the layering itself, the exact $2m^2$-dimensional family the assembled operator holds, and the condition under which a transition holds the summed quantity, which a trained basis reaches only by freezing its spectrum. On four grid benchmarks, KC holds mass at the round-off level of the rollout, whereas every diagonal baseline drifts and two mass correctors fail once the prediction they correct diverges, and the same layering carries to molecular dynamics, chemical kinetics, pharmacokinetics and power-grid swing dynamics.

Acknowledgments and Disclosure of Funding

Author contributions. Jiaqian Zhu conceived the project, developed the theory, designed the method, implemented the code, conducted all experiments, and wrote the manuscript. Yang Zhang provided guidance on the revision of the manuscript and contributed to discussions of the work. Xiaowei Yu provided the primary computing resources. Junhua Ding provided additional computing resources.

Funding. Experiments were run on an NVIDIA RTX PRO 6000 (96 GB) workstation provided by Xiaowei Yu at Missouri University of Science and Technology. Additional computing resources were provided by Junhua Ding at the University of Wyoming. Jiaqian Zhu was supported by a teaching assistantship at Missouri University of Science and Technology. This work received no other third-party funding.

Competing interests. The authors declare no competing interests.

References

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1120–1128. PMLR, 2016. URL https://proceedings.mlr.press/v48/arjovsky16.html.

Omri Azencot, N. Benjamin Erichson, Vanessa Lin, and Michael W. Mahoney. Forecasting sequential data using consistent Koopman autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 475–485. PMLR, 2020. URL https://proceedings.mlr.press/v119/azencot20a.html.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL https://doi.org/10.1017/9781009166164.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J. Bekkers, and Max Welling. Geometric and physical quantities improve E(3) equivariant message passing. In *International Conference on Learning Representations*, 2022.

William E. Chapman, John Schreck, and Yingkai Sha. Hard conservation correctors can hide a degrading model when training autoregressive emulators. *arXiv preprint arXiv:2607.18416*, 2026. URL https://arxiv.org/abs/2607.18416.

Zhengdao Chen, Jianyu Zhang, Martin Arjovsky, and Léon Bottou. Symplectic recurrent neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BkgYPREtPr.

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. URL https://openreview.net/forum?id=iE8tFa4Nq.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10041–10071. PMLR, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Florian Dörfler, Michael Chertkov, and Francesco Bullo. Synchronization in complex oscillator networks and smart grids. *Proceedings of the National Academy of Sciences*, 110(6):2005–2010, 2013. doi: 10.1073/pnas.1212134110.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 15379–15389, 2019. URL https://papers.nips.cc/paper/9672-hamiltonian-neural-networks.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems*, volume 35, pages 35971–35983, 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e9a32fade47b906de908431991440f7c-Abstract-Conference.html.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. In *Advances in Neural Information Processing Systems*, volume 35, pages 22982–22994, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9156b0f6dfa9bbd18c79cc459ef5d61c-Abstract-Conference.html.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=dPSTDbGtBY.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 2006. doi: 10.1007/3-540-30666-8.

Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. Springer, 2nd edition, 2015. doi: 10.1007/978-3-319-13467-3.

Stephen C. Harvey, Robert K.-Z. Tan, and Thomas E. Cheatham, III. The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. *Journal of Computational Chemistry*, 19(7):726–740, 1998. doi: 10.1002/(SICI)1096-987X(199805)19:7<726::AID-JCC4>3.0.CO;2-S.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled Cayley transform. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1969–1978. PMLR, 2018. URL https://proceedings.mlr.press/v80/helfrich18a.html.

Giancarlo Kerg, Kyle Goyette, Maximilian Puelma Touzel, Gauthier Gidel, Eugene Vorontsov, Yoshua Bengio, and Guillaume Lajoie. Non-normal recurrent neural network (nnRNN): learning long time dependencies while improving expressivity with transient dynamics. In *Advances in Neural Information Processing Systems*, volume 32, pages 13591–13601, 2019.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2688–2697. PMLR, 2018. URL https://proceedings.mlr.press/v80/kipf18a.html.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 26548–26560, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/df438e5206f31600e6ae4af72f2725f1-Abstract.html.

Kookjin Lee and Kevin T. Carlberg. Deep conservation: A latent-dynamics model for exact satisfaction of physical conservation laws. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 277–285, 2021. doi: 10.1609/aaai.v35i1.16102.

Mario Lezcano-Casado and David Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3794–3803. PMLR, 2019. URL https://proceedings.mlr.press/v97/lezcano-casado19a.html.

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

Chaoyu Liu, Yangming Li, Zhongying Deng, Chris Budd, and Carola-Bibiane Schönlieb. Adaptive correction for ensuring conservation laws in neural operators. *arXiv preprint arXiv:2505.24579*, 2025. URL https://arxiv.org/abs/2505.24579.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, 2018. doi: 10.1038/s41467-018-07210-0.

Reza Olfati-Saber, J. Alex Fax, and Richard M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007. doi: 10.1109/JPROC.2006.887293.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26670–26698. PMLR, 2023. URL https://proceedings.mlr.press/v202/orvieto23a.html.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8459–8468. PMLR, 2020. URL https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Adrian Sandu. Positive numerical integration methods for chemical kinetic systems. *Journal of Computational Physics*, 170(2):589–602, 2001. doi: 10.1006/jcph.2001.6750.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Ai8Hw3AXqks.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1596–1611, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0a9747136d411fb83f0cf81820d44afb-Abstract-Datasets_and_Benchmarks.html.

Abishek Thangamuthu, Gunjan Kumar, Suresh Bishnoi, Ravinder Bhattoo, N M Anoop Krishnan, and Sayan Ranu. Unravelling the performance of physics-informed graph neural networks for dynamical systems. In *Advances in Neural Information Processing Systems*, volume 35, pages 3691–3702, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/17b598fda495256bef6785c2b76c3217-Abstract-Datasets_and_Benchmarks.html.

J. J. P. Veerman, T. Whalen-Wagner, and Ewan Kummel. Chemical reaction networks in a Laplacian framework. *Chaos, Solitons & Fractals*, 166:112859, 2023. doi: 10.1016/j.chaos.2022.112859.

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3570–3578. PMLR, 2017.

Scott Wisdom, Thomas Powers, John Hershey, Jonathan Le Roux, and Les Atlas. Full-capacity unitary recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper/6327-full-capacity-unitary-recurrent-neural-networks.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Dissipative SymODEN: Encoding Hamiltonian dynamics with dissipation and control into deep learning. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020a. URL https://openreview.net/forum?id=knjWFNx6CN.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ODE-Net: Learning Hamiltonian dynamics with control. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=ryxmb1rKDS.

Appendix

Table of Contents

A Notation and Implementation Details **16**

B Extended Related Work **18**

C Proofs and numerical detail **19**

C.1 Preliminaries 19

C.2 Characterisation of the conserved set (Theorem 1) 20

C.3 Kronecker–Cayley decomposition (Theorem 2) 20

C.4 Layered conservation (Corollary 3) 20

C.5 The invariant family (Corollary 4) 21

C.6 Applicability (Corollary 5) 21

C.7 Lower bound for a unitary step (Proposition 6) 21

C.8 Conservation of the summed quantity (Proposition 7) 21

C.9 Numerical properties 21

C.10 Batched form 22

C.11 Full algorithm of the Kronecker–Cayley transition 23

D Additional Experiments **24**

D.1 Q1: Mass Is Held Where Every Diagonal Baseline Loses It 24

D.2 Q2: Correction at the Output Does Not Replace the Condition 26

D.3 Q3: Each Condition Holds Exactly One Quantity 27

D.4 Q4: Training Reaches the Condition Only by Freezing the Spectrum 29

D.5 Q5: The Boundary Is Fixed Before Training 30

D.6 Diagnostics on the Accuracy Cap 31

E Limitations and Future Directions **33**

E.1 Limitations 33

E.2 Future Directions 33

A Notation and Implementation Details

Table 4: **Notation.**

Symbol	Meaning	Introduced
Latent state and partition		
$N,\ m,\ d$	object subspaces, width of each subspace, and latent dimension, $d{=}Nm$	§3.1
$z{=}(z_1,\dots,z_N)\in\mathbb{C}^d$	latent state, with $z_i\in\mathbb{C}^m$ the subspace of object i	§3.1
$x_t,\ a_t,\ t,\ h$	observation, action, time index, and rollout step	§3.1
$E,\ D$	encoder $x\mapsto z$ and decoder $z\mapsto\hat{x}$, shared by every latent transition compared here	§3.1
The operator		
$w_{ij}{=}\mathrm{MLP}_\theta(\lVert z_i{-}z_j\rVert^2, a)$	learned pairwise coupling, symmetrised as $(w_{ij}{+}w_{ji})/2$. The only trained component of the transition	Eq. 4
$\mathbf{L}\in\mathbb{R}^{N\times N}$	graph Laplacian, $\mathbf{L}_{ij}{=}-w_{ij}$ off the diagonal and $\mathbf{L}_{ii}{=}\sum_{j\neq i}w_{ij}$, hence symmetric with zero row sums	Eq. 5
$\mathbf{H}{=}\mathbf{L}\otimes I_m$	isotropic pairwise generator: Hermitian, zero row sums, blocks proportional to the identity	Eq. 5
$\mathrm{Cay}(\mathbf{H}){=}(I{+}\frac{\mathrm{i}}{2}\mathbf{H})^{-1}(I{-}\frac{\mathrm{i}}{2}\mathbf{H})$	Cayley transform, unitary for every Hermitian $\mathbf{H}$	Eq. 6
$\mathbf{L}{=}P\Lambda P^\top$, $\tilde{z}{=}(P^\top\otimes I_m)z$	eigendecomposition of the Laplacian and the resulting modal coordinates	Thm. 2
$d_k{=}(1{-}\mathrm{i}\lambda_k/2)/(1{+}\mathrm{i}\lambda_k/2)$	modal eigenvalue, on the unit circle for every real λ_k	Eq. 7
$\lambda_k,\ k^*$	Laplacian eigenvalue and the index of the zero mode, $\lambda_{k^*}{=}0$ so $d_{k^*}{=}1$	Thm. 2
$\mathbf{1}/\sqrt{N}$	uniform vector, the eigenvector of $\mathbf{L}$ at eigenvalue zero whenever the row sums vanish	Prop. 7
Conserved quantities and the condition each one needs		
$\lVert z\rVert^2$	latent norm, preserved by $\lvert d_k\rvert{=}1$ alone and therefore insensitive to the row sums	Eq. 2
$p(z){=}\sum_i z_i$	momentum proxy, preserved by $d_{k^*}{=}1$, which holds only when the row sums vanish	Eq. 2
$\mathrm{Im}(z^\dagger\mathbf{J}z)$	angular proxy, preserved by the scalar-times-identity block structure of $\mathbf{H}$	Eq. 2
$\mathrm{drift}_Q@h$	$100\,\lVert Q(\hat{z}_h)-Q(z_0)\rVert/\lVert Q(z_0)\rVert$, in percent, at rollout step h, measured from the first decoded frame so that the encoder–decoder round trip is excluded. That round trip loses 2 percent of the dam-break mass once, at $t{=}1$	§4
tied readout	fixed linear map carrying total mass, or linear momentum on N-body systems, to the latent zero mode $(\mathbf{1}^\top\otimes I_m)z$, so that Corollary 5 applies. On the sphere each patch is scaled by the inverse of its mean $\cos(\mathrm{lat})$, so the zero mode is the area-weighted integral at patch resolution	Tab. 5
Basis, spectrum, and the conservation condition		
$O(N)\times(S^1)^N$	parameter space of a unitary diagonal transition, $U{=}(Q\,\mathrm{diag}(d)Q^\top)\otimes I_m$, carrying a modal basis and a spectrum on the torus	Prop. 7
$Q\in O(N)$	modal basis of the transition. Either the Laplacian eigenbasis, or a free orthogonal matrix in the learned-basis control	Prop. 7
$c{=}Q^\top\mathbf{1}/\sqrt{N}$	coefficients of the uniform vector in the modal basis. The summed quantity is preserved exactly when $(d_k{-}1)c_k{=}0$ for every k	Eq. 12
φ	angle between $\mathbf{1}/\sqrt{N}$ and the nearest column of Q, computed as $\arccos\max_j\lvert c_j\rvert$. Zero when the uniform vector is a column, at most $\arccos(1/\sqrt{N})$ when it is split equally	Tab. 15
$\lvert Q^\top\mathbf{1}/\sqrt{N}\rvert$	projection of the uniform vector on each mode. One entry equal to 1 means a single mode carries the summed quantity, and $1/\sqrt{3}\approx0.577$ in every entry means it is divided equally among three	Tab. 17
Operator perturbation		
ε, $\lVert\mathbf{L}'{-}\mathbf{L}\rVert_F/\lVert\mathbf{L}\rVert_F$	perturbation scale and the relative Frobenius change it produces	Tab. 13
compensated	diagonal corrected so that the row sums stay zero, which leaves the condition for momentum intact	Tab. 13
uncompensated	same Frobenius magnitude with no diagonal correction, so the row sums no longer vanish	Tab. 13
Cross-domain drift		
time-integrated mean	$\frac{1}{T}\sum_{t=1}^{T}\mathrm{drift}_Q@t$ over $T{=}2000$ steps. Does not saturate	Tab. 21
Evaluation		
$\mathrm{PSNR}@h$	decoded-frame peak signal-to-noise ratio at step h	Tab. 19
mean-frame baseline	PSNR of the training-set mean frame used as a constant prediction, reported per evaluation set in §4. Sets the level above which PSNR carries information about the dynamics	§4
s	seed, $s\in\{42,43,44\}$ unless the caption states otherwise. The alignment study extends to $s\in\{42,\dots,49\}$	App. D

