THE READOUT-GAUGE CONFLICT: ANCHORED EQUATION DISCOVERY WITH LEARNED KOOPMAN CLOSURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned Koopman representations can improve equation discovery, yet their flexibility can make the reported equation depend on the chosen coordinates: an auxiliary closure coordinate may absorb a named library term without changing the predicted vector field, even as the coefficient assigned to that term changes. We call this failure of the scientific readout the *readout-gauge conflict*. The Resolvent-based Tail-Residual Network (ReTRN) addresses this conflict by declaring the candidate equation library before fitting, adapting only auxiliary closure coordinates, and reading the equation from the same generator used to fit the lifted dynamics. We characterize the closure transformations that preserve this anchored readout and train the adaptive closure from sparse ordered observations through multiscale temporal moments, without numerical-derivative targets. Under the stated equation-space and conditioning assumptions, anchoring fixes the coefficient coordinates and controls the direct readout error. Across six systems evaluated with an unabridged common analytic library, ReTRN gives the best result in 11 of 12 primary comparisons, wins all three baseline rollout comparisons on each of 60 independent replicates, and generalizes under frozen configurations or selection rules to seven additional equation families.

1 INTRODUCTION

Equation discovery has two coupled aims: reproducing the dynamics and attaching coefficients to named functions—a cubic stiffness, a bilinear interaction, or a reaction term—that carry scientific meaning. These aims coincide for a fixed basis but can separate under representation learning, because an adaptive coordinate may improve the fitted field while absorbing part of a named term. Two equally predictive models can then report different equations, making the ambiguity scientific rather than merely internal to the predictor.

Koopman-based discovery makes this tension unavoidable: the observables that state the governing equation rarely close the generator. A degree-p vector field can map a degree-k observable to degree $k + p - 1$, so a cubic equation may require degree-five observables after one generator application and still higher degrees thereafter. The equation library may therefore be complete for the vector field yet incomplete as a finite Koopman representation, motivating learned dictionaries that supply the missing closure (Klus et al., 2020; Li et al., 2017; Takeishi et al., 2017; Lusch et al., 2018; Azencot et al., 2020; Pan & Duraisamy, 2020).

Let the scientific hypothesis and the richer adaptive Koopman space be

$$\mathcal{F}_{\mathrm{eq}} = \mathrm{span}(U) \subset \mathcal{F}_{\mathrm{K},\theta} = \mathrm{span}[U, V + r_\theta].$$

Here U is a possibly overcomplete library: in the usual degree-aligned construction it contains the lower-order named candidates, whereas $V + r_\theta$ supplies higher-order observables for Koopman closure. Inactive terms may be removed at readout, but every true mechanism must lie in U to retain an anchored coefficient.

This separation creates a consequential ambiguity. If $f = ub_U + tb_T$, with u a named term and t an auxiliary coordinate, replacing t by $t - au$ preserves the field while moving the coefficient of u to $b_U + ab_T$. Varying a thus yields equally predictive equations whose mechanisms differ in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

magnitude, sign, or support. We call this the *readout-gauge conflict*; unlike generic latent-coordinate non-identifiability (Locatello et al., 2019), it changes the model's scientific claim.

The conflict exposes a structural dilemma. Restricting the lift to U protects the coefficient meanings but asks the equation library to perform a second job for which it was not designed: closing the Koopman action. Learning every coordinate improves closure, but removes the reference frame in which the named coefficients are read. Post-hoc projection defines a different estimator: it can recover the equation when the learned full field is sufficiently accurate and the analytic design is well conditioned, but its error inherits both full-field approximation error and projection conditioning. The relevant question is therefore not merely how to learn a richer lift, but how to enlarge the Koopman space while retaining a direct readout in the declared equation coordinates.

Existing approaches address only one side of the dilemma. Sparse, weak-form, and symbolic regression return named equations without auxiliary Koopman closure (Bongard & Lipson, 2007; Schmidt & Lipson, 2009; Brunton et al., 2016; Fasel et al., 2022; Messenger & Bortz, 2021); fixed-dictionary Koopman estimators preserve an analytic readout but cannot adapt the observable space (Mauroy & Gonçalves, 2020; Meng et al., 2024); and learned Koopman models adapt that space without protecting named generator coefficients. What is missing is an estimator that fits adaptive closure without allowing it to redefine the semantic coordinates.

ReTRN learns closure around the equation coordinates rather than through them: the analytic head U remains fixed, a zero-initialized residual adapts only the auxiliary tail V, and the coefficients of U are read from the generator fitted in that representation rather than reconstructed afterward. This construction characterizes the transformations under which that direct readout is invariant. It does not posit a unique closure decomposition; it defines B_U as a same-fit coefficient estimand conditional on the declared scientific chart and treats equation–closure shears as different estimators.

Fixing the readout determines *what* is estimated but not how to learn closure from sparse observations. ReTRN uses exponentially weighted temporal moments to couple discrepancies across a segment at several time scales without numerical-derivative targets. Anchoring therefore preserves the meaning of the returned equation, while temporal moments improve the Koopman approximation from which it is read.

Contributions.

- We identify the readout-gauge conflict, prove that prediction-equivalent representations need not identify named coefficients, and characterize the largest class of closure reparameterizations that preserves their readout.
- We turn this invariance into an estimator. ReTRN fixes the complete candidate equation library, learns only its auxiliary Koopman closure, and fits that closure through multiscale temporal moments.
- We connect the design to evidence through field-preserving shears, equation–closure interventions, component ablations, and equation recovery on canonical, mechanistic, coupled, controlled, and measured dynamics.

2 Problem Setup and Koopman Readout

2.1 Equation Discovery from Short Trajectories

Consider a continuous-time system

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)), \tag{1}$$

where $\mathbf{x} \in \Omega \subseteq \mathbb{R}^d$, Ω is pre-compact, and $\mathbf{f}$ is locally Lipschitz.

Let $U = [u_1, \ldots, u_h]$ be a prescribed, possibly overcomplete equation library and suppose $\mathbf{f} = UB_\star$. The functions in U are linearly independent on the design domain, so the coefficients $B_\star$ have a fixed meaning before an adaptive representation is learned. We observe $\mathbf{x}_k = \mathbf{x}(kT_s)$ along many independently initialized segments and seek the unknown support and coefficients in U. Different segments are never concatenated into a fictitious long trajectory. Because a finite Koopman approximation may require more functions than the equation itself, we initialize it with additional analytic observables V and later allow only this auxiliary block to adapt.

Definition 2.1 (Equation-anchored readout)**.** Fix a named equation library U before fitting. A learned-lift estimator is *equation-anchored* relative to U if (i) the coordinates in U remain pointwise fixed, (ii) its reported coefficients are taken from the state-output block of the same fitted generator, and (iii), for a fixed fitted lifted field, this readout is invariant under the induced similarity transformations from invertible changes within the complementary closure coordinates.

By making the coefficient map part of the estimator, Definition 2.1 fixes the scientific coordinate system through condition (i), rules out a post-hoc surrogate that may answer a different regression problem through condition (ii), and permits harmless reparameterization of auxiliary closure through condition (iii). The invariance is a same-fit, tail-only coordinate statement; it neither covers equation–closure shears nor asserts invariance after retraining. These requirements also mark the boundary of what a direct generator readout can preserve, as the next section shows.

2.2 Koopman Generator Estimation

Let $\mathbf{\Phi}^t$ be the flow of Eq. (1). For an observable $g : \Omega \to \mathbb{R}$ in a Banach space $\mathcal{F}$, the Koopman operator is

$$\mathcal{K}^t g(\mathbf{x}) = g(\mathbf{\Phi}^t(\mathbf{x})), \tag{2}$$

which is linear in g for fixed t. Its infinitesimal generator is

$$\mathcal{L}g := \lim_{t\to 0} \frac{\mathcal{K}^t g - g}{t}. \tag{3}$$

We use $\mathcal{F} = \mathcal{C}(\overline{\Omega})$ with the supremum norm and generator domain $D(\mathcal{L})$. For g continuously differentiable near $\overline{\Omega}$, $\mathcal{L}g = \nabla g \cdot \mathbf{f}$ Brunton et al. (2022); choosing $g(\mathbf{x}) = x_i$ recovers the ith vector-field component.

Equation terms and Koopman closure. Suppose $\mathbf{f}$ is polynomial of degree $p > 1$ and g is a monomial of degree k. Then $\mathcal{L}g = \nabla g \cdot \mathbf{f}$ has degree at most $k + p - 1$ and, absent cancellation, contains terms of that degree. Consequently, even the full degree-p library used to report $\mathbf{f}$ is generally not invariant: one generator application reaches degree $2p - 1$, and repeated applications create a degree cascade. The auxiliary space $V + r_\theta$ approximates this missing Koopman closure. Its coordinates are not candidate governing terms unless the scientific hypothesis separately admits them into U.

Choosing $g(\mathbf{x}) = x_i$ makes equation discovery a generator-estimation problem. The Koopman Logarithm Method (KLM) (Mauroy & Gonçalves, 2020) combines $\mathcal{K}^t = e^{t\mathcal{L}}$ with extended dynamic mode decomposition (EDMD) on analytic observables:

$$\widehat{\mathbf{K}} = \arg\min_{\mathbf{K}} \|\mathbf{g}(\mathbf{X}_{t+1}) - \mathbf{g}(\mathbf{X}_t)\mathbf{K}\|_F^2, \qquad \widehat{\mathbf{L}} = T_s^{-1} \log(\widehat{\mathbf{K}}), \tag{4}$$

where the principal logarithm converts the fitted Koopman operator into a generator. Although this analytic fit preserves the meaning of its coordinates, its one-step least-squares defect does not directly constrain how errors accumulate from a common initial state. ReTRN retains the generator readout while changing how the complementary closure coordinates are represented and trained.

3 Anchored Equation Discovery with Adaptive Closure

ReTRN enlarges the Koopman space without changing the coordinates used to state the equation. Throughout representation, training, and readout, U remains fixed, only $V + r_\theta$ adapts, and the fitted generator returns the corresponding block B_U. This deliberate asymmetry reflects the different roles of the two coordinate sets: equation coordinates carry names and units, whereas closure coordinates are approximation variables. Figure 1 shows how the distinction is maintained throughout the estimator.

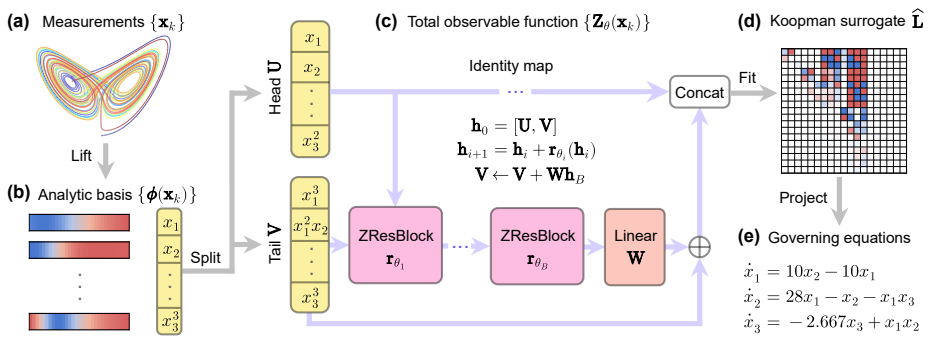


Figure 1: **Overview of ReTRN.** The equation coordinates U remain fixed while a residual network adapts the auxiliary closure coordinates V. A generator is fitted in the combined space, but the reported equation is read directly from its anchored U block.

3.1 Anchored Residualization Defines the Estimator

Let $\phi = [U, V]$, where U spans the prescribed equation class and V provides analytic coordinates for Koopman closure. ReTRN leaves U unchanged and residualizes only the closure block:

$$\begin{aligned} T_\theta &= V + r_\theta, \qquad Z_\theta = [U, T_\theta], \\ \widehat{f}_{\mathrm{eq}} &= UB_U, \qquad \widehat{f}_{\mathrm{lift}} = UB_U + T_\theta B_T, \end{aligned} \tag{5}$$

The returned equation $\widehat{f}_{\mathrm{eq}} = UB_U$ is read from the same generator that trains the lifted model, whereas the closure contribution $T_\theta B_T$ participates in fitting but remains a diagnostic rather than being relabeled as a governing term. Because $r_0 = 0$, ReTRN begins at the analytic-only estimator and can enrich its closure without replacing either the scientific coordinates or the object ultimately reported.

The possibly redundant U may contain inactive candidates but must contain every scientifically admissible term; sparse readout then discovers support within a fixed coordinate system. This does not assume known support: it permits inactive candidates while preventing closure from substituting for a missing anchored mechanism. Under this equation-space adequacy condition, $[B_\star^\top, 0^\top]^\top$ is the target state-output block.

Where the equation space ends. The boundary between head and tail is set by the scientific hypothesis, not by algebraic degree or by whether a feature is analytic or neural. A high-order or nonpolynomial term belongs in U whenever it is an admissible governing mechanism. Conversely, an analytic monomial introduced only because repeated generator action creates it belongs in closure. Protecting every analytic observable would turn auxiliary approximation functions into candidate mechanisms and change the scientific question. ReTRN instead protects the complete named hypothesis class—including inactive candidates—and adapts only the coordinates added for closure. Section 4.2 tests precisely what happens when a true term is placed on the wrong side of this boundary.

3.2 Temporal-Moment Closure Learning

Where anchoring fixes the meaning of the coefficients, temporal moments determine how auxiliary closure is learned. Instead of treating adjacent transitions as unrelated residuals, ReTRN aggregates the residual sequence of each trajectory at several exponential time scales, so early and late discrepancies interact and the closure can account for how local mismatch accumulates across the observed segment. The weights follow the bounded resolvent/Yosida construction used by the Resolvent-Type Method (RTM) (Meng et al., 2024); for μ beyond the semigroup growth bound,

$$\mathcal{L}_\mu = \mu^2(\mu\mathcal{I} - \mathcal{L})^{-1} - \mu\mathcal{I} = \mu^2 \int_0^\infty e^{-\mu t}\mathcal{K}^t dt - \mu\mathcal{I}. \tag{6}$$

ReTRN truncates this construction at the observed horizon and applies the resulting weights to transition defects rather than to raw observables; Eq. (10) in Appendix A.1 gives the truncation bound.

For trajectory r, let $\mathbf{E}_k^{(r)} = \mathbf{Z}_{k+1}^{(r)} - \mathbf{Z}_k^{(r)}\widehat{\mathbf{K}}$ and let $\widetilde{\mathbf{E}}^{(r)}(t)$ be its first-order-hold interpolation. The training objective is

$$\mathcal{L}_{\text{res}} = \frac{1}{N_{\text{traj}}|\mathcal{M}|} \sum_{r=1}^{N_{\text{traj}}} \sum_{\mu\in\mathcal{M}} \left\| \omega_\mu \int_0^{\tau_s} \mu^2 e^{-\mu t} \widetilde{\mathbf{E}}_{\theta_m}^{(r)}(t)\, dt \right\|_2^2 . \tag{7}$$

where $\mathcal{M}$ is a log-spaced set of temporal scales. Each moment couples the interior states and endpoints of a trajectory; Eq. (11) in Appendix A.1 gives the discrete identity. With residual matrix $\mathbf{E}$ and quadrature matrix W, the unnormalized objective is $\|W\mathbf{E}\|_F^2$. If W has full column rank,

$$\sigma_{\min}(W)\|\mathbf{E}\|_F \le \|W\mathbf{E}\|_F \le \sigma_{\max}(W)\|\mathbf{E}\|_F, \tag{8}$$

so $G = W^\top W$ defines a positive-definite temporal metric. Its off-diagonal entries couple different transitions in the segment; Appendix D analyzes this geometry.

Telescoping the fitted recursion bounds the j-step lifted discrepancy by $\gamma_j(\widehat{K})\|W\mathbf{E}\|_F/\sigma_{\min}(W)$, where $\gamma_j(K)^2 = \sum_{s<j}\|K\|_2^{2s}$. Thus Eq. (7) controls in-sample propagation without differentiating through $\widehat{K}^j$; Appendix I separates out-of-sample recovery (see also Appendix F).

Alternating operator and closure fitting. Zero initialization makes ReTRN begin at the analytic-only lift. Each outer iteration refits $\widehat{\mathbf{K}} = \arg\min_{\mathbf{K}} \sum_k \|\mathbf{Z}_{k+1} - \mathbf{Z}_k\mathbf{K}\|_F^2$, holds it fixed while updating the auxiliary residual blocks with Eq. (7), and finally refits the operator and reads B_U from the principal-log generator (Appendices E.1 and E.2). Gradients therefore never pass through least squares, the matrix logarithm, or recursive rollouts. The residual blocks act only on the auxiliary coordinates and project back to the tail (Fig. 1c); Algorithm 1 in Appendix A.1 gives the complete procedure.

3.3 Why Anchoring Is Necessary

Having defined the estimator, we now identify the ambiguity that motivates its anchored construction. Let $Z = [U, T]$ contain equation coordinates U and auxiliary closure coordinates T. Exact prediction determines the combined field $UB_U + TB_T$, but need not determine how that field is divided between the two blocks. The following proposition states the resulting readout criterion.

Proposition 3.1 (Maximal invariance of named equation readout)**.** *Let $Z = [U, T]$, with $h = \dim(U)$, $q = \dim(T)$, and the state coordinates contained in U. Assume that the sampled least-squares design has full column rank and that the principal logarithm exists. Write the state-output columns of the fitted generator as $(B_U^\top, B_T^\top)^\top$.*

(i) Tail reparameterization. *For every $Q \in \mathrm{GL}(q)$, replacing T by TQ leaves B_U unchanged.*

(ii) Equation–closure mixing. *For $A \in \mathbb{R}^{h\times q}$, replacing T by $T - UA$ represents the same state-output field but changes the named block to $B_U + AB_T$.*

(iii) Boundary. *Among invertible linear transformations that leave U pointwise fixed, tail-only changes of basis form the largest subgroup that preserves B_U for every state-output field. Whenever $A_0B_T \neq 0$, the shear path λA_0 leaves the field fixed while generating a continuum of named coefficients.*

Corollary 3.2 (Prediction does not identify named coefficients)**.** *If $AB_T \neq 0$, no rule that depends only on the represented state-output field can equal B_U in every representation and remain invariant to equation–closure shears.*

The proposition distinguishes harmless changes within the closure from changes that alter the scientific readout. Rotating or rescaling the internal closure basis leaves the named block unchanged, but allowing a closure coordinate to contain a named equation term creates an affine family along which the field stays fixed while that term's coefficient moves. Because every member of the family makes the same prediction, no improvement in rollout accuracy can select the intended coefficient.

The corollary makes a coordinate constraint necessary for direct named readout unless an independent rule selects one member of the shear orbit. ReTRN takes this constrained route by declaring the equation chart before fitting. Under $T \mapsto T - UA$, the moment defects satisfy $E_A = ED_A$; if

$WE = [M_U, M_T]$, then $\mathcal{J}(A) = \|M_U\|_F^2 + \|M_T - M_U A\|_F^2$, which is generally not constant. The declared chart and temporal metric select an operational representative whose direct readout is B_U: coefficient identification is conditional on the scientific chart, not on the predicted field alone. An equation–closure shear therefore defines a different estimator rather than an admissible symmetry. Appendix G.2 gives the proofs and corresponding risk decomposition.

3.4 Moment Loss Controls the Returned Readout

We next relate the achieved moment loss to the reported equation. For the learned lift, define $Y_j = [Z(\Phi^{jh}(x_r))]_{r=1}^R$, $E_k = Y_{k+1} - Y_k \widehat{K}$, and $E = [E_0^\top, \dots, E_{J-1}^\top]^\top$. With $h = T_s$, the first-order-hold matrix satisfies $\epsilon_W := \|(W \otimes I_R)E\|_F = (R|\mathcal{M}|\mathcal{L}_{\mathrm{res}})^{1/2}$.

Theorem 3.3 (Moment-to-readout control)**.** *Assume that W and Y_0 have full column rank, the lifted trajectory is twice differentiable, and $Y_0^\dagger Y_1$ and $\widehat{K}$ admit a common principal-log contour Γ. Let $s_Y = \sigma_{\min}(Y_0)$, $b_W = \|e_0^\top W^\dagger\|_2$, and $\sup_{0\le t\le h}\|Y''(t)\|_F \le M_2$. If $f = UB_\star$ on the sampled trajectories, then $\widehat{L} = h^{-1}\log\widehat{K}$ satisfies*

$$\left\| \widehat{L}_{:,1:d} - \begin{bmatrix} B_\star \\ 0 \end{bmatrix} \right\|_F \le \frac{C_\Gamma b_W}{h s_Y}\sqrt{R|\mathcal{M}|\mathcal{L}_{\mathrm{res}}} + \underbrace{\frac{hM_2}{2s_Y} + \delta_{\log,h}}_{\delta_h}, \tag{9}$$

where $\delta_{\log,h} = h^{-1}\|[\log(Y_0^\dagger Y_1) - (Y_0^\dagger Y_1 - I)]_{:,1:d}\|_F$ and C_Γ is the logarithm-stability constant. Fixed feature and state rescaling multiplies the right-hand side by $\|S^{-1}\|_2\|S_x\|_2$.

The bound separates the part controlled by training from the approximation terms that remain: its first term decreases with the achieved moment loss, while the rest reflects feature conditioning, temporal discretization, and the logarithm approximation. Restricting the left-hand side to its first h rows controls the anchored block B_U. This conditional guarantee fixes the readout coordinates; it does not require the learned tail to be functionally orthogonal to U. Because the statement concerns the returned checkpoint, it does not require neural optimization to reach a global minimum. If an active term is omitted from U and assigned to closure, an additional projection bias remains; Appendix G.2 derives that bias, and Appendix D verifies the rank and logarithm conditions numerically.

4 Experiments

4.1 Equation Recovery from Seven Observations

We first test equation recovery on Van der Pol, Duffing, Lorenz, and the ODEBench Competition, Reaction, and Magnet systems (Strogatz, 2015; d'Ascoli et al., 2023), and then examine the mechanism through interventions and ablations before extending to new equation families, coupled systems, known inputs, and measured data. Every reported coefficient belongs to UB_U; unless explicitly marked as a closure diagnostic, the neural tail is removed before rollout, so success requires the named equation itself to reproduce the dynamics.

The main benchmark contains 20 trajectory subsamples, each retaining 800 of 1,000 training trajectories and seven observations at $t = 0, .05, \dots, .30$. It evaluates complete-library normalized root-mean-square coefficient error (NRMSE), area under the precision–recall curve (AUPRC), and the three-second rollout mean-squared error (MSE) of the returned analytic equation (Table 1; Appendices A.2 and A.1). The library-based methods receive the same unabridged analytic candidate library, observation design, validation endpoint, and evaluator; no method is given a truth-trimmed equation space, and coefficient truth and long tests remain unavailable during selection.

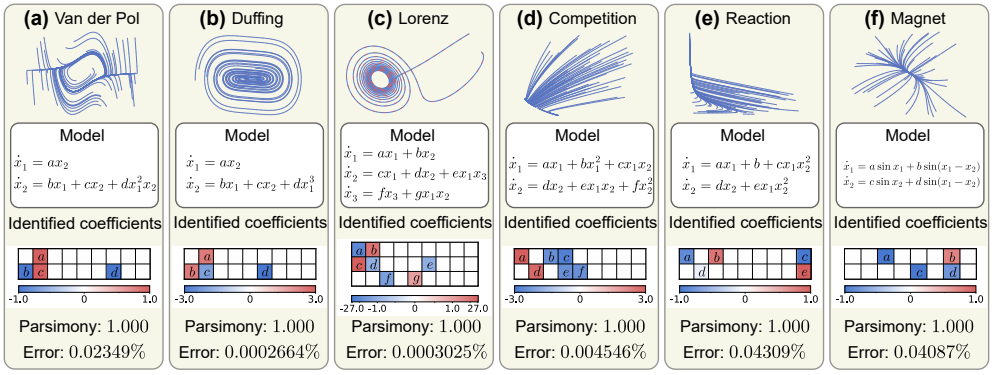


Figure 2: **Dense-grid equation recovery across six nonlinear systems.** The thresholded ReTRN equations accurately reproduce the reference trajectories; the seven-observation protocol is reported separately in Table 1.

Method	Van der Pol	Duffing	Lorenz	Competition	Reaction	Magnet
Complete coefficient NRMSE						
ReTRN	**.008095±.000566**	**(2.206±1.076)e−4**	**.005624±.000823**	**.005631±.001067**	**.002690±.000627**	**(5.176±2.517)e−4**
KLM	.01052±.00106	(.7185±.0996)$e-3$	.007790±.001128	.006322±.000099	.004983±.001600	(.6389±.2457)$e-3$
RTM	.1718±.0019	.13245±.00043	.04112±.00404	.13514±.00398	.01759±.00157	.002069±.000047
WSINDy	.009181±.001387	.01525±.00547	.04860±.00617	.7631±.4772	.03179±.00769	.000875±.000426
Three-second rollout MSE						
ReTRN	(3.251±.993)$e-4$	**(1.009±.796)e−5**	**2.379±.771**	**(2.392±1.074)e−5**	**(.7169±.2094)e−6**	**(5.519±1.926)e−11**
KLM	(.4326±.1222)$e-3$	(.2230±.0308)$e-3$	11.13±2.76	(2.996±.274)$e-5$	(2.043±.275)$e-6$	(1.588±1.068)$e-9$
RTM	1.339±.007	20.19±.06	63.15±7.19	.1037±.0017	(1.757±.089)$e-4$	(8.595±2.634)$e-8$
WSINDy	**(.1176±.0354)e−3**	.03592±.05152	39.69±3.25	(4.108±9.985)$e5$	(2.790±2.102)$e-5$	(4.037±2.641)$e-9$
PySR	–	–	60.13±1.67	–	(7.918±.720)$e-6$	(1.525±2.520)$e-5$

Table 1: **Equation discovery from seven observations per trajectory.** Library-based methods use the same unabridged analytic candidates; all reported equations are selected without coefficient truth or long-test access.

ReTRN has unit AUPRC: every active coefficient is ranked above the inactive candidates, although this does not imply exact support at the locked threshold. Under the common unabridged library, ReTRN has the lowest coefficient NRMSE on all six systems and the lowest mean rollout MSE on five, with WSINDy lower only on Van der Pol. This yields 11 of 12 primary system–metric wins without giving any comparator a support-informed candidate space.

The comparison spans distinct routes to a named equation: KLM and RTM share the analytic Koopman readout but cannot adapt its closure, WSINDy uses weak-form regression, and PySR searches free-form symbolic expressions. On the dense-grid benchmark, ReTRN's three-second MSE is $1.91e{-7}$, $4.22e{-8}$, $1.92e{-5}$, $2.53e{-8}$, $9.15e{-8}$, and $1.73e{-8}$ on Van der Pol, Duffing, Lorenz, Competition, Reaction, and Magnet. It is lower than SINDy, E-SINDy, PySR, and NeSymReS in 23 of 24 method–system comparisons; the sole exception is PySR on Magnet ($1.64e{-10}$). The complete table is in Appendix Table 24.

The advantage persists as the temporal record is shortened: with four observations, ReTRN obtains NRMSE .0442 on Reaction and .00171 on Magnet, compared with .2157 and .00888 for WSINDy (Appendix A.11). Across the broader robustness study, it also gives the lowest NRMSE at 9/10 temporal-downsampling levels and all 10 noise levels (Appendix C), even though each test changes the available information while leaving the scientific readout unchanged.

System	Metric	ReTRN	KLM	RTM	WSINDy
Lorenz	NRMSE	**.00494 ± .00080**	.00717 ± .00115	.04042 ± .00298	.04776 ± .00701
	MSE	**1.881 ± .781**	11.13 ± 2.73	60.04 ± 6.19	39.28 ± 4.09
Reaction	NRMSE	**.00344 ± .00145**	.00421 ± .00065	.01767 ± .00199	.03178 ± .00614
	MSE (10^{-6})	**.600 ± .143**	1.762 ± .219	159.91 ± 27.59	30.30 ± 20.04
Magnet	NRMSE (10^{-3})	**.524 ± .138**	.722 ± .295	2.073 ± .050	.860 ± .504
	MSE (10^{-9})	**.0802 ± .0236**	1.035 ± .851	82.04 ± 25.63	49.73 ± 33.01

Table 2: **Independent seven-observation simulations.**

Independent simulations rule out a favorable reuse of trajectories as the source of this pattern: on each of 60 independent data replicates (20 per system), ReTRN has lower rollout error than KLM, RTM, and WSINDy, yielding 180/180 pairwise wins. Mean MSE falls 83.1/65.9/92.3% relative to the strongest competitor on Lorenz, Reaction, and Magnet ($p = 3.43e{-}5$ after Holm correction). The corresponding NRMSE win counts are 20/20, 16/20, and 13/20, so the main empirical advantage is not simply a smaller average coefficient error but an equation that remains consistently better when integrated beyond the observed segment (Appendix A.4).

4.2 Anchoring Stabilizes the Named-Equation Readout

Because accurate forecasting alone cannot reveal whether a named coefficient is well defined, we intervene directly on the equation–closure boundary. Anchoring fixes the reported equation coordinates. The comparison keeps the represented functions and neural capacity available while changing only whether a true term is protected as an equation coordinate or left for the closure to absorb.

On Duffing, moving x_1^3 from the equation space into closure preserves the state-output field to $1.10e{-}16$ but changes the direct coefficient block by $70.79 \pm 5.65\%$. The models match in 2,016 parameters, data, and seeds; restoring x_1^3 to U reduces coefficient error $214\times$, rollout MSE $6{,}278\times$, and field-projection error $21.4\times$, while the neural contribution falls from 27.32% to $.0735\%$. Across four dynamics, Table 3 shows why redundancy is safer under support uncertainty: inactive candidates can be removed sparsely, whereas an omitted mechanism cannot be recovered as an anchored coefficient (Appendix A.9).

System	Head NRMSE increase	Rollout MSE increase
Duffing	$214\times$	$6{,}278\times$
Lorenz	$3.34\times$	$5.66e3\times$
Reaction	$150\times$	$8.15e5\times$
Magnet	$1{,}036\times$	$1.85e7\times$

Table 3: **Penalty from assigning a true equation term to closure.**

Learned-observable controls either distribute neural features across the representation or append them to the analytic lift (Li et al., 2017; Takeishi et al., 2017; Lusch et al., 2018). Against the distributed construction, ReTRN wins complete NRMSE, sparse NRMSE, and MSE in all 60 pairs, reducing complete NRMSE by 83.1%, 96.9%, and 99.9% on Lorenz, Reaction, and Magnet. Against appended features, it lowers coefficient error in all 60 pairs by 18.4%, 75.8%, and 72.6%, respectively; the larger appended lift has a significant rollout advantage only on Reaction, separating equation fidelity from universal forecast dominance (Appendix A.10).

These interventions isolate complementary parts of the claim: the exact shear changes the readout while holding the field fixed, moving a true term into closure measures the resulting error after fitting, and the learned-observable controls retain neural capacity while changing where it enters the representation. Their agreement attributes the benefit to anchoring during representation learning rather than to a smaller network or a more favorable final projection (Appendix A.8).

4.3 Temporal Moments Exploit Within-Trajectory Order

Because anchoring and temporal moments play different roles, we ablate them separately: replacing temporal moments with a pointwise transition objective preserves the representation and readout while removing coupling across the observed segment, whereas removing learned closure preserves the analytic coordinates but returns to a fixed lift.

Without temporal moments, coefficient NRMSE rises by $3.16/17.7/148/6.35/5.10/1.80\times$ and rollout MSE by $11.9/14.4/7{,}879/22.9/8{,}749/29.6\times$ on Van der Pol, Duffing, Lorenz, Competition, Reaction, and Magnet; removing learned closure likewise increases every coefficient and rollout error. Full ReTRN consequently wins all 24 system–metric comparisons in Table 4 (Appendix A.6). A diagonal-Gram control preserves marginal time weights while removing cross-transition products: moments have lower mean rollout error on all three tested systems, with significant paired separation on Lorenz (Appendix A.7).

	No temporal moments / full		No learned closure / full	
System	NRMSE	MSE	NRMSE	MSE
Van der Pol	3.16	11.9	2.02	1.75
Duffing	17.7	14.4	1.04	1.02
Lorenz	148	7,879	2.78	3.75
Competition	6.35	22.9	1.95	14.1
Reaction	5.10	8,749	1.68	31.4
Magnet	1.80	29.6	13.9	613

Table 4: **Error increase after removing temporal moments or learned closure.**

4.4 Frozen-Configuration Generalization Across Equation Families

Across seven additional oscillatory, chaotic, and trigonometric families, ReTRN attains unit AUPRC and NRMSE from $1.39e{-}7$ to .00402 on Rössler, Sprott A, Shimizu–Morioka, and damped pendulum, and NRMSE $4.29e{-}9$, $8.68e{-}5$, and $2.33e{-}4$ on Genesio–Tesi, Dadras, and Rabinovich–Fabrikant; three-second MSE ranges from $2.42e{-}15$ to $1.66e{-}5$. All selected equation spaces contain the true terms. When U must instead be chosen from nested spaces, an optional frozen adequacy gate accepts 39/40 individual deployments, including 9/10 Lorenz–84 fits, and every accepted space contains the true terms. Its stricter family-level rule abstains on Lorenz–84 after five conservative warnings (candidate-level specificity .9375); Appendix A.5 gives the protocol and per-family statistics.

4.5 Recovery in Coupled Systems and with Known Inputs

Increasing the library from degree 3 to 6 and the state from 3D to 6D/9D grows the analytic space from 19 to 219 functions and the neural tail from 4,902 to 582,102 parameters. For 3D orders 3–6, the tuples (terms, parameters, runtime in seconds, NRMSE) are $(19, 4{,}902, 79.1, 2.06e{-}6)$, $(34, 14{,}892, 91.8, 3.16e{-}6)$, $(55, 37{,}950, 103.4, 3.83e{-}6)$, and $(83, 85{,}158, 153.6, 4.05e{-}6)$; full-cubic 6D and 9D give $(83, 85{,}158, 226.8, 3.10e{-}6)$ and $(219, 582{,}102, 673.0, 2.83e{-}6)$. On a separately selected quadratic library, ReTRN also wins all 20 matched 3D pairs on both metrics. On a 9D coupled-Lorenz ring, it recovers all 39 active terms, including 18 cross-system couplings, with NRMSE $(9.6548 \pm .000003)e{-}6$ and MSE $(1.4405 \pm .0046)e{-}5$; five 6D fits recover all six couplings at both strengths, with coefficient MAE $1.35e{-}5$ and $2.50e{-}5$ (Appendix B).