Table 5: **Implementation details.**

Setting	Value	Scope
Public, grid		
Dam break	Shallow-water dam break of PDEBench [Takamoto et al., 2022], 1000 trajectories at 128×128, split $900/100$. The conserved field is the total mass	
Advection	Advection equation of PDEBench [Takamoto et al., 2022], 1000 trajectories at 1024 points, split $900/100$	
Sphere	Spherical shallow water of PDEArena [Gupta and Brandstetter, 2023], 96×192 latitude–longitude grid, 500 trajectories split $400/100$ over disjoint initial-condition seeds, pressure anomaly as the conserved field. Re-posed as in §D.1	
Allen–Cahn	Generated by us in the Rubinstein–Sternberg form on a 128×128 periodic grid, 1000 trajectories of 100 steps, with mass conserved to 10^{-16} and energy monotone decreasing	
Public, particles		
N-body, EGNN	Charged, gravitational and spring systems of the EGNN benchmark [Satorras et al., 2021, Brandstetter et al., 2022] at the standard $3000/2000/2000$ split	
Spring benchmark	Thangamuthu et al. [2022] under its published protocol: equal masses on a closed ring, two connections per body, unit stiffness and rest length, $q_0 \sim \mathcal{U}(0,2)$ and $p_0 \sim \mathcal{U}(0,0.1)$ after mean centring, Störmer–Verlet at $\Delta t{=}10^{-3}$, subsampled to 100 frames. Scored on its own three metrics	
Visual, in distribution		
N-body, 30 frames	Planar gravitational N-body, $N{=}3$, unit constant, softening 0.1, leapfrog at $\Delta t{=}0.01$. Positions $\mathcal{N}(0,0.5)$ and velocities $\mathcal{N}(0,0.3)$ with the centre of mass removed, so the physical momentum is zero. Rendered at 64×64, one colour channel per object, actions $\mathcal{N}(0,0.01)$, $2000/256$ split	
N-body, 101 frames	Same physics and renderer over a longer horizon, 256 trajectories. Used for every rollout beyond 29 steps	
Cross domain, operator level		
Molecular dynamics	$N{=}16$ particles of width $m{=}8$, state encoded as position plus i times momentum per axis. Conserved quantity is the total linear momentum $\sum_i m_i v_i$, read from the imaginary part of the first two channels. A Lennard-Jones simulation with velocity Verlet at $\Delta t{=}0.005$ in float64 supplies the symplectic floor and the fixed operator of the spectrum-invariant control	
Chemical reaction network	$N{=}10$ species of width $m{=}4$, conserving the total mass $\sum_i c_i$, read from the real part of the first channel	
Pharmacokinetics	$N{=}4$ compartments of width $m{=}4$, conserving the total amount across compartments	
Power-grid swing	$N{=}12$ nodes of width $m{=}4$, conserving the mass-weighted swing momentum $\sum_i M_i \dot{x}_i$ with $M_i \sim \mathcal{U}(1,5)$, the uniformly weighted zero mode at the operator level	
Architecture and training, one setting per dataset family		
Grid benchmarks	$N{=}64$ patches of width $m{=}8$, per-patch MLP encoder and tied per-patch linear decoder. Adam at 10^{-3}, batch 32, 100 epochs and 150 on the sphere. Rollouts of 100 steps on the dam break and Allen–Cahn, 200 on advection and 86 on the sphere	Dam break, advection, Allen–Cahn, sphere
N-body	$N{=}5$ or 10 particles of width $m{=}8$, per-particle MLP encoder and tied per-particle linear decoder. Adam at 10^{-3}, batch 32, 200 epochs, rollouts of 1000 steps	Charged, gravity, spring
Visual N-body	$z \in \mathbb{C}^{120}$, $N{=}3$ subspaces of width $m{=}40$. Four convolutional blocks with channels $[64, 128, 256, 512]$, a U-Net decoder with skip decay $K{=}10$, and 15M parameters in total. Rollouts of 99 steps, under the protocols below	All latent transitions
Spring benchmark	$N{=}3$ or 5 objects of width $m{=}40$ under the same convolutional encoder and decoder. Adam at 10^{-3}, batch 16, 40 epochs, rollouts of 25 steps	Spring benchmark
Coupling network, Cayley step	MLP $17 \to 64 \to 32 \to 1$ on $(\|z_i{-}z_j\|^2, a)$, symmetric in (i,j), 3,963 parameters with the action embedding and the only trained part of the transition. One float64 eigendecomposition of the $N\times N$ Laplacian, then diagonal phase rotations	Constructed basis
Training protocols, Adam throughout, no early stopping		
P1	2×10^{-4}, batch 16, 30 epochs, seeds $\{42, 43, 44\}$	Four-level ablation, main comparison
P1′	2×10^{-4}, batch 16, 40 epochs, seeds $\{42, \ldots, 49\}$	Alignment test and penalty sweep
P2	3×10^{-4}, batch 32, 80 epochs, seeds $\{42, 43, 44\}$	External recurrent baselines, basis ablation, long-horizon rollouts
P3	5×10^{-3}, batch 128, 200 epochs	Synthetic invariant task
P5	2×10^{-4}, batch 16, 30 epochs, stopping after 20 epochs without improvement and capped at 150, seeds $\{42, 43, 44\}$	Sample-size sweep trained to convergence
Controls		
Four levels	A leaves $\|d_k\|$ free, B adds $\|d_k\|{=}1$, C adds one mode per object fixed at unity in the coordinate basis, and D is the full construction. Each level adds one ingredient	Separates the spectrum from the basis
Learned orthogonal basis	Same architecture and objective with Q a free orthogonal matrix. Penalty weight $\lambda \in \{0, 0.1, 1, 10\}$, seeds $\{42, \ldots, 49\}$. Initialised either near the identity, as $Q{=}\,\mathrm{Cay}((W{-}W^\top)/2)$ with W i.i.d. $\mathcal{N}(0, 0.01^2)$ so that $\|Q{-}I\|_F \approx 0.03$, or Haar-randomly from the QR factorisation of a standard Gaussian	Isolates the source of the basis
Norm-preserving operators	A Haar-random unitary, a dense unitary from the Cayley transform of a skew-Hermitian matrix mixing all d modes, a diagonal unitary with an independent phase per mode, and the construction. All four have unit-modulus eigenvalues and are therefore stable	Separates stability from conservation
Spectrum-invariant read-out	One fixed Laplacian with its spectrum held constant, read along the eigenvector of smallest $\|\lambda\|$ and along that of largest $\|\lambda\|$, on the synthetic graph and on the molecular-dynamics operator	Isolates the read-out direction
External operators	FNO using the published implementation and default settings. CPFNO reimplemented after Liu et al. [2025], corrector $\hat{u}+(m_0{-}m(\hat{u}))A$ for a learned softmax weight A inside the forward pass and the loss on the corrected output, logged before and after the corrector at every step	Correction against condition

Table 5: **Implementation details** (continued).

Setting	Value	Scope
Compensated perturbation	Diagonal corrected so that the row sums stay zero, against an uncompensated perturbation of the same Frobenius magnitude	Separates the two conditions
Evaluation		
Rollout, precision	25 steps on the 30-frame sets, 99 and 2000 on the long-horizon sets, float32 throughout, which sets the floor on the reported drift. The symplectic reference is computed in float64, since float32 overflows on the Lennard-Jones force	All experiments
Determinism	Deterministic given the seed on the hardware used here, verified by regenerating all four benchmarks and the three N-body systems at seed 42 and recovering the stored values to three significant figures. FNO and CPFNO on advection are run under `torch.use_deterministic_algorithms`, since the tail of the rollout is sensitive to one-dimensional CUDA reductions	All experiments
Seeds	$\{42, 43, 44\}$ for every trained model, extended to $\{42, \ldots, 49\}$ for the alignment study	All experiments

B Extended Related Work

Parameterisations of the modal basis. The constructions that secure stability act on the spectrum and obtain the modal basis as a by-product. Unitary recurrence restricts the transition to a manifold, through a product of structured unitary factors [Arjovsky et al., 2016], a Stiefel-manifold gradient step on the full matrix [Wisdom et al., 2016], a scaled Cayley transform of a skew-symmetric parameter [Helfrich et al., 2018], or the exponential map of a Lie algebra [Lezcano-Casado and Martínez-Rubio, 2019], each resting on standard manifold optimisation [Absil et al., 2008, Hall, 2015, Boumal, 2023], and the eigenvectors are whatever the chosen parameterisation returns at the optimum. Koopman autoencoders instead learn the coordinates, advancing an encoded latent by a linear operator so that nonlinear dynamics become linear in the learned basis [Lusch et al., 2018, Azencot et al., 2020], which makes that basis a function of the data and of the optimiser. Our construction applies the transform of Equation 6 used by scoRNN and expRNN to a Laplacian rather than to a free skew-symmetric parameter, so that the eigenbasis is determined by the interaction structure before any value is fitted, and the vanishing row sums place the uniform vector at eigenvalue one, which is the mode that carries the summed quantity at every value of the coupling (Proposition 7).

Conservation through a learned generator. The generator-based designs name the conserved scalar and learn it. A Hamiltonian network learns the energy and advances the state along its symplectic gradient [Greydanus et al., 2019], a Lagrangian network removes the requirement of canonical coordinates [Cranmer et al., 2020], a symplectic recurrent network integrates the learned energy over several steps at once [Chen et al., 2020], and SymODEN extends the construction to systems with control inputs and to dissipation [Zhong et al., 2020b,a], while a related design states the quantity as a constraint set and solves it at each step [Lee and Carlberg, 2021]. Benchmarks across this family report that the benefit obtained depends on how closely the stated structure matches the system at hand [Thangamuthu et al., 2022]. A learned scalar is a general function of the state, so the induced flow is not linear in any fixed basis and the linear-time diagonal step is unavailable. Restricting the conserved set to quadratic forms is the price of retaining that step, and the commutant of Theorem 1 then returns the entire $2m^2$-dimensional family of Corollary 4 rather than one named quantity.