The same principle extends to known inputs by anchoring admissible control-affine mechanisms u, x_1u, and x_2u rather than treating them as anonymous closure features. With 1% observation noise, ReTRN recovers the bilinear-Duffing input coefficients $(1, .2, 0)$ as $(.9982, .1966, -.00068)$ and lowers rollout MSE by 41.6%, 49.4%, and 37.7% relative to the strongest competitor under constant, sinusoidal, and switching inputs (Appendix B.1), preserving the readout semantics in the controlled setting.

4.6 Long-Horizon Recovery on Measured Heave Decay

We finally test whether the returned equation remains useful beyond simulated dynamics. On held-out record 4 of the measured heave-decay data (Kramer et al., 2021), a test horizon $302\times$ longer than the ten-observation training segments, ReTRN attains MSE $1.340e{-}5$ compared with $6.796e{-}5$ for WSINDy, $5.699e{-}4$ for KLM, and $7.309e{-}4$ for RTM. Its head-only equation estimates the normalized stiffness of the LPF0 reference model within $.3449\%$ and adds compact nonlinear restoring and damping terms (Appendix N.7); because no neural closure enters this rollout, the long-horizon agreement evaluates the scientific readout rather than a hidden correction used only for prediction.

5 Discussion and Conclusion

ReTRN separates named equations from learned Koopman closure. Theory, gauge interventions, 11/12 primary wins with a common unabridged library, baseline wins across 60 independent replicates, and all 24 component ablations support the design. A redundant U must still be specified: inactive candidates can be pruned, but missing mechanisms cannot be anchored. Future anchored symbolic search could propose new named observables while reserving the learned tail for closure.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing, draft sections of the paper, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, and manage code and materials. We have not used generative AI tools for other tasks. All AI-assisted suggestions were treated as provisional. The authors inspected and tested the code, verified the reported experiments and machine-readable outputs, checked the mathematical derivations and citations against primary sources, and reviewed and approved the final text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The unified supplementary ZIP contains code, data, checkpoints, environments, and machine-readable outputs. Appendix A.1 maps the principal experiments to their entry points. Paired comparisons use identical trajectory resamples; separately frozen pools are explicitly excluded from paired statistics. The protocol-locked equation-space study freezes its rule and sources before fitting and serializes coefficient truth and long tests only after model and selection freeze. Each ReTRN result is obtained from an independently trained model rather than an ensemble.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019. doi: 10.1145/3292500.3330701.

Omri Azencot, N. Benjamin Erichson, Vanessa Lin, and Michael 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*, pp. 475–485. PMLR, 2020.

Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurelien Lucchi, and Giambattista Parascandolo. Neural symbolic regression that scales. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 936–945. PMLR, 2021.

Josh Bongard and Hod Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104(24):9943–9948, 2007. doi: 10.1073/pnas.0609476104.

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern Koopman theory for dynamical systems. *SIAM Review*, 64(2):229–340, 2022. doi: 10.1137/21M1401243.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

Stéphane d'Ascoli, Sören Becker, Alexander Mathis, Philippe Schwaller, and Niki Kilbertus. ODEFormer: Symbolic regression of dynamical systems with transformers. *arXiv preprint arXiv:2310.05573*, 2023.

Urban Fasel, J Nathan Kutz, Bingni W Brunton, and Steven L Brunton. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 478(2260): 20210904, 2022. doi: 10.1098/rspa.2021.0904.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, 2008. doi: 10.1137/1.9780898717778.

Stefan Klus, Feliks Nüske, Sebastian Peitz, Jan-Hendrik Niemann, Cecilia Clementi, and Christof Schütte. Data-driven approximation of the Koopman generator: Model reduction, system identification, and control. *Physica D: Nonlinear Phenomena*, 406:132416, 2020. doi: 10.1016/j.physd.2020.132416.

Morten Bech Kramer, Jacob Andersen, Sarah Thomas, Flemming Buus Bendixen, Harry Bingham, Robert Read, Nikolaj Holk, Edward Ransley, Scott Brown, Yi-Hsiang Yu, et al. Highly accurate experimental heave decay tests with a floating sphere: A public benchmark dataset for model validation of fluid–structure interaction. *Energies*, 14(2):269, 2021. doi: 10.3390/en14020269.

Qianxiao Li, Felix Dietrich, Erik M Bollt, and Ioannis G Kevrekidis. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the Koopman operator. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(10):103111, 2017. doi: 10.1063/1.4993854.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019.

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

Alexandre Mauroy and Jorge Gonçalves. Koopman-based lifting techniques for nonlinear systems identification. *IEEE Transactions on Automatic Control*, 65(6):2550–2565, 2020. doi: 10.1109/TAC.2019.2941433.

Yiming Meng, Ruikun Zhou, Melkior Ornik, and Jun Liu. Resolvent-type data-driven learning of generators for unknown continuous-time dynamical systems. *arXiv preprint arXiv:2411.00923*, 2024.

Daniel A Messenger and David M Bortz. Weak SINDy: Galerkin-based data-driven model selection. *Multiscale Modeling & Simulation*, 19(3):1474–1497, 2021. doi: 10.1137/20M1343166.

Shaowu Pan and Karthik Duraisamy. Physics-informed probabilistic learning of linear embeddings of nonlinear dynamics with guaranteed stability. *SIAM Journal on Applied Dynamical Systems*, 19 (1):480–509, 2020. doi: 10.1137/19M1267246.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009. doi: 10.1126/science.1165893.

Steven H Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview Press, 2 edition, 2015.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning Koopman invariant subspaces for dynamic mode decomposition. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zhexuan Zeng, Zuogong Yue, Alexandre Mauroy, Jorge Gonçalves, and Ye Yuan. A sampling theorem for exact identification of continuous-time nonlinear dynamical systems. *IEEE Transactions on Automatic Control*, 69(12):8402–8417, 2024. doi: 10.1109/TAC.2024.3409639.

A Experimental Protocols and Complete Results

A.1 Full ReTRN training procedure

Starting from $\mathbf{h}_0 = [\mathbf{U}, \mathbf{V}]$, the zero-initialized residual blocks enrich only the auxiliary coordinates and project back to the tail,

$$\mathbf{h}_{i+1} = \mathbf{h}_i + \tanh(\mathrm{LayerNorm}(\mathbf{h}_i))\mathbf{W}_i, \qquad \mathbf{T}_\theta = \mathbf{V} + \mathbf{h}_{N_b}\mathbf{W}.$$

If $\|\mathcal{K}^t\| \le Ce^{\kappa t}$, truncating the resolvent/Yosida integral at the observed horizon τ_s gives

$$\|\mathcal{L}_\mu - \mathcal{L}_{\mu,\tau_s}\| \le \frac{C\mu^2}{\mu-\kappa} e^{-(\mu-\kappa)\tau_s}. \tag{10}$$

For $\bar{\mu} = \mu T_s$, $\rho_\mu = e^{-\bar{\mu}}$, and $J = \tau_s/T_s$, the corresponding discrete moment is

$$\mathbf{S}_\mu^{(r)} = \sum_{k=0}^{J-1} \rho_\mu^k \mathbf{E}_k^{(r)} = \rho_\mu^{J-1}\mathbf{Z}_J^{(r)} - \mathbf{Z}_0^{(r)}\widehat{\mathbf{K}} + \sum_{j=1}^{J-1} \rho_\mu^{j-1}\mathbf{Z}_j^{(r)}(\mathbf{I} - \rho_\mu\widehat{\mathbf{K}}). \tag{11}$$

The complete alternating procedure is given below.

Algorithm 1 ReTRN training

Require: Short-trajectory data $\{\mathbf{x}_k\}$, interval T_s, duration τ_s; analytic basis $\boldsymbol{\phi}$; tail residual network $\mathbf{r}_\theta$; scale set $\mathcal{M}$ with weights $\{\omega_\mu\}$; learning rate $\eta > 0$.

1: Initialize θ at zero, yielding the nested analytic-only model.
2: **for** $m = 0, 1, 2, \ldots$ until convergence **do**
3: Compute $\{\mathbf{Z}_{\theta_m}(\mathbf{x}_k)\}$ as in Fig. 1c.
4: Fit $\widehat{\mathbf{K}}_{N,m}$ by least squares and set $\widehat{\mathbf{L}}_{N,m} = T_s^{-1}\log\widehat{\mathbf{K}}_{N,m}$, without gradients.
5: Compute $\mathcal{L}_{\mathrm{res}}$ using Eq. (7).
6: Hold $\widehat{\mathbf{K}}_{N,m}$ and $\widehat{\mathbf{L}}_{N,m}$ fixed; update $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{\mathrm{res}}(\theta)$.
7: **end for**
8: Refit $\widehat{\mathbf{K}}$ and $\widehat{\mathbf{L}} = T_s^{-1}\log\widehat{\mathbf{K}}$ using the final $\mathbf{Z}_\theta$.
9: Undo feature scaling to obtain $(B_U, B_T) = S^{-1}\widehat{L}_{:,1:d}S_x$; **return** B_U and report UB_U, retaining $T_\theta B_T$ and $r_\theta B_T$ as closure diagnostics.

This appendix specifies the data splits, selection rules, metrics, and complete results behind every main-text comparison. It covers the seven- and four-observation experiments, trajectory-coverage robustness, scaling and coupled dynamics, known observed inputs, temporal-moment analyses, heave decay, and dense-grid symbolic comparisons. The supplementary archive contains the implementations, processed data, configurations, independently trained checkpoints, per-run outputs, and a guide mapping each experiment to its reproduction command.

The principal comparison contains all 480 system–method–subsample records, with per-run configurations and fitted readouts. Automated selection uses disjoint short-validation rollouts; neither true coefficients nor final tests are available during selection. Table 5 distinguishes overlapping stability subsamples, independent paired replications, mechanism studies, and measured-data controls so that each aggregate is interpreted at the level supported by its design. Because the six primary families inform method development, Appendix A.5 separately evaluates frozen optimization settings on polynomial and trigonometric families outside that set.

The main grid is $t = 0, .05, \ldots, .30$ and therefore contains seven observations and six transitions. For five systems these nodes are exact subsamples of the available .01-spaced arrays. The Van der Pol array ends at .29 s, so its .30 s endpoint is generated by one additional .01-s RK4 step before subsampling. Twenty evaluation subsamples (seeds 1001–1020) each contain 800 of 1,000 trajectories. All methods use the same simulator, observation design, analytic library, validation endpoint, and long-rollout evaluator; their selection variables and search procedures remain method-specific. Their arrays are identical within each resample for Duffing, Lorenz, Competition, Reaction, and Magnet. The Van der Pol ReTRN row uses a separately frozen, identically sized resampling pool, so it enters only the descriptive development count and not paired statistics. Five tuning subsamples and an independent 200-trajectory short-validation set select each ReTRN configuration without coefficient

Protocol	Systems and repetitions	Method alignment	Purpose
Principal seven-point	Six; 20 trajectory subsamples	Common data for five systems; separate frozen Van der Pol pool	Descriptive stability and reported-mean ranking
Independent seven-point	Lorenz, Reaction, Magnet; 20 new replicates	All four methods paired within replicate	Primary paired inference
Frozen cross-family transfer	Rössler, Sprott A, Shimizu–Morioka; 10 fits each	One jointly selected ReTRN configuration for every family	Absolute frozen-configuration transfer
Cross-function transfer	Damped pendulum; 10 fits	Same frozen optimization configuration; prescribed trigonometric head	Function-class transfer evidence
Truth-free equation-space selection	Same four target families; 120 candidate fits	Fixed optimization configuration; family-level equation set/threshold selected only on short validation	Validation-based equation-space selection and sensitivity
Crossing-adequacy selection	Same four target families; 160 fresh fits	Four predeclared equation spaces; global and leave-one-seed-out short-validation selection	Fresh-resample sensitivity across 70 inadequate and 90 adequate candidates
Frozen adequacy gate	Four polynomial transfer families; 160 fits	Frozen warning/readout thresholds and disjoint short-data partitions; no baseline refit	Truth-free equation-space deployment with explicit abstention
Single-data-set gate	Same four families; 40 independent deployments	Each decision uses only that replicate's frozen warning partition	Per-data-set adequacy and abstention analysis
Dictionary construction	Lorenz, Reaction, Magnet; 20 new pairs	Same capacity, data, seeds, projection, and validation	Equation-anchoring mechanism evidence
Dense-grid component ablation	Six systems; one validation-selected fit per variant	Shared data/library/evaluator; each trainable objective selects its own architecture and learning rate on held-out validation	Component evidence for temporal moments and learned closure
Temporal/diagonal objective control	Lorenz, Reaction, Magnet; 20 new pairs	Equal-budget, objective-specific validation selection; paired final data and seeds	Isolate cross-transition products from marginal time weighting
Four-point stress test	Reaction, Magnet; 20 subsamples	Shared arrays and seven-point-selected settings for both methods	Symmetric transfer evidence
Coupled/control	Five optimization seeds per condition	Fixed-data method comparison	Scope diagnostics
Heave decay	Segment validation on records 1–3	Record 4 remains intact and is opened only for final evaluation	Measured-data long-horizon test

Table 5: **Map of the principal experimental protocols.** *Paired* denotes identical train, validation, and test arrays within a replication unit.

truth or long-test access. No denoising is applied, and feature normalization is fitted only on training data. Mean $\pm$ sample SD summarizes the 20 fits; paired win counts and paired tests are restricted to comparisons sharing arrays. The frozen PySR expressions use a separate twenty-resample pool and remain a free-form contextual reference.

A.2 Metrics and complete seven-point statistics

Let $B, \widehat{B} \in \mathbb{R}^{n\times d}$ be aligned coefficient matrices over the full analytic initialization. For ReTRN, $\widehat{B}$ contains the reported equation-space block and zeros on closure-reference rows; this zero padding compares all methods against the same prescribed coefficient coordinates without interpreting closure as equation terms. We use

$$\mathrm{NRMSE}(\widehat{B}) = \frac{\sqrt{(nd)^{-1}\sum_{i,j}(B_{ij}-\widehat{B}_{ij})^2}}{\|B\|_1/\|B\|_0}.$$

AUPRC is average precision obtained by scoring each coefficient with $|\widehat{B}_{ij}|$ against the exact binary support of B. It measures support ranking over all thresholds; AUPRC one means every active term is ranked above every inactive term, not that a particular threshold yields exact support. NRMSE and AUPRC are computed before sparsification, and ReTRN training contains no sparsity penalty. For rollout, every threshold is selected using only the corresponding short-validation trajectories. The deterministic matched comparators use $\{10^{-4}, 10^{-3}, 10^{-2}, .05, .1\}$ with per-fit validation selection. ReTRN jointly selects one system-level threshold and its configuration by mean validation rollout over tuning subsamples 2001–2005, then applies both unchanged to all 20 final fits; the locked thresholds are $5e{-}5, 3e{-}4, 0, 1e{-}4, 1e{-}4, 3e{-}6$ in system order, and the selected values and selection hashes are retained with each system's final record. Thus search spaces are method-specific, while the validation endpoint and final evaluator are shared; training and validation observations are identical for the five systems specified above. Neither the true coefficients nor the long-test rollouts enter

selection. Since PySR may generate terms outside the prescribed library, its distinct-pool reference is reported only through native rollout MSE; projected coefficient errors remain a separate diagnostic.

A.3 SEVEN-OBSERVATION RETRN CONFIGURATION AND OPTIMIZATION

All six seven-observation ReTRN configurations are selected by mean short-validation rollout over tuning subsamples 2001–2005. Candidate configurations vary the number of closure coordinates and residual blocks, learning rate, batch size, and log-spaced μ grid. Zero initialization supplies the exact analytic-only reference, which is reported separately on the same data for the five array-matched systems. An activity check finds a minimum learned-closure ratio of .00233 across the recorded selection fits and excludes none; all prespecified final fits remain in the summaries. The final settings are listed in Table 6. Selection follows each estimator's native degrees of freedom: KLM and RTM select the relative readout threshold, WSINDy also selects its test-function scale, and ReTRN selects neural architecture and optimization variables. For the five array-matched systems, training and validation observations and the final evaluator are identical; all six systems use the same validation endpoint and exclude coefficient truth and long tests. This aligns data and evaluation but does not equate method-specific search dimensions or compute, so Table 7 is interpreted descriptively. Each outer step recomputes $\widehat{K}$ from accumulated normal equations using `solve`, with `lstsq` as the singular fallback, and then holds $\widehat{K}$ fixed during an AdamW update (zero weight decay) of the neural closure. Cosine learning-rate decay approaches one tenth of the initial rate; the final operator and principal-log generator are recomputed after the selected checkpoint is restored.

System	Tail coordinates	Residual blocks	Learning rate	Batch / μ grid
Van der Pol	1	2	$1.5e{-4}$	128 / 30 in $[10, 10^4]$
Duffing	2	18	$1.440e{-4}$	256 / 30 in $[10, 10^4]$
Lorenz	10	8	$3.161e{-5}$	256 / 30 in $[10, 10^4]$
Competition	5	16	$2.5e{-3}$	256 / 30 in $[10, 10^4]$
Reaction	4	10	$1.09e{-4}$	256 / 30 in $[10, 10^4]$
Magnet	2	16	$1.083e{-3}$	256 / 30 in $[10, 10^4]$

Table 6: Validation-selected ReTRN settings for the seven-observation comparison. Here $T_s = .05$, so the reported $\mu \in [10, 10^4]$ is equivalently the dimensionless grid $\bar{\mu} = \mu T_s \in [.5, 500]$. Epoch budgets and selected checkpoints are recorded in the machine-readable configuration files.