Zero modes of a Laplacian in the applied sciences. For a symmetric Laplacian with vanishing row sums, $\mathbf{L}\mathbf{1} = 0$, so the flow $\dot{x} = -\mathbf{L}x$ satisfies $\frac{d}{dt}\mathbf{1}^\top x = -\mathbf{1}^\top \mathbf{L}x = 0$. The eigenvalue is exactly zero and the eigenvector is exactly the uniform vector for every connected weighted graph, so neither one depends on the coupling strengths and neither is perturbed when those strengths are estimated from data. Each applied field reads the resulting sum as its own invariant, the conserved mass of a reaction network [Veerman et al., 2023], the agreement value of average consensus [Olfati-Saber et al., 2007], the retained dose of a compartmental pharmacokinetic model, and the total angular momentum of the swing equations of a power network [Dörfler et al., 2013]. In each case the matrix is supplied by the physical model and the zero mode is read from it. The construction of Equation 5 proceeds in the opposite direction, assembling the matrix from learned couplings under a pattern that imposes the zero mode, which is the mechanism by which those four systems transfer at the operator level in §4.5 and lose their invariant once a matched perturbation breaks the row sums.

C Proofs and numerical detail

We prove the results of §3 in the order they appear, then record the numerical properties that set the floor against which the measured drift is read (§C.9) and the batched form used in the experiments (§C.10). Throughout, $\mathbf{L} \in \mathbb{R}^{N\times N}$ is the graph Laplacian of Equation 5, $\mathbf{H} = \mathbf{L} \otimes I_m$, $U = \mathrm{Cay}(\mathbf{H})$, $\mathbf{L} = P\Lambda P^\top$ with $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$, and k^* is the index of the zero eigenvalue.

C.1 Preliminaries

Lemma 8 (Structure of the assembled generator)**.** *For every value of θ the matrix $\mathbf{H}$ of Equation 5 satisfies*

(i) $\mathbf{H}^\dagger = \mathbf{H}$,

(ii) $\mathbf{H}(\mathbf{1} \otimes e) = 0$ *for every* $e \in \mathbb{C}^m$,

(iii) $[\mathbf{H},\ I_N \otimes C] = 0$ *for every* $C \in \mathbb{C}^{m\times m}$.

Proof. Property (i). The coupling network returns real values and the assembly symmetrises them, $w_{ij} = w_{ji}$, so $\mathbf{L}$ is real symmetric and $\mathbf{L} = \mathbf{L}^\top = \mathbf{L}^\dagger$. The Kronecker product of Hermitian factors is Hermitian, hence $\mathbf{H}^\dagger = \mathbf{H}$.

Property (ii). Row i of $\mathbf{L}$ sums to $\mathbf{L}_{ii} + \sum_{j\neq i} \mathbf{L}_{ij} = \sum_{k\neq i} w_{ik} - \sum_{j\neq i} w_{ij} = 0$, so $\mathbf{L}\mathbf{1} = 0$. Then $\mathbf{H}(\mathbf{1} \otimes e) = (\mathbf{L} \otimes I_m)(\mathbf{1} \otimes e) = (\mathbf{L}\mathbf{1}) \otimes e = 0$.

Property (iii). Using $(A \otimes B)(C \otimes D) = AC \otimes BD$,

$$(\mathbf{L} \otimes I_m)(I_N \otimes C) = \mathbf{L} \otimes C = (I_N \otimes C)(\mathbf{L} \otimes I_m),$$

which is the stated commutator identity. The three properties follow from the pattern of Equation 5 and hold whatever values the coupling network returns. □

Lemma 9 (Cayley transform of a Kronecker product)**.** *Let $A \in \mathbb{C}^{N\times N}$ be Hermitian. Then* $\mathrm{Cay}(A \otimes I_m) = \mathrm{Cay}(A) \otimes I_m$*, and* $\mathrm{Cay}(A)$ *is unitary.*

Proof. Write $S = \frac{\mathrm{i}}{2}A$, so $S^\dagger = -S$. The matrix $I + S$ is invertible, since $(I + S)v = 0$ would imply $v^\dagger v = -v^\dagger S v$, whose right-hand side is purely imaginary while the left is real and non-negative, forcing $v = 0$. For the Kronecker structure, $S \otimes I_m = \frac{\mathrm{i}}{2}(A \otimes I_m)$ and

$$\big(I_{Nm} + S \otimes I_m\big)^{-1} = \big((I_N + S) \otimes I_m\big)^{-1} = (I_N + S)^{-1} \otimes I_m,$$

using $(X \otimes I_m)^{-1} = X^{-1} \otimes I_m$. Multiplying by $(I_N - S) \otimes I_m$ and collecting the factors yields $\mathrm{Cay}(A \otimes I_m) = \mathrm{Cay}(A) \otimes I_m$. Unitarity follows from

$$\mathrm{Cay}(A)^\dagger\,\mathrm{Cay}(A) = (I-S^\dagger)^{-1}(I+S^\dagger)(I-S)(I+S)^{-1} = (I+S)^{-1}(I-S)(I-S)^{-1}(I+S) = I,$$

where the middle equality uses $S^\dagger = -S$ and the commutation of $(I \pm S)$ with each other. □

Lemma 10 (Matrices commuting with every Laplacian)**.** *Let $N \geq 2$ and let $\mathcal{L}_N$ be the set of real symmetric $N \times N$ matrices with zero row sums. A matrix $A \in \mathbb{C}^{N\times N}$ satisfies $[\mathbf{L}, A] = 0$ for every $\mathbf{L} \in \mathcal{L}_N$ if and only if $A \in \mathrm{span}\{I_N, \mathbf{1}\mathbf{1}^\top\}$.*

Proof. Sufficiency. I_N commutes with everything, and $\mathbf{L}\mathbf{1}\mathbf{1}^\top = (\mathbf{L}\mathbf{1})\mathbf{1}^\top = 0 = \mathbf{1}(\mathbf{L}\mathbf{1})^\top = \mathbf{1}\mathbf{1}^\top\mathbf{L}$ for every $\mathbf{L} \in \mathcal{L}_N$, using symmetry of $\mathbf{L}$.

Necessity. $\mathcal{L}_N$ is a real vector space spanned by the elementary Laplacians $E_{ij} = (e_i - e_j)(e_i - e_j)^\top$ for $i < j$, since every symmetric zero-row-sum matrix is a combination of them. Commuting with every $\mathbf{L}$ is equivalent to commuting with every E_{ij}. Write $u = e_i - e_j$, so $E_{ij} = uu^\top$. Then $[uu^\top, A] = u(u^\top A) - (Au)u^\top$, and this vanishes exactly when $Au = \mu u$ and $u^\top A = \mu u^\top$ for some scalar μ, that is when u is a right and left eigenvector of A with the same eigenvalue. Since this holds for every difference $e_i - e_j$, the subspace $\mathbf{1}^\perp$ spanned by those differences lies in a single eigenspace of A, and A acts on $\mathbf{1}^\perp$ as a scalar μ. Symmetry of E_{ij} also forces A to map $\mathbf{1}^\perp$ into itself and, by the same argument on the transpose, $\mathbf{1}$ into $\mathrm{span}\{\mathbf{1}\}$. Hence $A = \mu\,\Pi_\perp + \nu\,\Pi_\parallel$ with $\Pi_\parallel = \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ and $\Pi_\perp = I_N - \Pi_\parallel$, which is a combination of I_N and $\mathbf{1}\mathbf{1}^\top$. □

C.2 Characterisation of the conserved set (Theorem 1)

Proof. Let $U = \mathrm{Cay}(\mathbf{H})$ with $\mathbf{H}$ Hermitian, so U is unitary by Lemma 9. For a Hermitian G, the identity $Q_G(Uz) = Q_G(z)$ for every z is $z^\dagger U^\dagger G U z = z^\dagger G z$ for every z. Since $U^\dagger G U$ and G are both Hermitian, equality of the two quadratic forms on every z forces $U^\dagger G U = G$, and multiplying on the left by U yields $GU = UG$, that is $[U, G] = 0$.

It remains to show $[U, G] = 0 \Leftrightarrow [\mathbf{H}, G] = 0$. Write $f(\lambda) = (1 - \mathrm{i}\lambda/2)/(1 + \mathrm{i}\lambda/2)$, so $U = f(\mathbf{H})$ in the functional calculus, and f is injective on $\mathbb{R}$, since $f(\lambda) = f(\mu)$ implies $(1-\mathrm{i}\lambda/2)(1+\mathrm{i}\mu/2) = (1 - \mathrm{i}\mu/2)(1 + \mathrm{i}\lambda/2)$ and hence $\lambda = \mu$. Decompose $\mathbf{H} = \sum_k \lambda_k \Pi_k$ into spectral projectors over distinct eigenvalues, so $U = \sum_k f(\lambda_k)\Pi_k$ with the $f(\lambda_k)$ distinct. A matrix G commutes with $\mathbf{H}$ exactly when it commutes with every Π_k, and likewise for U, because in each case the projectors are polynomials in the matrix with distinct eigenvalues. The two commutants therefore coincide. □

C.3 Kronecker–Cayley decomposition (Theorem 2)

Proof. By Lemma 8(i) the matrix $\mathbf{L}$ is real symmetric, so Lemma 9 applies with $A = \mathbf{L}$ and yields $U = \mathrm{Cay}(\mathbf{L}) \otimes I_m$ together with unitarity.

Substituting $\mathbf{L} = P \Lambda P^\top$ and using that Cay acts on the spectrum,

$$\mathrm{Cay}(\mathbf{L}) = P \ \mathrm{diag}(d_1, \ldots, d_N)\, P^\top, \qquad d_k = \frac{1 - \mathrm{i}\lambda_k/2}{1 + \mathrm{i}\lambda_k/2}. \tag{13}$$

Each λ_k is real because $\mathbf{L}$ is symmetric, so $|d_k|^2 = d_k \overline{d_k} = \frac{1+\lambda_k^2/4}{1+\lambda_k^2/4} = 1$, so every modal eigenvalue lies on the unit circle. Since $\mathbf{L}\mathbf{1} = 0$ by Lemma 8(ii), the spectrum contains $\lambda_{k^*} = 0$, and Equation 13 yields $d_{k^*} = 1$. In the modal coordinates $\tilde{z} = (P^\top \otimes I_m)z$ the step is $\tilde{z}^{(k)}_{t+1} = d_k\, \tilde{z}^{(k)}_t$, which is diagonal.

For the cost, one step evaluates $\binom{N}{2}$ couplings at $O(N^2)$, eigendecomposes $\mathbf{L}$ at $O(N^3)$, applies $P^\top \otimes I_m$ and $P \otimes I_m$ at $O(N^2 m)$ each, and scales N modes of width m at $O(Nm)$, for $O(N^3 + N^2 m)$ in total. Forming $\mathrm{Cay}(\mathbf{H})$ directly would require inverting an $Nm \times Nm$ matrix at $O(N^3 m^3)$, which exceeds $O(N^3 + N^2 m)$ by a factor that grows as m^3/N once m exceeds N. □

C.4 Layered conservation (Corollary 3)

Proof. For the norm, Lemma 9 makes U unitary, so $\|Uz\|^2 = \|z\|^2$ for every z, and in modal coordinates the same statement is $\sum_k |d_k|^2 \|\tilde{z}^{(k)}\|^2 = \sum_k \|\tilde{z}^{(k)}\|^2$, which requires $|d_k| = 1$ and nothing further. For the sum, the quantity is $p(z) = (\mathbf{1}^\top \otimes I_m)z$, and by Lemma 8(ii) the subspace $\{\mathbf{1} \otimes e : e \in \mathbb{C}^m\}$ lies in the kernel of $\mathbf{H}$, which the Cayley transform maps to the eigenspace of U at eigenvalue one, since $\mathbf{H}v = 0$ implies $Uv = (I + 0)^{-1}(I - 0)v = v$, so $p(Uz) = p(z)$ for every z. For the angular pairing, Lemma 8(iii) puts every $I_N \otimes C$ in the commutant of $\mathbf{H}$, in particular $G = I_N \otimes \mathbf{J}$, and Theorem 1 then yields preservation of $\mathrm{Im}(z^\dagger G z)$. A generator of the form $\mathbf{L} \otimes I_m$ with $\mathbf{L}$ a Laplacian satisfies all three properties, so the rightmost condition of Equation 8 implies the other two. For the converses, two counterexamples suffice, each satisfying the conditions to its left and failing the one stated.