System	Method	Complete NRMSE	3 s MSE
Van der Pol	ReTRN	$\mathbf{.008095 \pm .000566}$	$(3.251 \pm .993)e{-4}$
	KLM	$.01052 \pm .00106$	$(.4326 \pm .1222)e{-3}$
	RTM	$.1718 \pm .0019$	$1.339 \pm .007$
	WSINDy	$.009181 \pm .001387$	$(.1176 \pm .0354)e{-3}$
Duffing	ReTRN	$(2.206 \pm 1.076)e{-4}$	$(1.009 \pm .796)e{-5}$
	KLM	$(.7185 \pm .0996)e{-3}$	$(.2230 \pm .0308)e{-3}$
	RTM	$.13245 \pm .00043$	$20.19 \pm .06$
	WSINDy	$.01525 \pm .00547$	$.03592 \pm .05152$
Lorenz	ReTRN	$.005624 \pm .000823$	$2.379 \pm .771$
	KLM	$.007790 \pm .001128$	11.13 ± 2.76
	RTM	$.04112 \pm .00404$	63.15 ± 7.19
	WSINDy	$.04860 \pm .00617$	39.69 ± 3.25
Competition	ReTRN	$.005631 \pm .001067$	$(2.392 \pm 1.074)e{-5}$
	KLM	$.006322 \pm .000099$	$(2.996 \pm .274)e{-5}$
	RTM	$.13514 \pm .00398$	$.1037 \pm .0017$
	WSINDy	$.7631 \pm .4772$	$(4.108 \pm 9.985)e5$
Reaction	ReTRN	$.002690 \pm .000627$	$(.7169 \pm .2094)e{-6}$
	KLM	$.004983 \pm .001600$	$(2.043 \pm .275)e{-6}$
	RTM	$.01759 \pm .00157$	$(1.757 \pm .089)e{-4}$
	WSINDy	$.03179 \pm .00769$	$(2.790 \pm 2.102)e{-5}$
Magnet	ReTRN	$(5.176 \pm 2.517)e{-4}$	$(5.519 \pm 1.926)e{-11}$
	KLM	$(.6389 \pm .2457)e{-3}$	$(1.588 \pm 1.068)e{-9}$
	RTM	$.002069 \pm .000047$	$(8.595 \pm 2.634)e{-8}$
	WSINDy	$.000875 \pm .000426$	$(4.037 \pm 2.641)e{-9}$

Table 7: Complete seven-point development results ($n = 20$ stability subsamples). Five systems use identical arrays across methods; Van der Pol uses a separately frozen, identically sized ReTRN pool. Its reported means enter the descriptive count but are excluded from paired statistics.

ReTRN has the smallest mean in 11 of the 12 coefficient-and-rollout cells. Relative to the strongest comparator in each winning cell, its coefficient-NRMSE reductions are $11.8\%/69.3\%/27.8\%/10.9\%/46.0\%/19.0\%$ in table order, and its rollout-MSE reductions are $95.5\%/78.6\%/20.2\%/64.9\%/96.5\%$ for Duffing through Magnet. Van der Pol rollout is the sole exception: WSINDy is 63.8% lower. Per-resample metrics are included in the artifact; paired statistics exclude Van der Pol and any other distinct-pool comparison.

On the RTX 3500 Ada GPU used for this study, mean ReTRN optimization times for Van der Pol, Duffing, Lorenz, Competition, Reaction, and Magnet are 38.6, 108.4, 104.4, 33.2, 26.5, and 109.7 s per fitted subsample. These times exclude hyperparameter search and final metric evaluation. KLM and RTM use closed-form matrix estimators, while WSINDy uses deterministic sparse regression after validation selection; their fitting times are therefore not directly comparable to neural optimization time.

Seven-observation PySR reference. We additionally evaluate PySR (Cranmer, 2023) on Lorenz, Reaction, and Magnet. Polynomial operators are used for Lorenz and Reaction and trigonometric operators for Magnet, with parsimony and expression size selected on validation data over 80 iterations and 16 populations. Across its separate 20-subsample pool, its rollout MSE is 60.13 ± 1.67, $(7.918 \pm .720)e{-}6$, and $(1.525 \pm 2.520)e{-}5$, respectively. The Magnet results are bimodal, with median $4.64e{-}10$ and a larger mean due to several inaccurate expressions. Since this pool is distinct, PySR is retained as a contextual symbolic reference and excluded from paired statistics and the 11/12 common-protocol count.

A.4 Independent seven-observation experiments

The Lorenz, Reaction, and Magnet arrays underlying Table 1 are stored to five significant digits, whereas the independent simulator output retains full floating-point precision. Table 2 therefore uses independently simulated, higher-precision data rather than attributing every change solely to random sampling. Configurations and one system-level ReTRN readout threshold are selected on five separate tuning replicates and disjoint short-validation trajectories, then locked before the 20 final fits and long-test generation.

For each of Lorenz, Reaction, and Magnet, we generate 20 independent data replicates, each with 800 training and 200 short-validation trajectories at $t = 0, .05, \ldots, .30$. Every method receives exactly the same arrays within a replicate. Initial states are uniform on $[-20, 20] \times [-30, 30] \times [0, 50]$ for Lorenz, $[0, 2]^2$ for Reaction, and $[-1.5, 1.5]^2$ for Magnet. Reference simulation uses RK4 with internal step $.001$ s. Halving this step on 16 training initial states per system changes the observed states by at most $4.85e{-}7$, $1.52e{-}11$, and $3.55e{-}15$, respectively; this is a limited integration check, not a bound for all trajectories.

The independent-replicate protocol selects estimator configurations separately from Table 6. ReTRN uses one model per replicate. Lorenz, Reaction, and Magnet use tail dimension/number of residual blocks/learning rate $3/8/1e{-}4$, $4/10/1.09e{-}4$, and $4/16/4.30e{-}3$; batch sizes are 128, 256, and 256. All three use 30 log-spaced μ values in $[10, 10^4]$, or $\bar{\mu} \in [.5, 500]$ at $T_s = .05$. Checkpoints are selected by residual on one half of the short-validation trajectories, while configuration and threshold selection use rollout on the disjoint half. ReTRN locks one system-level threshold across the 20 final fits; each deterministic comparator receives per-fit selection over the same threshold grid. All trained matrices and readouts are frozen before final-test generation. Each replicate then receives 32 new three-second test trajectories, sampled every $.01$ s. No parameter or readout choice is changed using these tests. This design separates independent-replicate inference from the overlapping development subsamples in Table 1.

The common equation evaluator uses RK4 with maximum internal step $.002$ s for both validation and test. MSE averages squared state error over trajectories, coordinates, and the 300 positive test times; the exactly shared initial state is excluded. Nonfinite states or maximum absolute state at least 10^6 are counted as failed trajectories and assigned a 10^{12} error for each remaining observation. There are no such failures among the 640 test trajectories per system and method. WSINDy's SR3 solver reaches its iteration limit in some fits; all returned models are retained, without asserting solver convergence. Table 2 reports every mean and sample SD. AUPRC is one throughout except Reaction RTM ($.9078 \pm .0129$) and WSINDy ($.8696 \pm .0359$).

System	Baseline	Metric	ReTRN wins / 20	Median paired difference	p_{Holm}
Lorenz	KLM	NRMSE	20	$-.002239$	$3.433e{-5}$
		MSE	20	-8.927	$3.433e{-5}$
	RTM	NRMSE	20	$-.03532$	$3.433e{-5}$
		MSE	20	-58.66	$3.433e{-5}$
	WSINDy	NRMSE	20	$-.04163$	$3.433e{-5}$
		MSE	20	-36.77	$3.433e{-5}$
Reaction	KLM	NRMSE	16	$-.001013$	$.05317$
		MSE	20	$-1.133e{-6}$	$3.433e{-5}$
	RTM	NRMSE	20	$-.01477$	$3.433e{-5}$
		MSE	20	$-1.656e{-4}$	$3.433e{-5}$
	WSINDy	NRMSE	20	$-.02696$	$3.433e{-5}$
		MSE	20	$-1.954e{-5}$	$3.433e{-5}$
Magnet	KLM	NRMSE	13	$-.0002298$	$.02189$
		MSE	20	$-5.488e{-10}$	$3.433e{-5}$
	RTM	NRMSE	20	$-.001534$	$3.433e{-5}$
		MSE	20	$-7.766e{-8}$	$3.433e{-5}$
	WSINDy	NRMSE	14	$-.0004623$	$.02189$
		MSE	20	$-3.890e{-8}$	$3.433e{-5}$

Table 8: Independent-replicate paired comparisons. Differences are ReTRN minus baseline, so negative is favorable to ReTRN. Two-sided Wilcoxon signed-rank tests use each independently simulated data replicate as one observation; Holm correction covers all 18 comparisons jointly. Inference targets the complete validation-selected pipelines.

A.5 Configuration transfer to additional equation families

Transfer with no target-family tuning. We first ask whether model selection can be completed without fitting or selecting on the target equation families. Van der Pol, Duffing, Lorenz, autocatalytic reaction, and competition supply disjoint seven-observation training, checkpoint, and selection trajectories. Six trials select one architecture and learning rate by the geometric mean of displacement-normalized short-rollout error over all five families; the same development objective selects one shared coefficient threshold. The resulting configuration has one residual-tail coordinate, 20 residual blocks, learning rate .0094614, batch size 256, and 30 logarithmically spaced scales in $[10, 10^4]$. It is frozen before any target-family observations are generated. The target vector fields are

$$\dot{x} = -y - z, \quad \dot{y} = x + .2y, \quad \dot{z} = .2 + z(x - 5.7) \quad \text{(Rössler)},$$
$$\dot{x} = y, \quad \dot{y} = -x + yz, \quad \dot{z} = 1 - y^2 \quad \text{(Sprott A)},$$
$$\dot{x} = y, \quad \dot{y} = (1 - z)x - .75y, \quad \dot{z} = x^2 - .45z \quad \text{(Shimizu–Morioka)}.$$

Each family has ten independent fits, each using 800 training and 200 checkpoint trajectories at $t = 0, .05, \ldots, .30$. Target validation applies only the fixed within-training checkpoint rule; it cannot alter any configuration or threshold. All 30 models are hash-frozen before 64 three-second test trajectories per replicate are generated. The constant-inclusive cubic lift has 20 coordinates, of which 19 form the fixed equation space; every true term lies in that space. The ten fits are evaluated separately rather than ensembled.

Unseen family	Complete NRMSE	3 s MSE	Coefficient AUPRC	Runtime (s)
Rössler	$(6.296 \pm 3.389)e{-7}$	$(1.432 \pm .679)e{-13}$	$1.000 \pm .000$	18.78 ± 6.21
Sprott A	$(3.123 \pm 1.218)e{-6}$	$(2.818 \pm 2.531)e{-10}$	$1.000 \pm .000$	16.39 ± 2.26
Shimizu–Morioka	$(1.387 \pm .374)e{-7}$	$(6.477 \pm 7.792)e{-13}$	$1.000 \pm .000$	$14.73 \pm .41$

Table 9: Frozen-configuration transfer (mean $\pm$ sample SD over ten independent fits). The complete NRMSE is measured before thresholding; AUPRC measures coefficient-magnitude ranking. Runtime is descriptive rather than a cross-method comparison.

All 1,920 rollouts remain finite. The learned closure is active: its RMS relative to its analytic closure reference is $.189 \pm .051$, $1.112 \pm .276$, and $1.006 \pm .244$, respectively. Although the shared development-selected zero threshold leaves the rollout equation dense, coefficient magnitudes rank every true term above every false term (AUPRC 1), and each of the five nonzero thresholds declared before target-data generation yields exact support in all 30 fits. An independent DOP853 reintegration changes rollout MSE by at most $.00783\%$; source, data, selection, model-freeze, and test-generation

hashes all pass independent verification. The experiment establishes coefficient ranking, sparse separability, and rollout accuracy under frozen configuration transfer.

Transfer across analytic function classes. We next retain the complete development-selected architecture, optimizer, moment grid, epoch budget, and checkpoint rule, but replace the cubic dictionary by the prescribed trigonometric dictionary appropriate to the damped pendulum $\dot{x} = y$, $\dot{y} = -\sin x - .25y$. This 20-term library contains the two state coordinates and first three sine/cosine harmonics of each coordinate and their difference; its last coordinate initializes closure, while every true term remains in the equation space. The protocol and configuration are locked before generating ten independent sets of 800 training and 200 checkpoint trajectories, each with seven observations. All model matrices are then frozen before generating 64 three-second tests per fit. Complete coefficient NRMSE is $.004015 \pm .001042$, rollout MSE is $(4.328 \pm 2.324)e{-}8$, AUPRC is $1.000 \pm .000$, learned-closure RMS ratio is $1.246 \pm .175$, and all 640 rollouts are finite. Every fit strictly separates true from false magnitudes, with a common post-hoc separating-threshold interval $(.0220, .2557]$; because the predeclared grid ends at .01, only one fit attains exact support on that grid. The interval is a diagnostic rather than a reselected threshold. This experiment transfers optimization settings across polynomial and trigonometric equation/closure spaces.

Validation-selected equation space within an adequate region. We separately test the nested equation-space selection rule stated in Section 3.1 on new data from these four target families. Before generating any short trajectory or fit, we hash-lock the implementation-order dictionaries, candidate closure dimensions $\{1, 5, 9\}$, relative-threshold grid $\{0, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$, and the selection rule. The optimizer, 20-block residual network, learning rate .0094614, batch size 256, 30-scale moment grid, epoch budget, and checkpoint rule are inherited from development families and remain fixed. For each family and ten seeds, 800 seven-observation trajectories train each candidate equation/closure split, 100 disjoint trajectories select the checkpoint, and a separate 40-trajectory partition contributes to the family-level mean normalized rollout used to select one split and threshold. Two other 40-trajectory partitions select and report the gauge diagnostic. This produces 120 ReTRN candidates; no baseline is fit.

All candidate checkpoints and readouts, four family-level selections, and 40 truth-free diagnostics are hash-frozen before standalone coefficient matrices or 301-point tests are created. The selected tail dimensions are $5/1/1/1$ for Rössler/Sprott A/Shimizu–Morioka/damped pendulum; selected relative thresholds are $10^{-4}/10^{-4}/10^{-4}/0$. Post-freeze truth evaluation gives:

Family	Split (q/h)	Dense NRMSE	3 s MSE	AUPRC/F1	Exact support	Warning
Rössler	$5/15$	$(3.466 \pm .905)e{-}7$	$(1.245 \pm .882)e{-}14$	$1/1$	$10/10$	$0/10$
Sprott A	$1/19$	$(3.315 \pm .329)e{-}6$	$(1.102 \pm .995)e{-}10$	$1/1$	$10/10$	$0/10$
Shimizu–Morioka	$1/19$	$(1.290 \pm .431)e{-}7$	$(5.886 \pm 3.100)e{-}14$	$1/1$	$10/10$	$0/10$
Damped pendulum	$1/19$	$.003591 \pm .000631$	$(7.185 \pm 4.930)e{-}8$	$1/.1395$	$0/10$	$0/10$

Table 10: Validation selection over the initial adequate equation-space grid (mean $\pm$ sample SD over ten independent fits). "Warning" applies the pre-existing neural-closure threshold without recalibration.

All 120 candidate equation spaces—not only the 40 selected ones—happen to contain every true row under post-freeze inspection. All 1,280 final rollouts are finite, and every family has AUPRC one. The three polynomial families attain exact support in all 30 fits. Pendulum instead selects the dense threshold on validation: it ranks every true coefficient above every false one but attains no exact support. Mean neural-closure ratios are $1.21e{-}7/8.18e{-}6/2.35e{-}7/8.05e{-}5$ and mean shear coefficient drift is at most $4.55e{-}4$; the transformed state-output field agrees to $9.35e{-}17$. This protocol demonstrates validation transfer among adequate equation spaces, but cannot test whether validation rejects a space that omits a true term.

Testing the equation-space boundary on a new candidate grid. We therefore run a separately locked confirmation on fresh seeds from the same four target families. The full analytic lift retains 20 coordinates, while candidate equation-space dimensions are $\{d, 6, 11, 19\}$; only the complementary closure dimension changes, giving 8,820–9,160 trainable parameters and the same 20×20 operator. All other architecture, optimizer, moment, data-volume, checkpoint, and threshold-grid settings

transfer unchanged. Before coefficient truth or long tests are generated, each candidate checkpoint is frozen and scored on disjoint short validation. Global selection averages the ten deployment seeds; leave-one-seed-out (LOSO) selection excludes the deployment seed from its aggregate. The grid contains 70 inadequate and 90 adequate fits under post-freeze support inspection. Global selection chooses dimension 19 and relative threshold 10^{-4} for every family. LOSO agrees for all polynomial deployments; on pendulum it chooses dimension 11 once and 19 nine times, with thresholds $3e{-}4$ once, $1e{-}4$ five times, and zero four times. Both selectors therefore choose an adequate equation space in all 40 deployments. Global-selection results are:

Family	Split (h/q)	Thresholded NRMSE	3 s MSE	Support F1	Exact support
Rössler	$19/1$	$(1.209 \pm .592)e{-}7$	$(9.187 \pm 3.612)e{-}15$	$1.000 \pm .000$	$10/10$
Sprott A	$19/1$	$(1.677 \pm .607)e{-}6$	$(5.414 \pm 3.359)e{-}11$	$1.000 \pm .000$	$10/10$
Shimizu–Morioka	$19/1$	$(4.365 \pm 1.346)e{-}8$	$(6.553 \pm 4.256)e{-}14$	$1.000 \pm .000$	$10/10$
Damped pendulum	$19/1$	$.003332 \pm .000921$	$(2.695 \pm 1.744)e{-}8$	$.582 \pm .055$	$0/10$

Table 11: Fresh validation selection across candidate equation spaces that cross adequacy (mean $\pm$ sample SD over ten independently trained models). AUPRC is one for every selected model.

All 1,280 selected-model trajectory predictions are finite. In contrast, 375 of 5,120 predictions across all candidates fail the locked divergence criterion, involving 27 of 160 models; every such model uses an inadequate equation space, and the reported candidate-level MSE includes the predeclared failure penalty. Applied independently of selection, the development-calibrated neural-closure warning gives 70 true positives, no false negatives, no false positives, and 90 true negatives. Independent verification recomputes all 960 short-validation scores and all 40 LOSO decisions from the frozen checkpoints.

On pendulum, short-validation selection chooses the 19-function equation space, whose thresholded NRMSE is $.003332$; the six-function space reaches $.0003368$, but coefficient truth was unavailable to selection. More generally, the state-only candidate has lower validation moment objective in all 40 matched comparisons yet is worse in coefficient NRMSE, equation- and complete-field NRMSE, rollout MSE, and neural-closure ratio in every pair. Hence the moment objective trains closure within a fixed lift, while the separate warning diagnostic supplies the equation-space adequacy signal tested here.

Validation-only equation-space selection with abstention. We then evaluate the warning gate on four polynomial vector fields not used in method design, configuration selection, warning calibration, or any preceding experiment. The equations and state domains are fixed during protocol design, while the warning cutoff and readout threshold are transferred from the preceding families without recalibration:

$$\begin{aligned}
&\dot{x} = y && \text{(Genesio–Tesi)},\\
&\dot{y} = z,\\
&\dot{z} = -6x - 2.92y - 1.2z + x^2;\\
&\dot{x} = y - 3x + 2.7yz && \text{(Dadras)},\\
&\dot{y} = 1.7y - xz + z,\\
&\dot{z} = 2xy - 9z;\\
&\dot{x} = -.25x - y^2 - z^2 + 2 && \text{(Lorenz–84)},\\
&\dot{y} = -y + xy - 4xz + 1,\\
&\dot{z} = -z + 4xy + xz;\\
&\dot{x} = y(z - 1 + x^2) + .1x && \text{(Rabinovich–Fabrikant)},\\
&\dot{y} = x(3z + 1 - x^2) + .1y,\\
&\dot{z} = -2z(.14 + xy).
\end{aligned}$$

Initial states are uniform on $[-1,1]^3$, $[-1,1]^3$, $[0,2] \times [-2,2]^2$, and $[-1,1]^2 \times [0,1]$, respectively. Every fit uses the same constant-inclusive 20-term cubic lift, candidate equation-space dimensions $\{3, 6, 11, 19\}$, 20 residual blocks, learning rate $.0094614$, batch size 256, and 30-scale moment grid.

For each family and ten seeds, 800 seven-observation trajectories train the model and 100 disjoint trajectories select the checkpoint. The 120 readout-validation trajectories are divided into mutually exclusive sets of 40 for the warning gate, shear-ridge selection, and diagnostic reporting. The fixed relative readout threshold is .003; the warning threshold is .00257429. The selector returns the smallest equation space with no warning or fit failure over the ten deployment seeds and otherwise abstains. Leave-one-seed-out selection applies the same rule to the other nine seeds.

All 160 fits, warning records, shear diagnostics, and choices are frozen before coefficient arrays and 32 independent 301-point, three-second test trajectories per seed are generated. The locked runner reached the predeclared abstention branch on Lorenz–84; an abstention-safe serializer records the unchanged choice rather than supplying a fallback. Independent verification recomputes every warning, diagnostic, choice, candidate evaluation, and checkpoint readout.

Family	Global split (h/q)	LOSO accepts	Adequate	TP/FN/FP/TN	Fit failures
Genesio–Tesi	$6/14$	$10/10$	$10/10$	$10/0/0/30$	0
Dadras	$11/9$	$10/10$	$10/10$	$20/0/0/20$	0
Lorenz–84	abstain	$1/10$	n/a	$20/0/5/15$	0
Rabinovich–Fabrikant	$19/1$	$10/10$	$10/10$	$30/0/0/10$	0

Table 12: Prospectively locked equation-space analysis. TP/FN/FP/TN compare the frozen warning with post-freeze inspection of whether the candidate space omits a true term. LOSO reports non-abstaining leave-one-seed-out deployments.

Family	Equation dim.	Coefficient NRMSE	3 s MSE	AUPRC/F1	Exact support
Genesio–Tesi	6	$(4.290 \pm .795)e{-9}$	$(2.424 \pm 1.677)e{-15}$	$1/1$	$10/10$
Dadras	11	$(8.677 \pm 3.165)e{-5}$	$(1.664 \pm 2.923)e{-5}$	$1/1$	$10/10$
Lorenz–84	abstain	n/a	n/a	n/a	n/a
Rabinovich–Fabrikant	19	$(2.328 \pm .481)e{-4}$	$(2.564 \pm 1.344)e{-7}$	$1/1$	$10/10$

Table 13: Post-freeze recovery of globally selected equation spaces (mean $\pm$ sample SD over ten independently trained models). NRMSE and support use the fixed .003 relative threshold. All 960 reported trajectory predictions are finite.

Across all candidates, the warning confusion matrix is $80/0/5/75$, giving empirical sensitivity one and specificity .9375 on these correlated finite-sample candidates. All five false positives occur on adequate Lorenz–84 equation spaces and force the global abstention; only the leave-one-out split excluding the single dimension-19 false positive accepts that space. This abstention states that the short-data diagnostic did not certify a candidate consistently across all ten fits, not that post-freeze inspection found the candidate equation space incomplete. The adequacy gate accepts three prospective families, and the selected equation space attains exact support in all 30 fits. The gate detects evidence of neural closure leaking into the equation readout rather than optimizing long-test error: on Dadras, the adequate 19-function space has NRMSE $(.7566 \pm .3583)e{-6}$ and MSE $(3.502 \pm 4.353)e{-9}$, lower than the selected 11-function space in all ten pairs, although both recover exact support. We therefore use the rule only as a conservative, truth-free safeguard when U must be chosen from data, not as a calibrated classifier or as a required component when U is scientifically prescribed.

The same frozen warning also admits a single-data-set decision. For each replicate independently, we choose the smallest successful equation space whose warning is negative, using only that replicate's fitted candidates and its disjoint warning partition; no other replicate, coefficient truth, support truth, or long test enters the decision. The same rule accepts 39/40 deployments and abstains once on Lorenz–84. Post-freeze support inspection finds that all 39 accepted spaces contain every true term. The 30 accepted deployments whose evaluations were generated under the locked protocol recover exact support in all 30 fits; the nine accepted Lorenz–84 spaces are used only for the adequacy analysis because global abstention precluded their long-test evaluation. This isolates deployment information at the level of one observed data set rather than a family aggregate.

A.6 DENSE-GRID COMPONENT ABLATION

We retain the six-system dense-grid data, analytic libraries, equation readout, and long-rollout evaluator and remove one ReTRN component at a time. The *moment-ablated* model keeps the hybrid analytic–neural lift and alternating operator update but trains its closure with ordinary pointwise defect energy. This trainable variant is not forced to use ReTRN's optimization hyperparameters: its closure dimension, residual depth, and learning rate are selected independently from held-out trajectories under its own training objective. The *network-ablated* model removes learned closure and is exactly the corresponding analytic-only estimator. No coefficient truth or long-test trajectory is used for model selection. The comparison therefore evaluates the complete temporal-moment objective against pointwise defect training under objective-specific validation selection; it is not presented as an ablation of only the off-diagonal entries of $W^\top W$.

System	Full ReTRN	No temporal moments	No learned closure
Van der Pol	$1.560e{-4}$	$4.926e{-4}$	$3.148e{-4}$
Duffing	$4.499e{-6}$	$7.964e{-5}$	$4.697e{-6}$
Lorenz	$2.213e{-6}$	$3.268e{-4}$	$6.156e{-6}$
Competition	$4.170e{-5}$	$2.646e{-4}$	$8.142e{-5}$
Reaction	$3.106e{-4}$	$1.585e{-3}$	$5.216e{-4}$
Magnet	$2.555e{-4}$	$4.600e{-4}$	$3.556e{-3}$

Table 14: Complete coefficient NRMSE in the dense-grid component ablation.

System	Full ReTRN	No temporal moments	No learned closure
Van der Pol	$1.669e{-7}$	$1.980e{-6}$	$2.921e{-7}$
Duffing	$4.158e{-8}$	$5.998e{-7}$	$4.247e{-8}$
Lorenz	$2.499e{-5}$	$1.969e{-1}$	$9.371e{-5}$
Competition	$1.948e{-8}$	$4.456e{-7}$	$2.751e{-7}$
Reaction	$3.114e{-8}$	$2.724e{-4}$	$9.782e{-7}$
Magnet	$4.810e{-9}$	$1.426e{-7}$	$2.947e{-6}$

Table 15: Three-second rollout MSE in the dense-grid component ablation. Full ReTRN has the lowest value in every system.

Full ReTRN is lower in all 24 component–system–metric comparisons. Relative to the independently selected no-moment variant, its coefficient NRMSE is 1.80–$148\times$ lower and its rollout MSE is 11.9–$8{,}749\times$ lower. Relative to the analytic no-closure variant, the respective ranges are 1.04–$13.9\times$ and 1.02–$613\times$. Table 4 summarizes the corresponding ablated/full ratios.

A.7 OBJECTIVE-SPECIFIC TEMPORAL/DIAGONAL CONTROL

The preceding ablation compares the complete temporal-moment objective with pointwise defect training. To isolate its cross-transition products, we also compare $G = W^\top W$ with $\mathrm{diag}(G)$, which preserves every marginal time weight but removes products between distinct ordered transitions. Each objective receives the same 18 candidate cores, five finalists, five tuning seeds, short-validation statistic, and final evaluator, while selecting its architecture and optimization settings independently before test access. Twenty final pairs per system share training, checkpoint-validation, readout-validation, optimization seed, library, threshold grid, and 32 new long tests. Coefficient truth and long tests are generated only after both objective-specific configurations and final readouts are frozen.

System	Temporal-moment MSE	Diagonal-Gram MSE	Temporal wins
Lorenz	2.130 ± 1.897	3.404 ± 2.666	$17/20$
Reaction	$(4.429 \pm 1.483)e{-7}$	$(4.569 \pm 1.918)e{-7}$	$11/20$
Magnet	$(1.819 \pm .708)e{-10}$	$(1.989 \pm .735)e{-10}$	$11/20$

Table 16: **Cross-transition objective control.** Three-second rollout MSE, mean $\pm$ sample SD over 20 independent pairs. Lower is better. Lorenz gives a two-sided paired Wilcoxon $p = .000483$ and Holm-adjusted $p = .00290$ across the six prespecified system–metric tests; the other five tests are not significant after correction.

Temporal moments have lower mean rollout MSE on all three systems. The isolated cross-transition effect is clearest on chaotic Lorenz; the Reaction and Magnet differences are smaller. This control does not replace the six-system component ablation: it distinguishes the full Gram geometry from matched diagonal weighting under separately validated optimization, whereas the ablation measures the complete moment objective against ordinary pointwise training.

A.8 Magnitude of the omitted neural derivative

For the equation/closure decomposition in Appendix G.2, we evaluate $r_\theta \widehat{\widehat{L}}_{T,1:d} S_x$ and $UB_U + VB_T$ for the exact 60 ReTRN checkpoints and readouts in Table 2. On all 1,400 independent short-validation states per checkpoint, neural-closure RMS is $.1372 \pm .0804$, $(2.030 \pm .862)e{-}4$, and $(2.190 \pm .188)e{-}5$ for Lorenz, Reaction, and Magnet; analytic-field RMS is 135.50 ± 4.65, $1.983 \pm .108$, and $.5171 \pm .0171$. We first take RMS over states and field components and then report mean $\pm$ sample SD over 20 replicates. Within-replicate neural-closure/analytic ratios have means $.101/.0103/.00424\%$ and maxima $.203/.0160/.00507\%$ in system order. No fit or readout changes, and neither field truth nor long tests enter this sampled-state decomposition diagnostic; it does not establish generator identifiability or global trajectory accuracy.

After fixing every benchmark result, we repeat the calculation on each replicate's full 32×301 test array as a post-hoc scope diagnostic. Omitted/analytic RMS ratios are $(.197 \pm .125)\%$, $(.0567 \pm .0283)\%$, and $(.00521 \pm .00084)\%$ for Lorenz, Reaction, and Magnet, with across-checkpoint maxima $.446/.0991/.00672\%$; omitted-term RMS is $.1256 \pm .0785$, $(.4531 \pm .2324)e{-}3$, and $(1.444 \pm .200)e{-}5$. The contribution therefore remains small over the three-second test distribution, without implying a structural guarantee beyond it.

We also test the finite-sample equation–closure gauge directly. Within each independent replicate, trajectories are divided deterministically into 50% projection fitting, 25% ridge selection, and 25% untouched reporting subsets. Let A be the resulting held-out projection of the raw neural residual onto the equation space. We apply $T' = T - UA$ and the exact similarity transform, for which $B'_U = B_U + AB_T$; hence the state-output field is unchanged although the reported equation readout may move. On Lorenz/Reaction/Magnet, held-out explained residual energy is $80.15 \pm 7.27/97.63 \pm .97/94.56 \pm 1.55\%$. Relative coefficient-matrix drift is $.0353 \pm .0313/.0210 \pm .0144/.0411 \pm .0218\%$, with per-replicate maxima $.0929/.0627/.0831\%$; analytic-field drift is $.0916 \pm .0523/.0103 \pm .0043/.00411 \pm .00056\%$, with maxima $.1775/.0161/.00536\%$. The largest discrepancy between original and transformed state-output fields is $3.86e{-}16$. Thus equation-space directions are present in the learned residual, as the transformation law permits, but their coupling into the state-output block is small on these held-out states. The test uses no coefficient truth or long-rollout data and establishes local empirical stability rather than global identifiability.

Readout stability of the primary fits. We repeat the projection diagnostic on the 120 frozen ReTRN fits underlying Table 1. Each projection uses the original 800 training trajectories; a ridge value is selected on half of the short-validation trajectories and reported on the other half. All projections are frozen before loading the original 500×301 long-test arrays. Table 17 reports mean $\pm$ sample SD across 20 models and the worst replicate in parentheses.

System	Omitted/analytic RMS (%)	Coefficient drift (%)	Analytic-field drift (%)
Van der Pol	$.0207 \pm .0174$ $(.0574)$	$.0136 \pm .0118$ $(.0410)$	$.0188 \pm .0159$ $(.0539)$
Duffing	$.0143 \pm .0012$ $(.0169)$	$.0559 \pm .0104$ $(.0735)$	$.00575 \pm .00076$ $(.00697)$
Lorenz	$.1375 \pm .0457$ $(.2127)$	$.0166 \pm .0064$ $(.0280)$	$.0932 \pm .0297$ $(.1406)$
Competition	$.0312 \pm .0136$ $(.0625)$	$.0707 \pm .0308$ $(.1486)$	$.0192 \pm .0069$ $(.0324)$
Reaction	$.0767 \pm .0422$ $(.1789)$	$.0281 \pm .0152$ $(.0653)$	$.0725 \pm .0401$ $(.1699)$
Magnet	$.00289 \pm .00096$ $(.00513)$	$.0162 \pm .0142$ $(.0486)$	$.00175 \pm .00052$ $(.00257)$

Table 17: Readout stability on the exact primary-table fits. The omitted neural field is relative to analytic-field RMS; the gauge is fitted without coefficient truth. Drift is relative to the original analytic readout.

The six fixed equation spaces contain all $4/4/7/6/5/4$ nonzero true coefficients in system order; no true term is assigned to closure or omitted from the full lift. This support check is post-hoc and does not select a partition. The state-output field is unchanged by the fitted gauge to at most $4.62e{-}16$.

Thus the analysis verifies equation-space adequacy and small data-supported projection sensitivity for these configurations; it does not establish invariance to arbitrary functional mixing or the other rank, smoothness, and spectral assumptions of Theorem 3.3.

A.9 Capacity-matched equation-space misspecification stress

Equation (32) predicts an additional projection term when an active scientific term is assigned to closure rather than the equation space. We test this boundary on 20 independent Duffing replicates with seven observations per trajectory. Both conditions use the same nine-term nonconstant two-state cubic analytic lift, $q_{\mathrm{tail}} = 4$, 2,016 trainable closure parameters, the same 800 training and 200 short-validation trajectories, optimization seed, residual network, temporal-moment objective, checkpoint rule, and three-second test. The misspecified condition places the true x_1^3 term among the final four closure references. The adequate condition exchanges x_1^3 with the zero-coefficient x_2^2 feature across the fixed equation–closure boundary. Coefficients are mapped back to canonical order, preserving the full analytic lift and model dimension. Simulator support is used only to construct this mechanism intervention; numerical coefficients and long-test errors remain excluded from training and readout selection.

All 40 models and validation-selected readouts are frozen before the coefficient truth and existing long-test arrays are accessed. Complete coefficient NRMSE and AUPRC use the zero-padded equation-space readout; support F1 and rollout use a relative threshold selected only on short validation. The neural-closure ratio is $\|r_\theta B_T\|/\|UB_U + VB_T\|$ on all long-test states. Gauge drift projects the learned residual onto the fixed equation space using training states, selects a ridge on half of short validation, and reports on the untouched half. Table 18 gives mean $\pm$ sample SD.