(a) *Unit modulus does not yield the sum.* Take $\mathbf{H}' = \mathrm{diag}(\mu_1, \ldots, \mu_N) \otimes I_m$ with all μ_k real and distinct and no μ_k equal to zero. Then $\mathrm{Cay}(\mathbf{H}')$ is unitary and every modal eigenvalue lies on the unit circle, while $\mathbf{L}'\mathbf{1} \neq 0$, so $\mathbf{1} \otimes e$ is not in the kernel and p is not preserved.

(b) *Vanishing row sums do not yield the angular pairing.* Take $\mathbf{H}'' = \mathbf{L} \otimes B$ with $\mathbf{L}$ a Laplacian and B Hermitian with $[B, \mathbf{J}] \neq 0$. Then $\mathbf{H}''(\mathbf{1} \otimes e) = 0$ for every e, so p is preserved, while $[\mathbf{H}'', I_N \otimes \mathbf{J}] = \mathbf{L} \otimes [B, \mathbf{J}] \neq 0$ and the angular pairing is not.

(c) *The layers are strict.* Counterexample (a) satisfies the leftmost condition and fails the middle one, and counterexample (b) satisfies the middle condition and fails the rightmost, so no two of the three conditions are equivalent.

The three conditions of Equation 8 are therefore distinct, and each is strictly stronger than the one to its left. □

C.5 The invariant family (Corollary 4)

Proof. By Theorem 1, a Hermitian G is held by $U = \mathrm{Cay}(\mathbf{L} \otimes I_m)$ exactly when $[\mathbf{L} \otimes I_m, G] = 0$, and at every $\mathbf{L} \in \mathcal{L}_N$ exactly when this commutator vanishes for every such $\mathbf{L}$. For sufficiency, take $G = I_N \otimes C_1 + \mathbf{1}\mathbf{1}^\top \otimes C_2$, so Lemma 10 yields $[\mathbf{L}, I_N] = [\mathbf{L}, \mathbf{1}\mathbf{1}^\top] = 0$ for every $\mathbf{L} \in \mathcal{L}_N$ and then $[\mathbf{L} \otimes I_m, G] = [\mathbf{L}, I_N] \otimes C_1 + [\mathbf{L}, \mathbf{1}\mathbf{1}^\top] \otimes C_2 = 0$. For necessity, write G in block form as $G = \sum_{a,b} E_{ab} \otimes G_{ab}$ with E_{ab} the $N \times N$ matrix units and $G_{ab} \in \mathbb{C}^{m \times m}$. The commutator $[\mathbf{L} \otimes I_m, G] = \sum_{a,b} [\mathbf{L}, E_{ab}] \otimes G_{ab}$ vanishes for every $\mathbf{L} \in \mathcal{L}_N$ only if, for every fixed C in a basis of $\mathbb{C}^{m \times m}$, the $N \times N$ coefficient matrix A_C obtained by contracting G against C commutes with every $\mathbf{L}$. By Lemma 10 each such A_C lies in $\mathrm{span}\{I_N, \mathbf{1}\mathbf{1}^\top\}$, so $G = I_N \otimes C_1 + \mathbf{1}\mathbf{1}^\top \otimes C_2$ for some C_1, C_2, which is Equation 9. Hermiticity of G forces C_1, C_2 Hermitian. For the dimension, the two matrices I_N and $\mathbf{1}\mathbf{1}^\top$ are linearly independent for $N \geq 2$, and the Hermitian $m \times m$ matrices form a real space of dimension m^2, so the family has real dimension $2m^2$. □

The summed quantity $p(z) = (\mathbf{1}^\top \otimes I_m)z$ is linear rather than quadratic and therefore not a member of Equation 9. It is nonetheless held, through the kernel of $\mathbf{H}$ as in the proof of Corollary 3, and its square $\|p(z)\|^2 = z^\dagger(\mathbf{1}\mathbf{1}^\top \otimes I_m)z$ is a member with $(C_1, C_2) = (0, I_m)$.

C.6 Applicability (Corollary 5)

Proof. With a linear decoder $y = Dz$, any function of y is a function of z, and q is held for every θ exactly when $q \circ D$ is held on the latent for every θ. By Corollary 4, the functions of the latent held for every θ that are quadratic forms are exactly those of Equation 9, which yields the equivalence for quadratic q. If q is not a quadratic form in y, then $q \circ D$ is not a quadratic form in z, since composition with a linear map preserves polynomial degree, so Corollary 4 makes no statement about it. □

C.7 Lower bound for a unitary step (Proposition 6)

Proof. For any vectors u, v, the reverse triangle inequality is $\|u - v\| \geq \big|\|u\| - \|v\|\big|$. Take $u = Uz_t$ and $v = z_{t+1}$, and use $\|Uz_t\| = \|z_t\|$ for unitary U. Equality holds when u and v are non-negative multiples of one another. □

C.8 Conservation of the summed quantity (Proposition 7)

Proof. Write the transition as $U = (Q\,\mathrm{diag}(d)Q^\top) \otimes I_m$ with $Q \in O(N)$ and $d \in (S^1)^N$, and set $M = Q\,\mathrm{diag}(d)Q^\top$. The summed quantity is $p(z) = (\mathbf{1}^\top \otimes I_m)z$, and

$$p(Uz) = (\mathbf{1}^\top \otimes I_m)(M \otimes I_m)z = \big((M^\top \mathbf{1})^\top \otimes I_m\big)z.$$

Requiring $p(Uz) = p(z)$ for every $z \in \mathbb{C}^{Nm}$ therefore requires $M^\top \mathbf{1} = \mathbf{1}$. Substituting $M^\top = Q\,\mathrm{diag}(\bar{d})Q^\top$ and multiplying on the left by $Q^\top$, this is $\mathrm{diag}(\bar{d})\,Q^\top \mathbf{1} = Q^\top \mathbf{1}$, that is $(\bar{d}_k - 1)c_k = 0$ for every k with $c = Q^\top \mathbf{1}/\sqrt{N}$, which is Equation 12 since $|d_k| = 1$ makes $\bar{d}_k - 1 = 0$ exactly when $d_k - 1 = 0$. The condition holds when every k with $c_k \neq 0$ has $d_k = 1$. For the assembly, $\mathbf{L}\mathbf{1} = 0$ makes $\mathbf{1}/\sqrt{N}$ the k^*th column of Q, so $c = e_{k^*}$, and $\lambda_{k^*} = 0$ makes $d_{k^*} = 1$, so the condition holds with every other d_k free. For a transition with $d_k = 1$ for every k, the condition holds for every c, and $M = I_N$. □

Remark 11 (The converse of Theorem 2). The forward direction of Theorem 2 is unconditional. The converse is not. Any unitary U without -1 in its spectrum can be written as $\mathrm{Cay}(\mathbf{H})$ for the Hermitian $\mathbf{H} = 2\mathrm{i}(U - I)(U + I)^{-1}$, and in the eigenbasis of that $\mathbf{H}$ the step is diagonal with unit-modulus entries. The summed quantity is preserved only when $\mathbf{H}$ additionally has vanishing row sums in the coordinates of p, which is a condition on the pair $(\mathbf{H}, p)$ and not on U alone.

C.9 Numerical properties

The guarantees above are exact in real arithmetic, and in floating-point the two invariants behave differently.

Proposition 12 (Floating-point floor)**.** *Let ϵ denote the unit round-off of the arithmetic used for the rollout, and let $\delta = |\lambda_{k^*}|$ be the residual the eigensolver returns in place of the exact zero eigenvalue.*

After T steps of the modal recurrence,

$$\textit{norm:} \qquad \big|\, \|z_T\|^2 - \|z_0\|^2 \,\big| \;\le\; 2TN\epsilon\, \|z_0\|^2 \;+\; O(T^2\epsilon^2), \tag{14}$$

$$\textit{sum:} \qquad \big\|\, p(z_T) - p(z_0) \,\big\| \;\le\; T\big(\delta + 2\epsilon\big)\, \|p(z_0)\| \;+\; O\big(T^2(\delta+\epsilon)^2\big). \tag{15}$$

The two bounds differ in their governing term. Equation 14 is governed by the round-off ϵ alone, and Equation 15 is governed by the eigensolver residual δ, which exceeds ϵ whenever the eigendecomposition is the less accurate of the two operations.

Proof sketch. For the norm, each step multiplies N modes by numbers of modulus one, and a floating-point product with a unit-modulus number introduces a relative error of at most ϵ per real multiplication. Summing over modes and steps and discarding second-order terms yields Equation 14. The bound does not grow with the accuracy of the eigendecomposition, since the computed $\mathrm{Cay}(\mathbf{L})$ is unitary to round-off regardless P and Λ were returned.

For the sum, the exact recurrence leaves p fixed because $d_{k^*} = 1$. The computed phase is $\hat{d}_{k^*} = (1 - \mathrm{i}\delta/2)/(1 + \mathrm{i}\delta/2)$, whose argument is $-2\arctan(\delta/2)$, so one step rotates the mode carrying p by an angle of magnitude at most δ and in addition incurs the round-off of the two modal transforms, bounded by 2ϵ. The errors compose over T steps, which yields Equation 15. □

The two invariants are not affected in the same way, and the asymmetry is visible in the measured drift. Norm preservation is a property of the map at each step, since $\mathrm{Cay}(\mathbf{H})$ is unitary whatever numerical error entered $\mathbf{L}$, so the norm error does not accumulate and remains at the round-off level for the whole rollout. Preservation of the sum requires λ_{k^*} to come out of the eigendecomposition as exactly zero. A numerical eigensolver returns a residual δ that is small but non-zero, the phase d_{k^*} then differs from one by $O(\delta)$, and Equation 15 accumulates the difference linearly in T. On the visual N-body system the energy residual is flat across the rollout, at $1.4\mathrm{e}{-6}$ percent at step 5 and $2.8\mathrm{e}{-6}$ at step 99, while the momentum residual grows from $6.4\mathrm{e}{-6}$ to $2.0\mathrm{e}{-5}$ over the same interval. Both lie at the float32 floor, and neither is comparable with a residual measured at a different horizon. We eigendecompose $\mathbf{L}$ in float64 and evaluate the rollout in float32, so the reported residuals are float32 quantities. The reference trajectories for the cross-domain systems are integrated in float64, since a float32 integration of the Lennard-Jones force overflows and returns a reference that is itself unusable.

C.10 Batched form

The assembly is applied independently to every element of a batch, so the eigendecomposition is batched over the leading dimension. With batch size B the step costs $O(BN^3 + BN^2m)$, and the $N \times N$ eigendecomposition is the only part that does not vectorise over m. For the sizes used here, N from 3 to 64 and m from 8 to 40, the modal transforms dominate the wall clock and the eigendecomposition contributes an overhead that grows with N and is independent of m. The diagonal of $\mathbf{L}$ is formed by summing the off-diagonal entries of its own row rather than by a separate parameter, so $\mathbf{L}\mathbf{1} = 0$ holds to the accuracy of that sum and not to the accuracy of a learned value. The modal transforms use the same P in both directions, computed once per step, so the round trip $(P \otimes I_m)(P^\top \otimes I_m)$ is the identity up to round-off rather than up to the agreement of two separate factorisations.

C.11 Full algorithm of the Kronecker–Cayley transition

Algorithm 2 The Kronecker–Cayley transition, training and rollout.