Metric	Term in equation space	Term in closure	Equation space better / tie
Validation moment objective	$.1461 \pm .0332$	$.2253 \pm .0386$	20/0
Coefficient NRMSE	$.00843 \pm .00407$	$1.801 \pm .288$	20/0
Coefficient AUPRC	$1.000 \pm .000$	$.8240 \pm .0229$	20/0
Validation-selected support F1	$.9689 \pm .0670$	$.6851 \pm .0656$	20/0
3 s analytic-equation MSE	$.00705 \pm .00576$	44.24 ± 14.82	20/0
Full-field projection NRMSE	$.00872 \pm .00402$	$.1863 \pm .0327$	20/0
Test projection error (%)	$.0486 \pm .0247$	13.84 ± 1.49	20/0
Omitted neural field (%)	$.0735 \pm .0303$	27.32 ± 2.82	20/0
Held-out coefficient drift (%)	$.341 \pm .204$	70.79 ± 5.65	20/0
Held-out analytic-field drift (%)	$.0345 \pm .0137$	17.05 ± 2.10	20/0

Table 18: Capacity-matched equation-space intervention. "Equation space better" means higher for AUPRC/F1 and lower otherwise. The state-output field remains invariant under the fitted gauge transform to at most $1.10e{-}16$.

Assigning the true cubic to closure increases the validation moment objective by 54.3%, mean NRMSE by $214\times$, and rollout MSE by $6{,}278\times$, while retaining every analytic coordinate in the Koopman lift. A two-sided paired sign test gives raw $p = 1.907e{-}6$ for both readout metrics and Holm-adjusted $p = 3.814e{-}6$. The training-only field projection also deteriorates, showing that the effect persists under a shear-invariant projection. This isolates the misspecification term predicted by Eq. (32): a rich Koopman lift does not repair an inadequate equation space.

Equation-space adequacy across systems. We repeat the capacity-matched intervention on Lorenz and Reaction with ten paired seeds per system. Lorenz exchanges the true x_1x_3 term with zero-coefficient $x_2^2x_3$ across the equation–closure boundary of a complete cubic lift, using $q_{\mathrm{tail}} = 3$ and 3,249 trainable parameters per model. Reaction analogously exchanges the true $x_1x_2^2$ term with zero-coefficient $x_1^3x_2$ in a complete quartic lift, using $q_{\mathrm{tail}} = 4$ and 2,610 parameters. Within each pair, the analytic lift, operator dimension, closure dimension, data, seed, optimization, selection, and evaluator are identical. Neural-closure and shear diagnostics use training states and disjoint short-validation projection-selection/report subsets; their values, checkpoints, and readouts are hash-frozen before coefficient truth and long tests are bound.

The neural-closure, coefficient-drift, and analytic-field-drift diagnostics increase in all ten pairs for both systems (exact two-sided sign test raw $p = .001953$ and Holm-adjusted $p = .01172$ for each of the six comparisons), while the fitted state-output field remains invariant under the shear to about 10^{-16}. Neural-closure and analytic-field-drift ranges do not overlap between adequate

System	Metric	Term in equation space	Term in closure	Closure larger
Lorenz	Validation moment	$1.0133 \pm .0368$	$1.0118 \pm .0368$	$2/10$
	Omitted field (%)	$.01010 \pm .00393$	$1.9597 \pm .7470$	$10/10$
	Coefficient drift (%)	$.00142 \pm .00059$	$.4221 \pm .2322$	$10/10$
	Coefficient NRMSE	$.005321 \pm .000905$	$.017756 \pm .000321$	$10/10$
	Coefficient AUPRC	$1.000 \pm .000$	$.8771 \pm .0000$	$10/10$
	3 s rollout MSE	5.253 ± 4.289	$(2.976 \pm .974)e4$	$10/10$
Reaction	Validation moment	$.002616 \pm .000379$	$.002541 \pm .000401$	$1/10$
	Omitted field (%)	$.01526 \pm .00537$	$3.0540 \pm .4236$	$10/10$
	Coefficient drift (%)	$.03384 \pm .01675$	29.8065 ± 11.4674	$10/10$
	Coefficient NRMSE	$.002952 \pm .000794$	$.443200 \pm .007357$	$10/10$
	Coefficient AUPRC	$1.000 \pm .000$	$.5319 \pm .0117$	$10/10$
	3 s rollout MSE	$(1.651 \pm 1.073)e{-6}$	$1.3461 \pm .1729$	$10/10$

Table 19: Capacity-matched equation-space adequacy test (mean $\pm$ sample SD). Diagnostics are fixed without coefficient truth or long-test access. "Closure larger" denotes the count among ten paired seeds; smaller is better for all listed metrics.

and misspecified conditions within either system; no warning threshold was predeclared, so this is descriptive separation rather than a calibrated classifier. With the declared zero-padded equation readout, coefficient NRMSE and rollout MSE worsen in every pair: by $3.34\times$ and $5.66e3\times$ on Lorenz, and by $150\times$ and $8.15e5\times$ on Reaction. Coefficient AUPRC also falls in all 20 pairs, from one to .8771 and .5319, respectively.

Held-out Magnet adequacy test. We lock ten new Magnet seed pairs and their protocol sources before generating any new trajectory or fit. Magnet exchanges the true $\sin(x_1 - x_2)$ term with zero-coefficient $\sin(3x_2)$ across the equation–closure boundary while retaining the original 20-term trigonometric lift, 16 equation coordinates, four closure coordinates, and 7,120 neural parameters. Each pair shares 800/100/100 independent seven-observation training/checkpoint/readout trajectories and every fit setting. All 20 checkpoints, readouts, and short-validation diagnostics are frozen before coefficient truth and 32 new 301-point long tests per seed are generated.

Metric	Term in equation space	Term in closure	Misspecified worse
Omitted-field ratio R	$.0000272 \pm .0000069$	$.1187 \pm .0417$	$10/10$
Coefficient drift	$.000629 \pm .000216$	$.7768 \pm .6469$	$10/10$
Analytic-field drift	$.0000257 \pm .0000069$	$.09252 \pm .02666$	$10/10$
Validation moment objective	$.001361 \pm .000123$	$.001816 \pm .000377$	$10/10$
Coefficient NRMSE	$.000529 \pm .000204$	$.5479 \pm .3385$	$10/10$
Coefficient AUPRC	1.000 ± 0	$.3760 \pm .1618$	$10/10$
Validation-selected support F1	$.2944 \pm .0878$	$.1355 \pm .0345$	$10/10$
3 s rollout MSE	$(7.914 \pm 4.505)e{-10}$	$.014666 \pm .004423$	$10/10$

Table 20: Held-out Magnet equation-space intervention, mean $\pm$ sample SD over ten pairs. Every reported diagnostic and downstream endpoint worsens when the true coupling term is assigned to closure.

No fit or rollout fails, and all planned seeds remain in the denominators. Neural-closure ratio, both shear drifts, coefficient NRMSE, rollout MSE, AUPRC, and support F1 worsen in every pair; the state-output field remains shear-invariant below $1.25e{-15}$. The validation moment objective also separates all ten pairs. This is one held-out, capacity-matched stress test of the equation-space boundary, not universal certification of equation-space adequacy.

A.10 Matched learned-dictionary construction control

We isolate equation anchoring from neural capacity by comparing ReTRN with a distributed learned dictionary on 20 independent seven-observation replicate pairs for each of Lorenz, Reaction, and Magnet. Both constructions retain exact observed-state coordinates. ReTRN residualizes designated analytic closure references while fixing the equation space; the control applies the same residual network through a fixed dense matrix with orthonormal rows over all non-state analytic features. This mixing has no trainable parameters and preserves residual energy. Every pair shares data, optimization seed, residual network and parameter count, optimizer and budget, alternating refits, temporal objective, analytic library, threshold grid, and evaluator, thereby isolating closure placement.

For both variants, we project the learned state-vector field onto the common analytic library by unregularized least squares on training states only. Thresholds are selected on disjoint short-validation trajectories. All 120 model/readout pairs are frozen before 60 new final tests are generated; source, array, parameter-count, freeze, and evaluation hashes agree in every pair. The comparison therefore tests where the residual capacity is placed, without decoder or test-selection confounding.

System	Construction	coefficient NRMSE	sparse NRMSE	3 s rollout MSE
Lorenz	anchored equation space	$.005549 \pm .000970$	$.005537 \pm .000966$	10.186 ± 4.494
	distributed dictionary	$.032779 \pm .005576$	$.032741 \pm .005736$	82.874 ± 14.711
Reaction	anchored equation space	$.003109 \pm .000650$	$.003109 \pm .000650$	$(.8690 \pm .3950)e{-6}$
	distributed dictionary	$.101462 \pm .051684$	$.101462 \pm .051684$	$(8.464 \pm 4.586)e{-6}$
Magnet	anchored equation space	$(.2186 \pm .0654)e{-3}$	$(.1985 \pm .0803)e{-3}$	$(3.360 \pm 2.480)e{-9}$
	distributed dictionary	$.260255 \pm .016858$	$.260255 \pm .016858$	$(75.13 \pm 47.10)e{-9}$

Table 21: Parameter-matched construction control, mean $\pm$ sample SD over 20 independent pairs. ReTRN is lower in all 20 pairs for all three metrics and systems. Each two-sided Wilcoxon test gives $p = 1.91e{-6}$; Holm adjustment across the nine tests gives $1.72e{-5}$.

The mean relative residual of the training-only analytic projection is $(.4553e{-3}, .00542e{-3}, .00977e{-3})$ for ReTRN and $(9.935e{-3}, 1.060e{-3}, .7983e{-3})$ for the distributed dictionary in system order. The fixed equation coordinates therefore keep learned state dynamics substantially closer to the prescribed analytic span.

We also test the natural alternative suggested by protecting every analytic observable: retain the complete N-term analytic dictionary as equation coordinates and append q learned closure coordinates. The comparator uses the same residual network and trainable-parameter count, system-specific q, optimization schedule, temporal objective, data, validation, analytic projection, threshold grid, and evaluator as ReTRN, while receiving the larger $(N + q)$-dimensional Koopman matrix. Its output gain is selected from $.01, .1, 1$ on five development replicates and then frozen. All 120 model/readout records are fixed before 60 new tests are generated, isolating the equation/closure construction at matched neural capacity.

System	Construction	coefficient NRMSE	3 s rollout MSE
Lorenz	ReTRN	$.005549 \pm .000970$	12.126 ± 5.166
	full analytic + appended closure	$.006802 \pm .001623$	13.190 ± 6.277
Reaction	ReTRN	$.003109 \pm .000650$	$(.8999 \pm .3690)e{-6}$
	full analytic + appended closure	$.012847 \pm .005767$	$(.4976 \pm .6389)e{-6}$
Magnet	ReTRN	$(.2186 \pm .0654)e{-3}$	$(3.497 \pm 2.576)e{-9}$
	full analytic + appended closure	$(.7991 \pm .3837)e{-3}$	$(2.661 \pm 1.628)e{-9}$

Table 22: Complete-analytic-plus-appended-closure control, mean $\pm$ sample SD over 20 independent pairs. ReTRN has lower complete and sparse coefficient NRMSE in every pair for all three systems ($p_{\mathrm{Holm}} = 1.72e{-5}$ for each of six tests). Rollout wins are 12/20, 1/20, and 10/20; only Reaction differs significantly, favoring the appended model ($p_{\mathrm{Holm}} = .00212$).

For an appended lift $[\boldsymbol{\phi}, \psi]$, the complete analytic library remains fixed and the learned coordinates enlarge the Koopman operator. This is a strong predictive alternative, but it does not impose ReTRN's fixed-dimensional residual parameterization of the designated closure block. ReTRN reports coefficients only in the predeclared U space and uses $V + r_\theta$ for closure at fixed lifted dimension. It reduces coefficient NRMSE by 18.4/75.8/72.6% in system order, while the appended model gives the lower Reaction rollout. Together with Table 21, this separates equation fidelity from forecasting capacity: the placement of adaptive closure changes coefficient estimation even when analytic projection and neural capacity are controlled, whereas a larger appended lift can help prediction.

A.11 Four-sample temporal resolution

At $t = 0, .1, .2, .3$, Reaction and Magnet test transfer from seven-observation validation, not resolution-specific retuning. Both ReTRN and WSINDy retain the configurations and sparsity choices selected using five seven-point tuning subsamples. The four- and seven-point evaluations share trajectory indices, analytic libraries, and long tests; the 20 final subsamples are matched across

methods. These are study-specific transfer settings, not the independently selected primary configurations in Table 6. Reaction NRMSE/MSE is $.0442 \pm .0039/.002878 \pm .000082$ for ReTRN and $.2157 \pm .0075/(2.15 \pm 3.39)e9$ for WSINDy; Magnet gives $.00171 \pm .00040/(.943 \pm .718)e{-}6$ and $.00888 \pm .00248/(20.62 \pm 31.36)e{-}6$, respectively. The two systems therefore support an advantage under transferred settings, not a ranking after four-point retuning.

B Scaling, Coupled Dynamics, and Known Inputs

Each setting in Table 23 uses one fit with the same 31-observation Lorenz-selected configuration on an RTX 3500 Ada GPU. Absolute times are comparable within this table only. The separately selected quadratic-library comparison below uses twenty matched fits.

Setting	Terms	Tail parameters	Time (s)	NRMSE	MSE
3D Lorenz/order 3	19	4,902	79.1	$2.06e{-}6$	$4.70e{-}5$
3D/order 4	34	14,892	91.8	$3.16e{-}6$	$4.73e{-}5$
3D/order 5	55	37,950	103.4	$3.83e{-}6$	$1.36e{-}3$
3D/order 6	83	85,158	153.6	$4.05e{-}6$	$2.05e{-}4$
6D/full cubic	83	85,158	226.8	$3.10e{-}6$	$3.83e{-}4$
9D/full cubic	219	582,102	673.0	$2.83e{-}6$	$1.07e{-}4$

Table 23: Scaling record. The 6D/9D rows are independent Lorenz direct sums and therefore measure cost, not cross-state interaction.

Coefficient error remains small throughout the scaling experiment, while dictionary size, neural parameter count, and runtime increase with polynomial order and state dimension. Complete NRMSE contains the factor $1/\sqrt{nd}$, where n is the library size and d the state dimension. We therefore also recover the all-entry relative Frobenius error $\|\widehat{B} - B\|_F/\|B\|_F$ exactly from the same reported matrices. In table order it is $3.79e{-}6, 7.75e{-}6, 1.20e{-}5, 1.55e{-}5, 1.19e{-}5, 1.76e{-}5$. This metric removes dimensional dilution while retaining errors on inactive candidates; the 9D result remains small but is $4.66\times$ the 3D-cubic value. The 3D Lorenz order-2 library uses a configuration selected independently on validation data. Across 20 matched evaluation subsamples, ReTRN obtains coefficient NRMSE $.001622 \pm .000056$ and rollout MSE $.1077 \pm .0512$, compared with KLM values $.001733 \pm .000036$ and $.5274 \pm .1546$. ReTRN is lower on both metrics in all 20 subsamples. Since the order-2 configuration is selected separately, its runtime is not included in the fixed-configuration order sweep of Table 23.

For the coupled-system experiment, two Lorenz systems interact according to $\dot{x}^{(1)} = f_L(x^{(1)}) + \epsilon(x^{(2)} - x^{(1)})$ and symmetrically. The full-cubic library contains 83 terms. Each method selects a relative coefficient threshold on short-validation rollout, after which the selected equation is integrated on 32 independent tests. At $\epsilon = .2$, the ReTRN readout retains all six coupling terms among 29 selected terms; at $\epsilon = .5$, it retains the 20 true terms. Across five ReTRN fits, ReTRN/KLM/RTM complete coefficient NRMSE is $(1.165173 \pm .0000002)e{-}5/1.165494e{-}5/1.285e{-}4$ at $\epsilon = .2$ and $(1.191 \pm .030)e{-}5/1.182e{-}5/1.226e{-}4$ at $\epsilon = .5$. Corresponding rollout MSE is $(5.9248 \pm .0003)e{-}4/7.4465e{-}4/2.934$ and $.06445 \pm .01753/.05035/1.414$. The six coupling coefficients have ReTRN mean absolute errors $(1.3509 \pm .0001)e{-}5$ and $(2.497 \pm .474)e{-}5$, and fitting takes 32.0 ± 1.0 s and 80.1 ± 37.1 s, respectively. On these fixed-data point estimates ReTRN is lower on both metrics at $\epsilon = .2$; the analytic-only comparator is narrowly lower at $\epsilon = .5$. ReTRN SD reflects optimization variability, whereas KLM and RTM are single deterministic fits; no inferential ranking is made.

The 9D coupling test joins three Lorenz subsystems by

$$\dot{x}^{(b)} = f_L(x^{(b)}) + .2\{x^{(b-1)} + x^{(b+1)} - 2x^{(b)}\}, \qquad b \in \{1, 2, 3\},$$

with cyclic indices. It uses 1,000 31-observation training trajectories, disjoint 100-trajectory checkpoint and threshold-selection sets, and 32 independent 301-observation tests. The full-cubic lift has 219 terms and the true vector field has 39 nonzero coefficients, including 18 cross-system terms. One validation-locked ReTRN configuration is evaluated at five optimization seeds. The zero-padded equation-space readout gives coefficient NRMSE $(9.654781 \pm .000003)e{-}6$ and 3 s

MSE $(1.4405 \pm .0046)e{-}5$ in 74.1 ± 1.4 s. Prediction and support are evaluated separately: the former uses the threshold chosen by short-validation MSE, whereas the latter fixes a per-equation relative threshold of 10^{-4}. The support analysis retains 39 terms with 39 true positives, no false positives or negatives, all 18 cross-system terms, coupling-coefficient MAE $(1.50207 \pm .00001)e{-}5$, and sparse-equation MSE $(5.7077 \pm .00014)e{-}3$. No baseline is refitted for this scope extension; the 6D study above remains the matched estimator comparison.