Require: observations $x_{0:T}$ with actions $a_{0:T-1}$, objects N, subspace width m, encoder E_ϕ, decoder D_ψ, coupling network MLP_θ, readout $r \in \mathbb{R}^N$, horizon T, batch B, learning rate η

Ensure: parameters (θ, ϕ, ψ) and a rollout whose conserved family $\mathcal{Q}$ of Corollary 4 is the same at every value of θ

Fixed before training.

$r \leftarrow \mathbf{1}$, so that $r^\top \otimes I_m$ reads the latent zero mode ▷ tied readout, Table 5

Training.

for each minibatch $(x_{0:T}, a_{0:T-1})$ of size B **do**

  $\hat{x}_{1:T} \leftarrow \textsc{Rollout}(x_0, a_{0:T-1})$

  $\mathcal{L} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \|\hat{x}_t - x_t\|^2$ ▷ reconstruction over the rollout

  $(\theta, \phi, \psi) \leftarrow \mathrm{Adam}\big((\theta, \phi, \psi),\ \nabla\mathcal{L},\ \eta\big)$

end for

Rollout.

function $\textsc{Rollout}(x_0, a_{0:T-1})$

  $z_0 \leftarrow E_\phi(x_0) \in \mathbb{C}^{B\times Nm}$, split as $(z_{0,1}, \dots, z_{0,N})$ with $z_{0,i} \in \mathbb{C}^{B\times m}$

  for $t = 0, \dots, T-1$ **do**

    $w_{ij} \leftarrow \mathrm{MLP}_\theta\big(\|z_{t,i} - z_{t,j}\|^2,\ a_t\big)$ for $i < j$, $\quad w_{ji} \leftarrow w_{ij}$ ▷ Eq. 4, the only learned line

    $\mathbf{L}_{ij} \leftarrow -w_{ij}$ for $i \neq j$, $\quad \mathbf{L}_{ii} \leftarrow \sum_{k\neq i} w_{ik}$ ▷ Eq. 5, $\mathbf{L}\mathbf{1} = 0$ by the row sum

    $(\Lambda, P) \leftarrow \textsc{SymEig}(\mathbf{L})$ in float64 ▷ $\mathbf{L} = P\Lambda P^\top$, rollout in float32

    $d_k \leftarrow (1 - \mathrm{i}\lambda_k/2)/(1 + \mathrm{i}\lambda_k/2),\ k = 1, \dots, N$ ▷ Eq. 7, $|d_k| = 1$ and $d_{k^*} = 1$

    $\tilde{z} \leftarrow (P^\top \otimes I_m)\, z_t$ ▷ into the modal basis

    $\tilde{z}^{(k)} \leftarrow d_k\, \tilde{z}^{(k)},\ k = 1, \dots, N$ ▷ Thm. 2, one phase per mode

    $z_{t+1} \leftarrow (P \otimes I_m)\, \tilde{z}$ ▷ the P of line 12, identity to round-off

    $\hat{x}_{t+1} \leftarrow D_\psi(z_{t+1})$

  end for

  return $\hat{x}_{1:T}$

end function

D Additional Experiments

D.1 Q1: Mass Is Held Where Every Diagonal Baseline Loses It

Table 2 summarises three seeds. Table 6 gives every seed separately, so that the spread across seeds can be read against the gap between rows. Table 7 adds the drift of the latent zero mode, which is held at the same order whether the decoder is tied or free, and locates the loss of physical mass at the read-out. Table 8 places every model against persistence and the constant field with the spatial mean removed, which is the evidence for the accuracy cap discussed in §D.6. Figure 7 then follows the drift and the prediction error per rollout step, which shows that the residual of KC accumulates smoothly rather than appearing at any one step.

Table 6: **Mass drift and nRMSE at every seed.** Final-horizon values behind Table 2. A dash marks a drift the rollout does not define, CPFNO on the dam break at seeds 43 and 44, where it leaves the axis at steps 48 and 62 and its drift at $T = 100$ is therefore read before divergence in Table 9 instead.

	Mass drift $\lvert\Delta M\rvert$			**nRMSE**		
Model	seed 42	seed 43	seed 44	seed 42	seed 43	seed 44
Dam break						
level B	27.1	28.1	28	0.221	0.23	0.227
level C	30.2	61.9	39.4	0.223	0.228	0.229
free	956	93.8	261	0.166	0.149	0.155
LRU	3.4e3	2.1e3	1.6e3	1.71	0.519	0.743
FNO	3.6e6	2.3e2	5.4e11	1.8e3	0.105	4.1e7
CPFNO, uncorrected	4.4e18	—	—	∞	∞	∞
CPFNO, corrected	1.3e6	—	—	∞	∞	∞
KC, free decoder	32	403	117	0.226	0.215	0.221
KC	9.3e−3	1.2e−3	0.0498	0.114	0.208	0.114
Advection						
level B	457	492	483	1.09	1.11	1.11
level C	376	357	510	0.998	0.926	1.14
free	343	456	477	1	1.01	1
LRU	3.1e3	1.4e3	1.4e3	1.91	1.49	1.82
FNO	5.8e4	898	1.5e4	135	2.52	52.1
CPFNO, uncorrected	398	395	395	1.3	1.89	2.36
CPFNO, corrected	3.3e−4	2.6e−4	2.6e−4	1.3	1.89	2.36
KC, free decoder	6.2	5.75	10.7	0.832	0.507	0.843
KC	3.0e−5	0.0222	0.0256	0.924	1.9	2.55
Allen–Cahn						
level B	4.14	6.49	3.86	0.818	0.819	0.803
level C	4.65	2.9	2.6	0.809	0.816	0.807
free	39.5	66.3	59.8	0.724	0.69	0.708
LRU	17.9	18.7	19.3	1.06	1.06	1.1
FNO	66.5	55	62.2	1.32	2.04	3.24
CPFNO, uncorrected	1.2e3	5	806	1.66	1.83	2
CPFNO, corrected	6.4e−4	6.3e−5	4.6e−4	1.66	1.83	2
KC, free decoder	25.7	52.3	27.8	1.94	2.35	1.92
KC	4.4e−6	1.8e−5	1.0e−4	1.02	0.84	1.02
Sphere						
level B	5.9e4	3.7e4	9.0e3	1.75	1.65	1.65
level C	1.3e4	5.1e4	2.1e4	1.6	1.57	1.68
free	6.5e10	2.5e10	4.1e12	2.1e6	2.1e6	1.2e8
LRU	3.3e4	8.6e4	1.1e5	2.51	3.5	3.02
KC, free decoder	561	1.2e3	703	1.03	1.11	1.23
KC	337	442	759	1.05	1.03	1.08

Table 7: **Latent zero-mode drift at every seed.** The invariant the operator holds, before the decoder. It stays between 10^{-6} and 10^{-2} under both read-outs, whereas the physical mass separates them by three to six orders of magnitude on the three flat benchmarks.

	KC			**KC, free decoder**		
Benchmark	seed 42	seed 43	seed 44	seed 42	seed 43	seed 44
Dam break	7.8e−3	8.0e−5	7.1e−3	6.7e−3	2.1e−5	1.7e−3
Advection	4.5e−5	5.3e−3	6.7e−3	1.4e−4	2.3e−4	7.1e−4
Allen–Cahn	1.7e−5	3.2e−6	4.5e−5	5.1e−5	2.8e−6	7.4e−5
Sphere	8.3e−4	4.8e−4	1.5e−3	7.4e−4	6.7e−4	9.7e−4

The spherical benchmark. The PDEArena release distributes a 25-trajectory subset whose pressure field is identically zero at frame 0, so a model taking that frame as input receives the same field on every trajectory. Every model we trained under that protocol, diagonal or Fourier, collapsed to a single latent, and its mass drift of exactly zero came from a constant output. We therefore re-pose the benchmark as recorded in Table 5, and verify that no model collapses by checking that the maximum pairwise latent distance across five test initial frames exceeds 10^{-4} of the mean latent norm for every baseline. The physical mass on a sphere is the $\cos(\text{lat})$-weighted integral, so a readout that assigns

Table 8: **Trivial baselines and the evolving-field error, seed 42.** Persistence outputs the first decoded frame at every step and the constant field is its spatial mean. nRMSE_{ev} removes the spatial mean, so a constant field scores 1 by construction.

	Dam-break		Advection		Allen–Cahn	
Model	nRMSE	nRMSE_{ev}	nRMSE	nRMSE_{ev}	nRMSE	nRMSE_{ev}
Trivial baselines						
persistence	0.241	2.32	0.006	0.009	4.81	4.82
constant field	0.103	1.00	0.702	1.00	0.999	1.00
Diagonal recurrences						
level B	0.221	2.13	1.09	1.44	0.818	0.818
free	0.166	1.51	1.00	**1.00**	**0.724**	**0.724**
LRU	1.71	17.6	1.91	4.01	1.06	1.06
Neural operators						
FNO	1784	1.8e4	3.44	4.18	1.32	1.32
CPFNO	∞	∞	1.78	3.63	1.66	1.66
KC	**0.114**	**1.10**	**0.924**	1.23	1.02	1.02

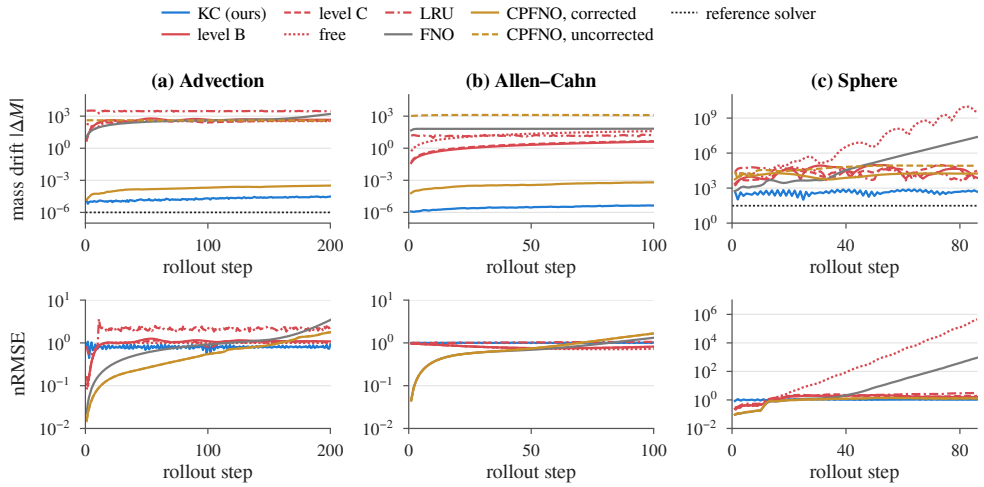


Figure 7: **Drift and prediction error per rollout step.** Advection, Allen–Cahn and the sphere, seeds 42 to 44, with the band spanning the three. The residual of KC rises smoothly from the first step, the signature of accumulated round-off rather than of a learned bias, and each corrector holds mass only while its uncorrected prediction stays bounded. FNO and CPFNO are absent from the sphere panel, where the released trainer does not run, so that panel compares the diagonal transitions alone.

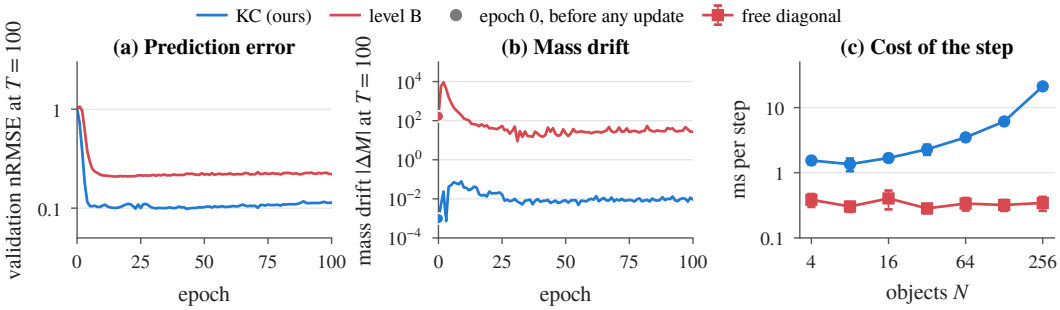


Figure 8: **Training and cost.** Dam break, seed 42. **(a)** Validation nRMSE and **(b)** mass drift at $T = 100$ after every epoch, where the dot marks the untrained model. KC starts at 10^{-3} before the first update and stays below 10^{-1}, whereas level B reaches 10^{4} at epoch 1 and settles at 27. **(c)** Wall-clock per rollout step against N: 3.5 ms at $N = 64$ and 21 ms at $N = 256$, against 0.3 ms for a free diagonal.

the unweighted patch sum to the zero mode holds a different quantity: at seed 42 it holds that mode to 4.7×10^{-4} while the physical mass drifts by 2.4×10^{3}, which is Corollary 5 in its negative direction. Scaling each patch by the inverse of its mean $\cos(\text{lat})$ instead makes the zero mode the area-weighted integral at patch resolution, held to 1.2×10^{-2} against a reference of 7.9×10^{3}. Under the per-pixel metric of Table 2, which weights every grid point by its own $\cos(\text{lat})$, the residual is 483 over three seeds against a reference of 30.8, the variation of $\cos(\text{lat})$ across a patch spanning 22 degrees of latitude amplified near the poles by the $1/\cos$ scaling. CPFNO's corrector targets the unweighted sum, so its corrected output drifts by 1.8×10^{4} under that metric.

D.2 Q2: Correction at the Output Does Not Replace the Condition

Figure 3 shows the median over five dam-break seeds. The per-seed values behind it are given in Table 9 for KC, CPFNO and FNO, and in Table 10 for FNO with the fixed uniform corrector, where the last column also re-targets CPFNO's corrector on the sphere to the area-weighted integral so that both correctors and KC enforce the same quantity. Table 11 then carries the two neural operators to the remaining three benchmarks, and Figures 9 and 10 show the decoded fields on two of them. In every case the corrector holds the quantity only while the uncorrected prediction stays bounded, which is the conclusion of §4.2 on every benchmark reported here.

Table 9: **Dam-break shallow water across five seeds,** $T = 100$**.** Mass drift, nRMSE, and the first step at which nRMSE exceeds 100. For CPFNO the drift is reported at $t \leq 50$, before the median seed diverges.

Metric	**s42**	**s43**	**s44**	**s45**	**s46**
CPFNO					
corrected drift, $t \leq 50$	2e−3	2e−3	2e−3	2e−3	5.5e−3
uncorrected drift, $t \leq 50$	1.6e4	1.6e4	1.6e4	1e4	1.1e4
nRMSE	∞	∞	∞	∞	**0.103**
diverges at	83	48	62	65	never
FNO					
mass drift	3.6e6	2.3e2	5.4e11	4.6e7	1.6e17
nRMSE	1783	**0.105**	4.1e7	5.7e3	3.1e13
diverges at	69	never	94	88	83
KC, mass drift	**9.3e−3**	**1.2e−3**	**5.0e−2**	**2.0e−2**	**2.6e−2**
KC, nRMSE	**0.114**	0.208	**0.114**	**0.115**	0.115
KC, diverges at	never	never	never	never	never

Table 10: **FNO with a fixed uniform corrector.** The corrector adds the missing mass uniformly over the grid at inference, without training. Dam-break across five seeds at $T = 100$, the other three benchmarks at seed 42. The last column is CPFNO on the sphere with a corrector that targets the $\cos(\text{lat})$-weighted integral.

Metric	**s42**	**s43**	**s44**	**s45**	**s46**
Dam-break, five seeds					
uncorrected drift	6.7e5	4.0e14	3.1e6	2.7e24	1.3e18
corrected drift	6.0e−2	1.6e7	7.6e−2	1.2e17	2.2e10
nRMSE	403	3.9e10	357	∞	1.1e14
diverges at	95	58	97	60	73

	Advection	**Allen–Cahn**	**Sphere**	**Sphere, CPFNO**
Other benchmarks, seed 42				
uncorrected drift	451	7.5e3	4.1e4	1.8e4
corrected drift	2.3e−4	2.2e−3	4.0e−2	1.3e−2
nRMSE	2.10	1.81	1.72	1.55

Table 11: **CPFNO and FNO on the remaining three benchmarks, seed 42.** Advection values are deterministic.

Metric	Advection	Allen–Cahn	Sphere
CPFNO			
corrected drift	3.2e−4	held	1.8e4
uncorrected drift	391	1290	8.3e4
nRMSE	1.78	1.66	1.29
FNO			
mass drift	1.6e3	66.5	2.4e7
nRMSE	3.44	1.32	diverges

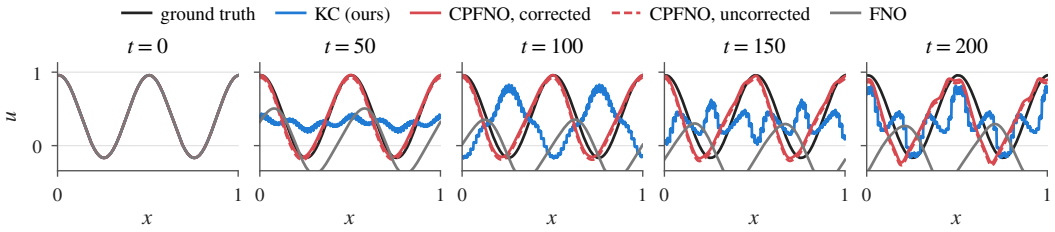


Figure 9: **Advection, field at five steps, seed 42.** Ground truth, KC, CPFNO before and after its corrector, and FNO. KC holds the integral of u to 3×10^{-5} while its phase lags the reference, the cap of the per-patch encoder, whereas FNO loses amplitude and mass.

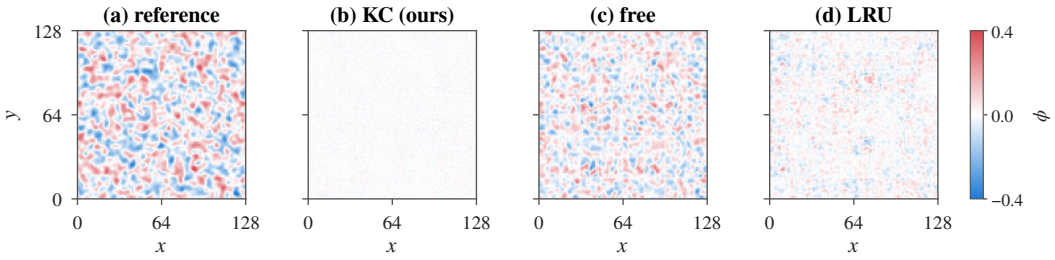


Figure 10: **Allen–Cahn, field at the final horizon, seed 42.** Reference, KC, free diagonal and LRU on one colour scale. KC holds mass to 4×10^{-6} with a near-constant field, whereas the free diagonal and LRU keep more texture while losing mass by 39.5 and 17.9.

D.3 Q3: Each Condition Holds Exactly One Quantity

Figure 4 and Table 3 report the three controls as means over three seeds. The per-seed values of the four levels are given in Table 12, followed per step in Figure 11 and decoded into frames in Figure 12. The matched perturbation is tabulated at every ε in Table 13, and Table 14 isolates the read-out direction by holding the operator and its spectrum fixed.

Table 12: **Four levels of the construction, per seed.** Visual N-body, drift at step 99 for seeds 42/43/44. Levels A to D are the spectrum free, unit modulus on every eigenvalue, B with one mode per object fixed at unity, and the full assembly.

Level	Condition present			Drift at step 99 (%), seeds 42 / 43 / 44		
	$\lvert d_k\rvert=1$	$\mathbf{L1}=0$	$\mathbf{L}\otimes I_m$	energy	momentum	angular
A	✗	✗	✗	100, 100, 100	100, 100, 100	100, 100, 100
B	✓	✗	✗	4.0e−6, 4.1e−6, 5.1e−6	145, 142, 135	411, 225, 222
C	✓	✗	✗	4.1e−6, 4.5e−6, 4.5e−6	138, 138, 140	1060, 786, 412
D (KC)	✓	✓	✓	**2.7e−6, 2.4e−6, 3.0e−6**	**1.7e−5, 1.6e−5, 1.9e−5**	**7.8e−5, 6.1e−5, 2.2e−4**

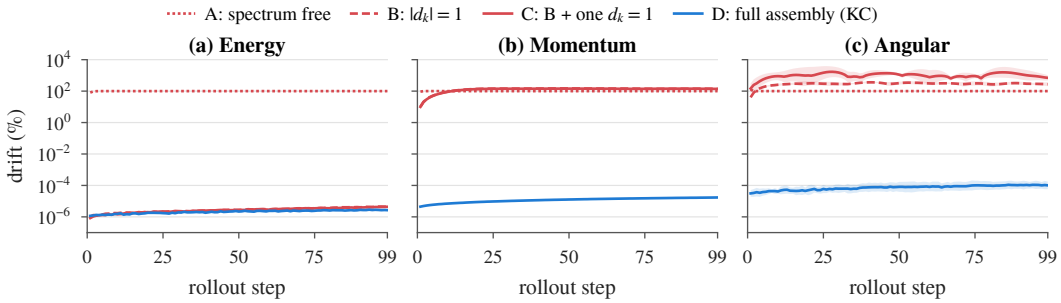


Figure 11: **Drift per rollout step under the four levels.** Visual N-body, seeds 42/43/44, geometric mean with one standard deviation in log space shaded. Energy is held from level B onward, momentum and the angular pairing only by the full assembly.

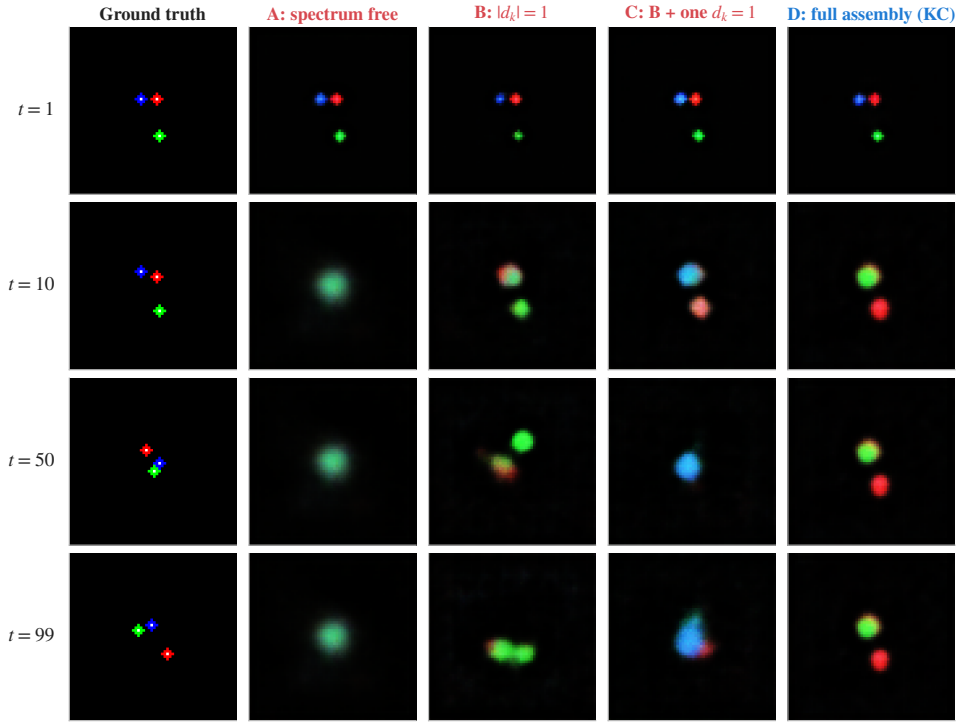


Figure 12: **Decoded frames of the visual N-body rollout under the four levels.** Test trajectory 0, seed 42, at four steps. None of the four follows the reference positions under the shared encoder (§D.6). Level A collapses to a single blob by $t = 10$, levels B and C merge objects and exchange their colours, and the full assembly keeps separate, colour-stable objects over the full horizon.