B.1 Known observed control-affine input

We augment the analytic head with prescribed u, x_1u, x_2u features and keep $u(t)$ exogenous. The bilinear Duffing test uses 1% state-scaled observation noise and is defined by

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = x_1 - 3x_1^3 - .5x_2 + u + .2x_1u.$$

Training uses constant controls varying across 1,000 segments; disjoint validation and final suites contain unseen constants, sinusoids, and switching controls. Over five independent ReTRN fits, ReTRN/KLM/RTM coefficient NRMSE is $.005681 \pm .000037/.005527/.017043$. ReTRN estimates u, x_1u, x_2u as $.99822 \pm .00033, .196589 \pm .000003, -.000679 \pm .000008$; final constant/sinusoidal/switching MSE is $(2.139 \pm .013)e{-}4/(2.051 \pm .010)e{-}4/(3.389 \pm .031)e{-}4$, compared with KLM's $3.661e{-}4/4.055e{-}4/5.442e{-}4$ and RTM's $2.930e{-}3/1.487e{-}3/2.277e{-}3$. On these fixed data, the reported ReTRN rollout values are 41.6%, 49.4%, and 37.7% below KLM, while KLM's coefficient NRMSE is 2.8% lower; deterministic baselines are single fits, so this is descriptive rather than inferential. The experiment considers known, observed control-affine inputs represented in a prescribed input library.

C Robustness and Mechanism Diagnostics

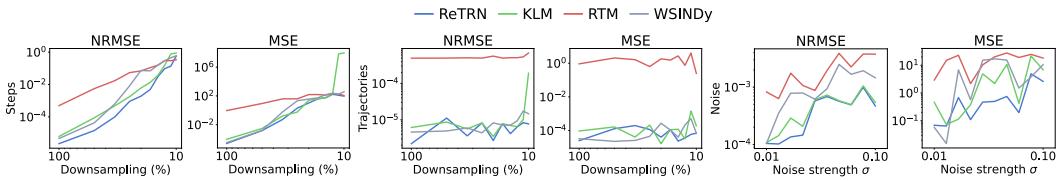


Figure 3: **Lorenz robustness.** Coefficient NRMSE and rollout MSE under reduced temporal resolution, fewer trajectories, and observation noise.

The original robustness experiment perturbs only the Lorenz training data while leaving the held-out evaluation set unchanged. It separately decreases the number of time samples at fixed duration, decreases the number of training trajectories, and adds Gaussian observation noise. ReTRN has the lowest NRMSE at 9/10 temporal-downsampling levels and all 10 noise levels, and the lowest rollout MSE at 9/10 and 6/10 levels, respectively. At the most severe 10% trajectory-count setting, ReTRN obtains NRMSE $7.72e{-}6$ and MSE $6.90e{-}5$, versus KLM's $1.91e{-}4/1.91e{-}4$, RTM's $6.84e{-}4/.252$, and WSINDy's $1.47e{-}5/6.96e{-}5$.

The neural closure is exactly zero at initialization, but it becomes active. Across five held-out Lorenz subsets, trained-closure rollout MSE is 28.24 ± 2.69, versus 35.10 ± 4.36 when zeroed and 34.45 ± 1.47 when shuffled; on Magnet the corresponding values are $7.54e{-}8/2.61e{-}7/2.83e{-}7$. The zeroed and shuffled variants are post-fit perturbations of each selected checkpoint, not independently optimized estimator families; they test information carried by the learned closure. Positive held-out defect R^2 with approximately zero shuffled R^2 supports auxiliary closure information. The reported equation is the head projection UB_U of the fitted state-output channels.

KLM's OLS fit uses every adjacent pair but does not explicitly minimize common-initial finite-horizon response. On three Lorenz partitions, adjacent state MSE is $4.20e{-}6$, whereas common-initial MSE at steps 1/2/3/5/10/20/30 is $1.06e{-}5/6.45e{-}4/8.01e{-}3/.214/13.7/122.9/89.0$. The diagnostic separates local one-step fitting from consistency over repeated propagation.

D EQUATION-READOUT AND TEMPORAL-MOMENT DIAGNOSTICS

Effect of the equation-space projection. As a separate diagnostic, we set KLM coefficients associated with ReTRN's closure-reference positions to zero. Across the six dense-grid systems, projected/full-library KLM NRMSE is $.314758e{-}3/.314759e{-}3$, $.00469655e{-}3/.00469655e{-}3$, $.00616657e{-}3/.00616658e{-}3$, $.0812875e{-}3/.0814277e{-}3$, $.515254e{-}3/.520632e{-}3$, and $3.50455e{-}3/3.55856e{-}3$. These small changes quantify the coefficients carried by the auxiliary analytic positions in KLM. ReTRN, by construction, reports the zero-padded equation-space block; KLM's full and projected values are shown only to make that readout distinction explicit.

Temporal-moment rank and conditioning. For the seven-observation protocol, the normalized 30×6 FOH weight matrix W at $\mu \in [10, 10^4]$ has full column rank six and condition number $5.35e3$. Consequently, Eq. (8) controls the complete ordered defect stack, while the off-diagonal entries of $W^\top W$ couple different transitions during neural closure learning. The six-system ablation in Appendix A.6 evaluates this objective against independently validation-selected pointwise training.

Conditioning of the learned feature and moment designs. We evaluate the feature and moment conditioning of all 60 frozen ReTRN models from the independent study. For each model, its saved training-scaled observable and training trajectories reconstruct $\mathbf{Z}_-$, $\widehat{\mathbf{K}}$, and $\widehat{\mathbf{E}}$ without updating neural parameters. All feature matrices have full column rank. With the shared 30-value μ grid, $\sigma_{\min}(W) = .001016$ and $\sigma_{\max}(W) = 5.437$. Define

$$Q_W = \frac{\|\mathscr{W}\widehat{\mathbf{E}}\|_F}{\sigma_{\min}(W)}, \qquad A_W = \frac{Q_W}{\|\widehat{\mathbf{E}}\|_F}.$$

Across the 60 independently frozen Lorenz, Reaction, and Magnet models, every initial design and full feature matrix has full column rank and admits a common principal-log contour. The sharper first-defect constant in Theorem 3.3 is $b_W = \|e_0^\top W^\dagger\|_2 = .272901$. Reconstructing the moment objective reproduces the recorded training loss to relative discrepancy below $4.93e{-}8$, and recomputing the operator from each fixed checkpoint reproduces its saved analytic coefficient matrix to relative discrepancy below $3.88e{-}9$.

We verify the matrix logarithm for all 60 frozen operators. A conservative acceptance rule routes nonzero `scipy.linalg.logm` error estimates to an eigendecomposition fallback; the zero-jitter estimates are $1.77e{-}15$–$6.05e{-}15$. No eigenvalue is floored, no pseudoinverse is used, and no returned entry is clipped. The fallback and SciPy principal-log generators differ by at most $1.58e{-}13$ relatively, their exponential reconstruction residuals are at most $1.44e{-}13$, and the discarded imaginary-to-real Frobenius ratio is at most $4.57e{-}14$. The minimum eigenvalue modulus is .173, and the smallest angular distance from the negative real axis is 1.512 radians. Thus the implemented readout is numerically equivalent to the principal real logarithm on every independent fit.

E LOGARITHM: PRINCIPAL BRANCH AND SAMPLING-INDUCED AMBIGUITY

This section clarifies two issues underlying KLM-style methods: (i) when the *principal* matrix logarithm is well-defined, and (ii) how branch ambiguity (aliasing) depends on the sampling period T_s.

E.1 PRINCIPAL MATRIX LOGARITHM: EXISTENCE AND POTENTIAL DISCONTINUITY

KLM recovers a continuous-time generator via $\hat{\mathbf{L}} = \frac{1}{T_s}\log(\hat{\mathbf{K}})$, where $\log(\cdot)$ denotes the *principal* matrix logarithm. For an eigenvalue $\mu = re^{i\theta}$ of $\hat{\mathbf{K}}$ with $\theta \in (-\pi, \pi)$, the principal scalar log is $\mathrm{Log}(\mu) = \log r + i\theta$. The principal matrix logarithm selects the corresponding principal arguments.

The principal logarithm is well-defined and continuous as long as the spectrum of $\hat{\mathbf{K}}$ does not intersect the branch cut on the non-positive real axis. Consequently, the log step may become ill-posed or numerically discontinuous when $\hat{\mathbf{K}}$ acquires eigenvalues near/onto the branch cut (e.g., close to negative real values), which may happen under finite-sample noise or ill-conditioning.

Numerical policy. The numerical implementation first evaluates SciPy's Schur-based principal logarithm. It accepts the result only when the returned error estimate is exactly zero, every entry is finite, and $\|\exp(\log \widehat{K}) - \widehat{K}\|_F / \|\widehat{K}\|_F \le 10^{-3}$. Otherwise it retries after diagonal jitters $10^{-12}, \dots, 10^{-9}$ and then uses an eigendecomposition fallback. That fallback floors eigenvalue magnitudes below 10^{-8}, uses a pseudoinverse only if the eigenvector matrix is singular, takes the real part, and clips nonfinite or extreme entries. The first gate is intentionally conservative: for all 60 independently frozen models it rejects error estimates of only $1.77e{-}15$–$6.05e{-}15$ and selects the fallback, while none of the flooring, pseudoinverse, or clipping safeguards activates. The resulting generators agree with SciPy's real principal logarithm to relative error at most $1.58e{-}13$ (Appendix D). These rules specify numerical failure handling; they neither select a nonprincipal physical branch nor identify a materially complex logarithm.

E.2 Branch ambiguity and aliasing: relationship to the sampling period T_s

Even when the principal logarithm exists, mapping from discrete-time to continuous-time is not unique in general. For an eigenvalue λ of a continuous-time generator $\mathcal{L}$, the corresponding discrete-time eigenvalue satisfies

$$\mu = e^{\lambda T_s}. \tag{12}$$

Hence all

$$\lambda^{(k)} = \lambda + \frac{2\pi i k}{T_s}, \qquad k \in \mathbb{Z}, \tag{13}$$

produce the same μ. The principal logarithm selects the representative whose imaginary part lies in $(-\pi/T_s, \pi/T_s)$, which is the Koopman-spectrum analogue of the Nyquist band.

Therefore, if intrinsic oscillation frequencies exceed π/T_s, the principal-log mapping necessarily folds frequencies (aliasing), and different continuous-time dynamics may become indistinguishable from sampled data Zeng et al. (2024); sufficiently small T_s is required for unambiguous spectral identification. ReTRN uses the same principal-log readout and therefore does not remove branch ambiguity or solve aliasing. Its learned closure and weighted residual may reduce finite-sample sensitivity of the fitted discrete-time surrogate before the logarithm step, but the experiments do not establish sampling-invariant spectra or a branch-selection guarantee.

F Moment Control of Empirical Propagation and Compression

Freeze an observable Z and fitted transition matrix $\widehat{K}$. For R trajectories, write $Y_j = [Z(\Phi^{jh}(x_r))]_{r=1}^R$, $E_k = Y_{k+1} - Y_k \widehat{K}$, and stack $E = [E_0^\top, \dots, E_{J-1}^\top]^\top$. Let $\epsilon_W = \|(W \otimes I_R)E\|_F$ and assume W has full column rank, with $s_W = \sigma_{\min}(W) > 0$.

Proposition F.1 (Moment control of empirical propagation and compression)**.** *For $1 \le j \le J$,*

$$\|Y_j - Y_0 \widehat{K}^j\|_F \le \frac{g_j(\widehat{K})}{s_W} \epsilon_W, \qquad g_j(\widehat{K})^2 = \sum_{s=0}^{j-1} \|\widehat{K}^s\|_2^2. \tag{14}$$

If Y_0 has full column rank, set $s_Y = \sigma_{\min}(Y_0)$ and $C_j = Y_0^\dagger Y_j$. For arbitrary scalar weights a_j,

$$\left\| \sum_{j=0}^{J} a_j (C_j - \widehat{K}^j) \right\|_F \le \frac{\epsilon_W}{s_W s_Y} \sum_{j=1}^{J} |a_j| g_j(\widehat{K}). \tag{15}$$

Let b_j be the standard nodal FOH hat functions on $0, h, \dots, Jh$, including the final-state function b_J. With $a_j = \mu^2 \int_0^{Jh} e^{-\mu t} b_j(t) dt$, Eq. (15) bounds the empirical compressed and fitted FOH-discretized finite-horizon resolvent surrogates; the same bound holds after subtracting μI.

Proof. Repeated substitution gives $Y_j - Y_0 \widehat{K}^j = \sum_{k=0}^{j-1} E_k \widehat{K}^{j-1-k}$. Cauchy–Schwarz over time blocks and the singular values of $W \otimes I_R$ prove Eq. (14). Full column rank gives $C_j - \widehat{K}^j = Y_0^\dagger (Y_j - Y_0 \widehat{K}^j)$; applying $\|Y_0^\dagger\|_2 = 1/s_Y$ and the triangle inequality proves Eq. (15). The FOH conclusion follows from linearity of interpolation and integration. □

For the initial defect, full column rank gives the sharper identity

$$E_0 = [(e_0^\top W^\dagger) \otimes I_R]\,[(W \otimes I_R)E], \qquad \|E_0\|_F \le b_W \epsilon_W, \quad b_W = \|e_0^\top W^\dagger\|_2 \le s_W^{-1}.$$

Combining this identity with $C_1 - \widehat{K} = Y_0^\dagger E_0$ gives the moment-to-compression step in Theorem 3.3. This is an empirical, sample-defined compression. Passing to a population projection requires a separate approximation term; continuous-time resolvent comparison additionally requires the FOH interpolation remainder.

G Population Finite-Horizon Decomposition

For this population statement, assume that $\mathcal{K}^t$ is a strongly continuous semigroup on the stated function space, generated by a complete forward flow on a forward-invariant $\overline{\Omega}$. Let $\mathcal{V}_{N,m} \subset D(\mathcal{L})$ be the learned span and $\mathcal{P}_{N,m}$ a bounded projection. Define $\widehat{\mathcal{L}}_{\mu,\tau_s,N,m} := \mu^2 \int_0^{\tau_s} e^{-\mu t} e^{t\widehat{\mathcal{L}}_{N,m}}\,dt - \mu I$, $\mathcal{L}_{\mu,\tau_s,N,m} := \mathcal{P}_{N,m}(\mu^2 \int_0^{\tau_s} e^{-\mu t}\mathcal{K}^t dt - \mu I)$, and $\mathcal{L}_{N,m} := \mathcal{P}_{N,m}\mathcal{L}$.

Theorem G.1 (Conditional finite-horizon decomposition for the compressed generator)**.** *For any $v \in \mathcal{V}_{N,m}$, suppose $\|\widehat{\mathcal{L}}_{\mu,\tau_s,N,m}v - \mathcal{L}_{\mu,\tau_s,N,m}v\| \le \varepsilon_{\mathrm{res}}(v;\mu,\tau_s)$. Then*

$$\left\|\widehat{\mathcal{L}}_{\mu,\tau_s,N,m}v - \mathcal{L}_{N,m}v\right\| \le \varepsilon_{\mathrm{res}}(v;\mu,\tau_s) + \varepsilon_{\mathrm{bridge}}(v;\mu,\tau_s), \tag{16}$$

where, for $\|\mathcal{K}^t\| \le Ce^{\kappa t}$ and $\mu > \kappa$,

$$\varepsilon_{\mathrm{bridge}} = \|\mathcal{P}_{N,m}\|\frac{C\mu^2}{\mu-\kappa}e^{-(\mu-\kappa)\tau_s}\|v\| + \eta_\mu(v), \qquad \eta_\mu(v) \to 0 \quad \text{as } \mu \to \infty.$$

Here $\varepsilon_{\mathrm{res}}$ is an explicit premise, not the empirical loss. Proposition F.1 can instantiate an empirical analogue only for the same sample-defined projection and with interpolation and conditioning terms retained.

Proof. Fix an outer iteration index m, the lifting dimension N, the resolvent parameter μ, and an arbitrary $v \in \mathcal{V}_{N,m} \subset D(\mathcal{L})$. Algorithm 1 returns the matrix-log generator $\widehat{\mathcal{L}}_{N,m}$; for the theorem we construct from it the distinct finite-horizon surrogate $\widehat{\mathcal{L}}_{\mu,\tau_s,N,m}$ defined above. The target compressed generator on the learned span is $\mathcal{L}_{N,m}$.