Table 13: **Matched perturbation, visual N-body, step 25, seeds 42 to 44.** One symmetric draw of scale ε, applied once with the diagonal corrected so that the row sums stay zero and once uncorrected, with the relative Frobenius change beside each.

	row sums kept			row sums broken		
ε	$\|\Delta\mathbf{L}\|_F/\|\mathbf{L}\|_F$	momentum	energy	$\|\Delta\mathbf{L}\|_F/\|\mathbf{L}\|_F$	momentum	energy
0	0	2.3e−5	1.7e−6	0	2.3e−5	1.7e−6
0.01	4.4e−3	3.3e−5	1.8e−6	4.6e−3	4.06	1.9e−6
0.05	2.2e−2	3.3e−5	1.7e−6	2.3e−2	20.2	1.7e−6
0.1	4.4e−2	3.3e−5	1.8e−6	4.6e−2	39.5	1.9e−6
0.2	8.7e−2	3.2e−5	1.8e−6	9.1e−2	72.7	1.9e−6
0.5	2.2e−1	3.5e−5	1.8e−6	2.3e−1	109.7	2.0e−6

Table 14: **Same operator, different read-out, 2000 steps, seeds 42/43/44.** Maximum drift in percent when the summed quantity is read along the null-mode eigenvector against the eigenvector of largest $|\lambda|$, with the operator and its spectrum held fixed.

System	null mode	non-null mode
synthetic graph, $N = 3$	$\mathbf{3.98e{-3} \pm 3.95e{-3}}$	$200.0 \pm 2.4e{-5}$
molecular dynamics, $N = 16$	$\mathbf{3.32e{-4} \pm 2.85e{-4}}$	247.4 ± 43.8

D.4 Q4: Training Reaches the Condition Only by Freezing the Spectrum

Figure 5 summarises sixteen trained seeds. The angle of each basis to the uniform vector before and after training is given in Table 15 for the near-identity initialisation and in Table 16 for the Haar-random one. Table 17 then reports the spectral activity of every seed and, for three representative seeds, the two factors of Equation 12, which shows that the two conserving seeds satisfy the condition by driving every d_k to 1. Tables 18 and 19 give the penalty sweep and the sample-size sweep behind Figures 5b and 5c, where no penalty weight reaches the condition and KC holds it at every training-set size.

Table 15: **Angle of a learned basis to the uniform vector, eight seeds.** φ at initialisation and after training under a conservation penalty at $\lambda = 10$, with momentum drift at step 25. The basis is initialised near the identity, at which φ takes its maximum $\arccos(1/\sqrt{3}) = 54.74°$.

seed	42	43	44	45	46	47	48	49	assembled
φ, init	54.1°	54.2°	54.3°	54.5°	53.6°	54.3°	53.2°	54.3°	—
φ, final	44.9°	48.1°	54.7°	54.7°	54.7°	54.7°	54.7°	54.7°	**0.01°**
momentum drift	159	932	174	43.9	174	44.0	174	44.4	$\mathbf{10^{-5}}$

Table 16: **The same test under Haar-random initialisation, eight seeds.** φ at initialisation and after training at $\lambda = 10$, with momentum drift at step 25. Seeds 47 and 49 conserve, by freezing the spectrum (Table 17).

seed	42	43	44	45	46	47	48	49
φ, init	26.3°	41.8°	42.2°	44.8°	39.1°	35.4°	36.4°	37.0°
φ, final	29.3°	42.0°	42.1°	44.9°	37.0°	32.5°	37.5°	35.6°
momentum drift	137	373	401	329	127	**0.090**	88.3	0.196

Table 17: **Spectral activity and the conservation condition.** Above, mean $|\lambda_k|$ over the modes that do not carry the uniform vector, for all sixteen trained seeds. Below, the two factors of Equation 12 for three representative seeds. The conserving seeds satisfy the condition by driving every d_k to 1.

init	seed	mean $\lvert\lambda_k\rvert$	drift (%)
Free orthogonal basis, trained under the penalty			
Haar	42, 43, 44	2.85, 2.83, 2.85	137, 373, 401
Haar	45, 46, 48	1.13, 1.80, 1.74	329, 127, 88.3
Haar	47, 49	2.8e−5, 1.3e−4	0.090, 0.196
near-identity	42, 43	2.18, 1.83	159, 932
near-identity	44, 46, 48	1.59, 1.59, 1.59	174, 174, 174
near-identity	45, 47, 49	0.77, 0.77, 0.77	43.9, 43.9, 44.5
assembled	42	**2.87**	**3.5e−5**

	d_k			c_k			$\lvert(d_k - 1)c_k\rvert$		
Haar 47	1	1	1	0.46	0.27	−0.84	10^{-5}	10^{-5}	10^{-5}
Haar 42	$e^{1.92i}$	$e^{1.92i}$	$e^{0.74i}$	0.35	−0.35	0.87	0.57	0.57	0.63
assembled	$e^{1.92i}$	$e^{1.92i}$	1	0	0	1	0	0	0

Table 18: **Penalty sweep.** Momentum drift at step 25, per seed.

λ	42	43	44	PSNR (dB)
0	**86.5**	**99.2**	**42.0**	18.11
0.1	173.7	388.8	173.6	**19.37**
1	177.4	436.5	173.6	19.35
10	159.3	932.0	173.6	19.36

Table 19: **Learned conservation at convergence.** Momentum drift at step 25 against training-set size.

n_{train}	KC	HNN	GNN	PSNR gap
50	**1.29e−5**	16.9	247	−0.31
100	**1.30e−5**	12.1	247	−0.26
250	**1.30e−5**	6.46	246	−0.22
500	**1.30e−5**	2.63	246	−0.12
1000	**1.38e−5**	1.49	248	−0.08
5000	**1.41e−5**	1.52	248	−0.10

D.5 Q5: The Boundary Is Fixed Before Training

Table 20 reports linear and angular momentum on the three N-body systems and Table 21 the four operator-level systems from other fields, the set behind Figure 6 of the main text. The decoded total momentum of one charged test trajectory is drawn per transition in Figure 13. The reference floor of each system is given in Table 22, which shows that the integrator itself conserves energy only on the spring system, and Table 23 therefore adds the energy column on that system alone. Table 24 then tests the quantity boundary inside the latent, by comparing the zero mode of the family with five random Hermitian forms outside it under the same trained operator, and Table 25 tests the dynamics boundary on a stiff system that relaxes to equilibrium, where Proposition 6 predicts that a unitary step fits less accurately than one that can contract. Finally, Table 26 reports the spring benchmark of Thangamuthu et al. [2022] on its own three metrics.

Table 20: N**-body conservation at** $T = 1000$**, seeds** 42 **to** 44**.** Drift in absolute units, geometric mean over the seeds that stayed finite. An asterisk marks a row where the free diagonal left the axis on one seed.

System	B	C	free	LRU	KC-fd	KC
Linear momentum						
charged	3.29	2.66	1.9e4	6.41	3.14	**2.2e−6**
gravity	4.06	3.92	4.7e3*	4.70	7.28	**3.6e−6**
spring	1.25	1.36	5.64	3.79	1.49	**1.6e−6**
Angular momentum						
charged	10.4	9.10	20.0*	26.1	13.7	**1.22**
gravity	17.4	15.7	5.7e7*	22.8	20.0	**1.31**
spring	0.978	1.77	6.06	4.40	1.60	**0.327**
Energy						
spring	0.323	0.369	12.6	1.53	0.927	**0.124**

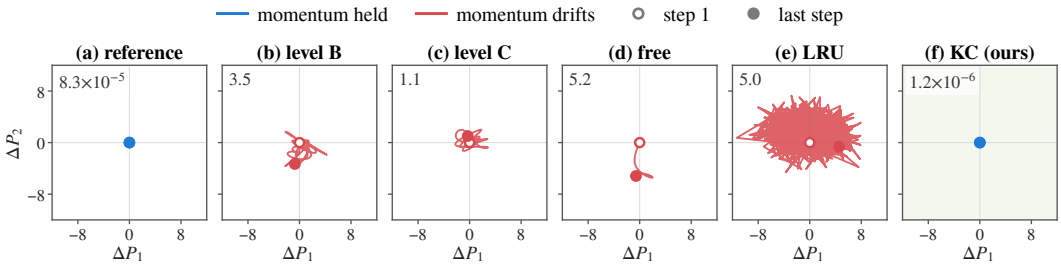


Figure 13: **Total linear momentum of the charged** N**-body rollout, one test trajectory, seed** 42**.** Each path is $P(t) - P(1)$ over 1000 steps in the plane carrying most of its own variation, so distance from the origin is the drift. Each panel states $\|P(1000) - P(1)\|$ against $\|P(1)\| = 1.20$. The momentum stays at a point under the reference integrator and the full assembly, and leaves it under every transition that constrains the spectrum alone.

Table 21: **Transfer to four dynamical systems, operator level,** 2000 **steps, seeds** 42/43/44**.** Time-integrated mean drift of the conserved quantity in percent, with the row sums of the Laplacian intact and broken by a matched perturbation (Figure 6).

System	Conserved quantity	intact↓	broken
molecular dynamics, N=16	total momentum	**1.4e−4**	100.5 ± 0.5
reaction network, N=10	total mass	**6.3e−3**	126.9 ± 0.3
pharmacokinetics, N=4	total drug amount	**1.0e−3**	127.8 ± 0.8
power-grid swing, N=12	angular momentum	**5.3e−3**	127.0 ± 0.3

Table 22: **Reference conservation on the three N-body systems, 20 runs.** Absolute drift over the full trajectory, with the initial magnitude in parentheses.

System	Energy	Linear momentum	Angular momentum
Charged, $N = 5$, 3D	1.26 (0.69)	1.7e−4 (1.05)	0.018 (1.47)
Gravity, $N = 10$, 3D	7.2 (13.0)	4.0e−14 (≈ 0)	7.9e−14 (6.58)
Spring, $N = 5$, 2D	9.8e−5 (0.63)	3.4e−15 (1.09)	2.4e−14 (0.43)

Table 23: **N-body drift at $T = 1000$, every seed.** Values behind Table 20. A dash in the first two blocks marks a seed on which the free diagonal left the axis, and energy is reported on the spring system alone, the one whose reference conserves it.

	charged			gravity			spring		
Model	s42	s43	s44	s42	s43	s44	s42	s43	s44
Linear momentum									
level B	3.32	3.13	3.42	4.11	4.26	3.81	1.21	1.38	1.16
level C	3.13	2.31	2.61	3.83	4.43	3.54	1.03	1.66	1.46
free	3.56	4.00	5.1e11	—	3.30	6.7e6	3.47	7.03	7.38
LRU	7.61	6.24	5.55	4.15	4.78	5.22	3.93	4.44	3.13
KC, free decoder	2.27	4.31	3.15	6.25	7.74	7.96	0.514	2.69	2.38
KC	1.9e−6	2.1e−6	2.6e−6	3.0e−6	4.3e−6	3.5e−6	8.6e−7	2.5e−6	1.8e−6
Angular momentum									
level B	11.5	8.84	10.9	14.8	19.9	18.0	0.867	1.26	0.859
level C	10.4	9.28	7.82	18.3	11.6	18.4	1.43	1.36	2.89
free	16.7	24.0	—	—	32.8	9.7e13	3.44	2.85	22.7
LRU	24.7	28.0	25.7	15.6	16.0	47.1	2.72	6.20	5.05
KC, free decoder	18.6	14.0	9.89	16.3	22.1	22.3	0.741	2.60	2.12
KC	1.77	1.37	0.753	3.53	0.757	0.846	0.429	0.309	0.263
Energy, spring system only									
level B			—				0.259	0.431	0.300
level C			—				0.305	0.404	0.409
free			—				5.13	17.9	21.8
LRU			—				1.57	1.51	1.52
KC, free decoder			—				0.143	2.78	2.00
KC			—				0.0928	0.222	0.0927

Table 24: **Family and outside-family drift at $T = 1000$, KC, seed 42.** Absolute drift of the latent zero mode, which lies in the family of Corollary 4, beside five random Hermitian forms outside it, under the same trained operator.

Quadratic form	charged	gravity	spring
Outside the family of Corollary 4			
outside form 1	28.2	101.4	12.4
outside form 2	26.6	109.3	14.8
outside form 3	31.5	108.9	10.0
outside form 4	30.6	117.2	12.3
outside form 5	33.5	100.2	13.6
family zero mode	**7.2e−6**	**1.3e−5**	**5.0e−6**

Table 25: **The dynamics boundary on the Robertson reaction network, $T = 50$.** A stiff system that relaxes to equilibrium, on which Proposition 6 bounds the per-step error of any unitary step. Means over seeds 42/43/44. The latent zero mode is held, and the baselines that can contract fit the state more accurately.

Model	state rel-L_2	$\sum y$ drift (%)	latent zero mode (%)
HNN	0.009	<u>0.099</u>	99.9
GNN	**0.002**	0.162	99.9
free	0.006	**0.057**	100.1
LRU	<u>0.003</u>	0.157	100.7
KC	0.128	0.224	**5.6e−5**

D.6 Diagnostics on the Accuracy Cap

On advection, Allen–Cahn and the sphere, prediction error lies near 1.0 for KC and for the free diagonal alike (Table 6), so the cap is not in the transition, and Table 27 rules out the remaining candidates. Table 8 confirms with a mean-removed error that no model, KC included, predicts the evolving field better than a constant on the three flat benchmarks. A wider latent lowers nRMSE on the dam break from 0.114 to 0.095, whereas neither a wider latent nor a nonlinear decoder lowers it below 0.3 on the other two benchmarks, and mass remains at the floor under both, so the conservation guarantee is unaffected by capacity or decoder choice and the accuracy cap lies in the shared per-patch encoder, which maps each patch independently with no spatial context. On the dam break, whose radial dynamics a linear patch readout can represent, the same encoder reaches 0.114, and Figure 3 shows the decoded field at that accuracy, a near-uniform cross-section that holds the total while the true solution spreads into a ring. A context-aware encoder is therefore the natural extension and does not touch the transition.

Table 26: **Spring benchmark of Thangamuthu et al. [2022], seeds 42/43/44.** Momentum and energy drift at step 50, beside the accuracy metric of the benchmark at step 25. GRU and MLP have no modal latent and the summed quantity is undefined for them.

Model	**PSNR@25 (dB)**	**momentum (%)**	**energy (%)**
$N = 3$			
Hamiltonian	**21.07**	1.09	0.16–0.25
graph network	20.78	494	290
GRU	20.68	—	—
MLP	20.87	—	—
KC	20.90	**1.8e−5**	**10^{-6}**
$N = 5$			
Hamiltonian	**18.11**	1.16	0.16–0.25
graph network	18.06	491	290
GRU	18.09	—	—
MLP	18.08	—	—
KC	18.10	**1.9e−5**	**10^{-6}**

Table 27: **Two diagnostics on the accuracy cap, KC, seed 42.** The latent width is raised from $m = 8$ to $m = 32$, and the linear decoder is replaced by one that is linear only on the zero mode and an MLP elsewhere.

	linear, $m = 8$		**linear,** $m = 32$		**split,** $m = 8$	
Benchmark	nRMSE	mass	nRMSE	mass	nRMSE	mass
dam-break	0.114	9.3e−3	**0.095**	**5.5e−4**	0.225	9.2e−3
advection	**0.924**	**3.0e−5**	1.33	5.2e−2	1.07	4.3e−5
Allen–Cahn	1.02	**4.4e−6**	**0.870**	7.6e−6	1.31	4.3e−5

E Limitations and Future Directions

E.1 Limitations

Scope of the guarantee. The guarantee of the Kronecker–Cayley transition is exact but narrow, and its limits follow from the same algebra that supplies it. Corollary 5 admits a physical quantity only if it is a quadratic form of the latent under a linear decoder, so a potential of the form $\sum_{i<j} \|r_i - r_j\|^{-1}$ is not held, and energy is conserved only on the spring system (Table 20). Angular momentum is a quadratic form, but the linear tie places only the summed quantity on a held mode, so KC improves it by 3 to 12 times without reaching the floor. A unitary step also cannot contract, and Proposition 6 turns this into a per-step error that no choice of $\mathbf{L}$, θ or decoder removes. On the Robertson network KC fits the state at a relative error of 0.128, against 0.002 to 0.009 for baselines that can contract (Table 25), so dissipative dynamics lie outside the construction. Another limitation is that conservation does not imply accuracy. On advection, Allen–Cahn and the sphere no model, KC included, predicts the evolving field better than a constant at the mass-matched mean, a cap that Appendix D.6 locates in the shared per-patch encoder rather than in the transition.

Read-out, cost and evaluation. The guarantee reaches the output only through the decoder. An unconstrained decoder holds the latent invariant equally tightly yet drifts the physical mass by 7 to 115 (§4.1), and on the sphere the area-weighted read-out still leaves a residual of 483 against a reference of 30.8 from the variation of $\cos(\text{lat})$ within a patch (Appendix D.1). Each step requires one $N \times N$ eigendecomposition, measured at 3.5 ms for $N = 64$ and 21 ms for $N = 256$ against 0.3 ms for a free diagonal (Figure 8), and because the couplings depend on the latent state the rollout is sequential and does not admit a parallel scan. The construction further assumes that the latent splits into a known, fixed number N of object subspaces of equal width. Finally, most trained results rest on three seeds, every reported drift is bounded below by the float32 rollout, and the four systems from molecular dynamics, chemical kinetics, pharmacokinetics and power-grid swing dynamics are tested at the operator level, without an encoder or decoder trained on data from those fields.

E.2 Future Directions

Extending the construction. Proposition 7 requires only $d_{k^*} = 1$ on the mode that carries the uniform vector, so allowing $|d_k| \leq 1$ on the remaining $N - 1$ modes would give a transition that contracts while still holding the summed quantity, at the cost of the norm, and would bring dissipative dynamics within reach. Read-outs that place a chosen bilinear quantity, such as angular momentum, on the conserved family, together with generators for nonquadratic invariants, would widen the set of physical quantities the construction holds. Accuracy under the guarantee can be raised without touching the transition, by replacing the per-patch encoder with one that sees spatial context, and finer patches or a read-out weighted per grid point would shrink the residual on curved domains such as the sphere.

Scaling and evaluation. Sparse coupling graphs and warm-started or low-rank eigenvalue updates between consecutive steps would lower the per-step cost, and couplings computed from the input rather than the state would let the rollout admit a parallel scan. A learned partition of the latent would extend the construction to systems whose objects are not known in advance or change over time. On the evaluation side, training encoders and decoders end to end on molecular dynamics, chemical kinetics, pharmacokinetics and power-grid data would test whether a trained model reaches the operator-level drift there, and larger benchmarks, real observational data and more seeds would test how far the results generalise.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three claims, that the conserved quantities of a diagonal recurrence are layered, that the Kronecker–Cayley transition holds exactly a $2m^2$-dimensional family at every parameter value, and that a trained basis reaches the summed-quantity condition only by freezing its spectrum. These are proved in §3 and Appendix C and tested in §4, and the introduction also states the scope limits, namely that the construction does not represent dissipative dynamics and does not by itself hold nonquadratic quantities.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix E (Limitations and Future Directions) collects the limitations and the direction each suggests: only quadratic quantities are held (Corollary 5), dissipative dynamics fall outside a unitary step (Proposition 6, Table 25), conservation does not imply accuracy under the shared per-patch encoder, the guarantee reaches the output only through a tied decoder, each step costs one $N \times N$ eigendecomposition with a sequential rollout, the latent requires a fixed object partition, and the evaluation rests mostly on three seeds with the cross-domain systems tested at the operator level.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Every theorem, corollary and proposition in §3 states its assumptions (Hermitian generator, symmetric Laplacian with zero row sums, $N \geq 2$, a linear decoder for Corollary 5). Complete proofs are in Appendix C, together with the supporting lemmas and a floating-point bound on the measured drift, and the main text gives a short argument after each result.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix A (Table 5) gives every dataset with its split, the architecture and width of each encoder, decoder and coupling network, the optimiser, learning rate, batch size, epochs, rollout length, numerical precision and seeds, along with the construction of every control and baseline. Appendix C.11 gives the full training and rollout algorithm, and the CPFNO reimplementation and the re-posing of the spherical benchmark are described in Table 5 and Appendix D.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and the scripts that reproduce every table and figure, including the generators for the data we created (Allen–Cahn, the visual N-body simulator and the four operator-level systems), are released at `https://github.com/Jiaqian-Janelle/NeurIPS26-Kronecker-Cayley-Stable-State-Space-Dynamics`. The public benchmarks (PDEBench, PDEArena, the EGNN N-body systems) are downloaded from their original sources, as described in Table 5.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Data splits, hyperparameters, optimiser and training protocols for each dataset family and each control are listed in Table 5 (Appendix A). Shared settings are used across all baselines within a family, FNO is run with its published defaults, and the main-text setup paragraphs of §4 summarise what is needed to read each result.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Every trained model is run over seeds 42–44, the dam-break correction study over five seeds and the alignment study over eight seeds per initialisation, and every per-seed value is tabulated in Appendix D. Figure 3 shows the median with the interquartile range, Figure 4 the mean with min–max bars, Figure 11 the geometric mean with one standard deviation in log space, and Table 21 the mean $\pm$ one standard deviation over three seeds. Drift spans many orders of magnitude, so Table 2 reports a geometric mean over seeds rather than a symmetric error bar, and no formal significance tests are run.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run on NVIDIA RTX PRO 6000 GPUs with 96 GB of memory. The per-step wall-clock cost of the transition against N is reported in Figure 8, and the coupling network has 3,963 parameters, with roughly 15M in the largest shared encoder and decoder. We did not log total GPU hours across the project, including preliminary runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public scientific simulation benchmarks and data we simulated ourselves, involves no human subjects or personal data, and conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational work on the structure of learned recurrent transitions, evaluated on physical simulation benchmarks, with no direct path to a harmful application. One indirect risk, which the paper itself documents in §4.1 and §4.2, is that a surrogate which conserves a quantity exactly can still predict the field poorly, so conservation should not be taken as evidence of accuracy when such models are used in scientific or engineering decisions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released models are small dynamics models trained on simulated physical systems and carry no meaningful risk of misuse, and no data is scraped from the internet.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets are cited where they are used (§4, Table 5): PDEBench [Takamoto et al., 2022] (code MIT, data CC BY 4.0), PDEArena [Gupta and Brandstetter, 2023] (code and shallow-water data MIT), the EGNN N-body generator [Satorras et al., 2021] (MIT), FNO [Li et al., 2021] via the neuraloperator library (MIT), and the spring benchmark protocol of Thangamuthu et al. [2022], which we regenerate from its published description. CPFNO [Liu et al., 2025] is reimplemented from the paper. All assets are used within their licence terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets are the KC code, the Allen–Cahn dataset, the visual N-body simulator and the four operator-level systems. Each is specified in Table 5. No human data is involved, so no consent was required. The release documents each asset with its generation script, seeds and a licence, at `https://github.com/Jiaqian-Janelle/NeurIPS26-Kronecker-Cayley-Stable-State-Space-Dynamics`.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves neither crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves neither crowdsourcing nor research with human subjects, so no IRB approval was required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the method or the experiments; they were used only to edit the writing, which does not require declaration.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.