By adding and subtracting the compressed finite-horizon target $\mathcal{L}_{\mu,\tau_s,N,m}v$, we have

$$\begin{aligned}\left\|\hat{\mathcal{L}}_{\mu,\tau_s,N,m}v - \mathcal{L}_{N,m}v\right\| \leqslant{}& \left\|\hat{\mathcal{L}}_{\mu,\tau_s,N,m}v - \mathcal{L}_{\mu,\tau_s,N,m}v\right\| \\ &+ \|\mathcal{L}_{\mu,\tau_s,N,m}v - \mathcal{L}_{N,m}v\|\,.\end{aligned} \tag{17}$$

The first term is the residual approximation error and is bounded directly by the premise of Theorem G.1, namely

$$\left\|\hat{\mathcal{L}}_{\mu,\tau_s,N,m}v - \mathcal{L}_{\mu,\tau_s,N,m}v\right\| \leqslant \varepsilon_{\mathrm{res}}(v;\mu,\tau_s). \tag{18}$$

It remains to bound the second term in (17). We insert the infinite-horizon compressed Yosida target $\mathcal{L}_{\mu,\infty,N,m}$ and obtain

$$\begin{aligned}\|\mathcal{L}_{\mu,\tau_s,N,m}v - \mathcal{L}_{N,m}v\| \leqslant{}& \|\mathcal{L}_{\mu,\tau_s,N,m}v - \mathcal{L}_{\mu,\infty,N,m}v\| \\ &+ \|\mathcal{L}_{\mu,\infty,N,m}v - \mathcal{L}_{N,m}v\|\,.\end{aligned} \tag{19}$$

The first term in (19) is the finite-horizon truncation error. By the definitions of $\mathcal{L}_{\mu,\tau_s,N,m}$ and $\mathcal{L}_{\mu,\infty,N,m}$,

$$\mathcal{L}_{\mu,\infty,N,m}v - \mathcal{L}_{\mu,\tau_s,N,m}v = \mu^2 \int_{\tau_s}^{\infty} e^{-\mu t}\,\mathcal{P}_{N,m}\mathcal{K}_t v\,dt. \tag{20}$$

Assume that the Koopman semigroup satisfies the growth bound

$$\|\mathcal{K}_t\| \leqslant Ce^{\kappa t}, \qquad t \geqslant 0,$$

and take $\mu > \kappa$. Since $\mathcal{P}_{N,m}$ is bounded, (20) gives

$$\begin{aligned}\|\mathcal{L}_{\mu,\infty,N,m}v - \mathcal{L}_{\mu,\tau_s,N,m}v\| &\leqslant \mu^2 \int_{\tau_s}^{\infty} e^{-\mu t}\,\|\mathcal{P}_{N,m}\mathcal{K}_t v\|\,dt \\ &\leqslant \|\mathcal{P}_{N,m}\|\,C\,\mu^2 \int_{\tau_s}^{\infty} e^{-(\mu-\kappa)t}\,dt\,\|v\| \\ &= \|\mathcal{P}_{N,m}\|\frac{C\mu^2}{\mu-\kappa}e^{-(\mu-\kappa)\tau_s}\|v\|. \end{aligned} \tag{21}$$

The second term in (19) is the Yosida approximation error on the learned span. Using the infinite-horizon resolvent representation,

$$\mathcal{L}_{\mu,\infty,N,m}v = \mathcal{P}_{N,m}\left(\mu^2 R(\mu;\mathcal{L})v - \mu v\right),$$

and the generator-first compression

$$\mathcal{L}_{N,m}v = \mathcal{P}_{N,m}\mathcal{L}v,$$

we have

$$\begin{aligned}\|\mathcal{L}_{\mu,\infty,N,m}v - \mathcal{L}_{N,m}v\| &= \left\|\mathcal{P}_{N,m}\left(\mu^2 R(\mu;\mathcal{L})v - \mu v - \mathcal{L}v\right)\right\| \\ &=: \eta_\mu(v). \end{aligned} \tag{22}$$

Since $v \in D(\mathcal{L})$, the standard Yosida approximation for the generator $\mathcal{L}$ gives

$$\mu^2 R(\mu;\mathcal{L})v - \mu v \longrightarrow \mathcal{L}v, \qquad \mu \to \infty.$$

Because $\mathcal{P}_{N,m}$ is bounded, (22) implies

$$\eta_\mu(v) \to 0, \qquad \mu \to \infty. \tag{23}$$

Combining (17), (18), (19), (21), and (22), we obtain

$$\left\|\hat{\mathcal{L}}_{\mu,\tau_s,N,m}v - \mathcal{L}_{N,m}v\right\| \leqslant \varepsilon_{\mathrm{res}}(v;\mu,\tau_s) + \|\mathcal{P}_{N,m}\|\frac{C\mu^2}{\mu-\kappa}e^{-(\mu-\kappa)\tau_s}\|v\| + \eta_\mu(v). \tag{24}$$

Equivalently, if we define

$$\varepsilon_{\mathrm{bridge}}(v;\mu,\tau_s) := \|\mathcal{P}_{N,m}\|\frac{C\mu^2}{\mu-\kappa}e^{-(\mu-\kappa)\tau_s}\|v\| + \eta_\mu(v),$$

then

$$\left\|\hat{\mathcal{L}}_{\mu,\tau_s,N,m}v - \mathcal{L}_{N,m}v\right\| \leqslant \varepsilon_{\mathrm{res}}(v;\mu,\tau_s) + \varepsilon_{\mathrm{bridge}}(v;\mu,\tau_s).$$

Together with (23), this proves the claim. □

G.1 An Exact OLS Defect Identity and Matrix-Log Stability

Fix the learned observable after training and stack all R trajectories into $\mathbf{Z}_- \in \mathbb{R}^{RJ\times m}$ and $\mathbf{Z}_+ \in \mathbb{R}^{RJ\times m}$. The returned OLS operator is $\widehat{\mathbf{K}} = \mathbf{Z}_-^\dagger \mathbf{Z}_+$. For any reference matrix $\mathbf{K}_\star$, define $\widehat{\mathbf{E}} = \mathbf{Z}_+ - \mathbf{Z}_-\widehat{\mathbf{K}}$ and $\mathbf{E}_\star = \mathbf{Z}_+ - \mathbf{Z}_-\mathbf{K}_\star$.

Proposition G.2 (OLS defect identity and matrix-log stability for a fixed lift)**.** *Assume $\mathbf{Z}_-$ has full column rank. Then $\mathbf{Z}_-^\dagger\widehat{\mathbf{E}} = 0$ and $\widehat{\mathbf{K}} - \mathbf{K}_\star = \mathbf{Z}_-^\dagger\mathbf{E}_\star$. Suppose further that the spectra of $\widehat{\mathbf{K}}$ and $\mathbf{K}_\star$ lie inside a common positively oriented contour Γ whose interior and boundary lie in $\mathbb{C}\setminus(-\infty,0]$. With $\widehat{\mathbf{L}} = T_s^{-1}\log\widehat{\mathbf{K}}$ and $\mathbf{L}_\star = T_s^{-1}\log\mathbf{K}_\star$, both using the principal logarithm,*

$$\|\widehat{\mathbf{L}} - \mathbf{L}_\star\|_F \leq \frac{C_\Gamma}{T_s}\|\mathbf{Z}_-^\dagger\mathbf{E}_\star\|_F \leq \frac{C_\Gamma\|\mathbf{E}_\star\|_F}{T_s\,\sigma_{\min}(\mathbf{Z}_-)}, \tag{25}$$

where

$$C_\Gamma := \frac{\mathrm{len}(\Gamma)}{2\pi}\max_{z\in\Gamma}|\log z|\,\max_{z\in\Gamma}\|(zI-\widehat{\mathbf{K}})^{-1}\|_2\max_{z\in\Gamma}\|(zI-\mathbf{K}_\star)^{-1}\|_2. \tag{26}$$

In particular, $\mathbf{E}_\star = 0$ implies $\widehat{\mathbf{K}} = \mathbf{K}_\star$ and $\widehat{\mathbf{L}} = \mathbf{L}_\star$.

Proof. Full column rank gives $\mathbf{Z}_-^\dagger \mathbf{Z}_- = I$. Therefore

$$\mathbf{Z}_-^\dagger \widehat{\mathbf{E}} = \mathbf{Z}_-^\dagger \mathbf{Z}_+ - \widehat{\mathbf{K}} = 0, \qquad \mathbf{Z}_-^\dagger \mathbf{E}_\star = \widehat{\mathbf{K}} - \mathbf{K}_\star.$$

The holomorphic functional calculus for the principal logarithm and the resolvent identity give (Higham, 2008)

$$\log \widehat{\mathbf{K}} - \log \mathbf{K}_\star = \frac{1}{2\pi i} \int_\Gamma (\log z)(zI - \widehat{\mathbf{K}})^{-1}(\widehat{\mathbf{K}} - \mathbf{K}_\star) \cdot (zI - \mathbf{K}_\star)^{-1}\, dz. \tag{27}$$

Taking Frobenius norms bounds the right-hand side by $C_\Gamma \|\widehat{\mathbf{K}} - \mathbf{K}_\star\|_F$. Dividing by T_s, using the exact OLS identity, and then applying $\|\mathbf{Z}_-^\dagger\|_2 = 1/\sigma_{\min}(\mathbf{Z}_-)$ proves Eq. (25). □

The fixed-lift result depends on feature conditioning, reference closure, and separation from the principal-log branch cut. The empirical residual does not add a term because OLS makes it orthogonal to the feature design. The moment loss acts earlier, through optimization of the lift; Eq. (8) describes its residual geometry, while Section 4.2 tests whether that geometry improves held-out recovery. Thus the proposition directly addresses the implemented readout without converting a training objective into a statistical identification guarantee.

G.2 Equation-Space Readout and Dependence on the Learned Lift

The synthetic-study readout selects the state-output columns and then retains only the rows associated with the prescribed equation space. If S is the diagonal matrix of training feature scales and S_x its state block, the unscaled full generator block is $B = S^{-1}\widehat{\widetilde{L}}_{:,1:d} S_x = [B_U^\top, B_T^\top]^\top$. For $\widetilde{Z} = [U, V + r_\theta]S^{-1}$, the learned state-output field is $UB_U + (V + r_\theta)B_T$, whereas the reported equation is UB_U. We therefore evaluate both the full closure contribution and its learned component; a small moment loss alone does not bound either diagnostic.

Write the unscaled fitted observable as $\mathbf{Z} = [\mathbf{U}, \mathbf{T}]$, where $\mathbf{T} = \mathbf{V} + \mathbf{r}_\theta$, and assume the sampled design $\mathbf{Z}_-$ has full column rank. For $S = \mathrm{diag}(I, Q)$ with invertible Q, the least-squares identity $(\mathbf{Z}_- S)^\dagger = S^{-1}\mathbf{Z}_-^\dagger$ gives $\widehat{\mathbf{K}}' = S^{-1}\widehat{\mathbf{K}}S$ and, whenever the same principal logarithm is defined, $\widehat{\mathbf{L}}' = S^{-1}\widehat{\mathbf{L}}S$. Consequently,

$$\widehat{\mathbf{L}}'_{1:N-q_{\mathrm{tail}},\,1:d} = \widehat{\mathbf{L}}_{1:N-q_{\mathrm{tail}},\,1:d}.$$

This invariance concerns the head subblock only, not the entire reported coefficient matrix. More generally, full column rank implies that any invertible linear change that leaves $\mathbf{U}$ pointwise fixed has the block form $D = \left[\begin{smallmatrix} I & A \\ 0 & Q \end{smallmatrix}\right]$ with Q invertible. Since $D^{-1} = \left[\begin{smallmatrix} I & -AQ^{-1} \\ 0 & Q^{-1} \end{smallmatrix}\right]$, its head block is $B'_U = B_U - AQ^{-1}B_T$. If $A \neq 0$, some B_T makes this difference nonzero; hence $A = 0$ is the maximal subgroup preserving every head coefficient. The complementary directions add analytic functions to the tail. Define $D_A = \left[\begin{smallmatrix} I & -A \\ 0 & I \end{smallmatrix}\right]$ and $\mathbf{Z}_A = \mathbf{Z}D_A = [\mathbf{U}, \mathbf{T} - \mathbf{U}A]$. The similarity law gives

$$B_A = D_A^{-1}B = \begin{bmatrix} B_U + AB_T \\ B_T \end{bmatrix}, \qquad \mathbf{Z}_A B_A = \mathbf{Z}B. \tag{28}$$

Thus a head–tail shear may change the named head coefficients while preserving the learned state-output field.

The coordinate readout above is distinct from projecting the complete learned field onto a fixed analytic design. On fixed fitting states, let Φ contain the full analytic library, assume it has full column rank, write $Z = \Phi + [0, R]$, and define $C_{\mathrm{proj}} = \Phi^\dagger Z B$.

Corollary G.3 (Field projection versus coordinate readout)**.** *The fixed-sample least-squares field projection satisfies*

$$C_{\mathrm{proj}} = B + \Phi^\dagger R B_T, \qquad C_{\mathrm{proj},A} = \Phi^\dagger Z_A B_A = C_{\mathrm{proj}}, \tag{29}$$

for every exact shear $Z_A = ZD_A$, $B_A = D_A^{-1}B$, although the coordinate readout B_A may differ from B.

Proof. Substitute $Z = \Phi + [0, R]$ and use $\Phi^\dagger \Phi = I$ for the first identity. The second follows from $Z_A B_A = ZB$. □

Corollary G.4 (Non-uniqueness of coordinate-block readout at fixed learned field)**.** *Under Proposition 3.1, suppose $B_T \neq 0$ and the admissible representation family contains $\{\mathbf{Z}_{\lambda A_0} : \lambda \in I\}$ for an interval I with nonempty interior and some A_0 satisfying $A_0 B_T \neq 0$. Consider an objective*

$$\mathcal{R}(\lambda) = \Psi(\{\mathbf{Z}_{\lambda A_0}(x) B_{\lambda A_0} : x \in \mathcal{X}\}, \mathcal{D}),$$

where $\mathcal{D}$ is independent of λ. Then $\mathcal{R}$ is constant on I, although $B_U + \lambda A_0 B_T$ contains a continuum of distinct head-coefficient matrices. Hence observations or losses depending only on the learned state-output field cannot identify the coordinate head block within this gauge family.

Proof. Equation (28) gives $\mathbf{Z}_{\lambda A_0} B_{\lambda A_0} = \mathbf{Z}B$ for every $\lambda \in I$, whereas $A_0 B_T \neq 0$ makes $B_U + \lambda A_0 B_T$ nonconstant. Substitution into Ψ proves the claim. □

Corollary G.4 is an obstruction result, not an identifiability claim for ReTRN. The method fixes the equation coordinates $\mathbf{U}$, the analytic closure references $\mathbf{V}$, feature scaling, tail dimension, and the parameterization $\mathbf{T}_\theta = \mathbf{V} + \mathbf{r}_\theta$, with $\mathbf{r}_0 = 0$ as the analytic reference. It evaluates the closure objective in this predeclared coordinate system. Indeed, shear-transformed transition defects satisfy

$$\mathbf{E}_{k,A} = \mathbf{E}_k D_A, \qquad (W \otimes I_R)\mathbf{E}_A = \big((W \otimes I_R)\mathbf{E}\big) D_A, \tag{30}$$

so their Frobenius loss is generally not shear invariant. In particular, write $M = (W \otimes I_R)\mathbf{E} = [M_U, M_T]$. Then

$$\mathcal{J}(A) = \|M D_A\|_F^2 = \|M_U\|_F^2 + \|M_T - M_U A\|_F^2. \tag{31}$$

If M_U has full column rank, this quadratic has the unique minimizer $A_\star = (M_U^\top M_U)^{-1} M_U^\top M_T = M_U^\dagger M_T$. Otherwise its minimizers are $M_U^\dagger M_T + N$ with $M_U N = 0$, and $M_U^\dagger M_T$ is the minimum-norm representative. This states the residual symmetry exactly: the risk distinguishes a shear only through $M_U A$. Fixed residualization, the coordinate metric, and zero reference therefore define an operational estimator convention for the equation-space readout; they do not make the analytic–neural closure decomposition unique. In particular, the network can still contain head functions. The off-diagonal entries of $W^\top W$ separately couple defects from different ordered transitions: they organize observed FSST transitions into a closure-training signal but add neither observations nor structural identification. At finite sampling intervals, the equation readout can depend on closure whenever B_T is nonzero. This dependence vanishes for an exact generator representation of a field lying in the equation space, because its closure-to-state block is zero. The following statement makes the corresponding small-sampling-interval condition precise.

Equation-space adequacy is substantive. If instead the true field is represented only in the full analytic initialization, $f = \boldsymbol{\phi} B_\star$, with an active term assigned to the closure block, let $R_0 = [r_\theta(x_i)]_i$ denote the sampled neural tail and $(B_\star)_{T,:}$ the analytic-closure rows. For $D_n = Y_0^\dagger Y'(0)$, the state columns satisfy the general identity

$$(D_n)_{:,1:d} = Y_0^\dagger Y_0 B_\star - Y_0^\dagger R_0 (B_\star)_{T,:}. \tag{32}$$

When Y_0 has full column rank, $Y_0^\dagger Y_0 = I$ and only the second term remains as bias. Thus assigning a true equation term to closure introduces a projection term that can persist as $T_s \to 0$. Its magnitude depends on the learned residual and the active closure coefficient; neither zero transition loss nor tail-only gauge invariance forces it to vanish. The main theorem deliberately assumes equation-space adequacy, while the experiments report neural-closure magnitude and finite-sample gauge drift for the chosen partitions.

Corollary G.5 (Equation-space misspecification bias)**.** *Retain the assumptions of Theorem 3.3, except replace equation-space adequacy by $f = \boldsymbol{\phi} B_\star$ with an active closure-block row. Let*

$$\widehat{B} = \widehat{L}_{:,1:d}, \qquad \Pi_\theta = Y_0^\dagger R_0 (B_\star)_{T,:}, \qquad a_h = \frac{C_\Gamma b_W \epsilon_W}{h s_Y} + \delta_h,$$

where δ_h is defined in Eq. (9)*. Then*

$$\|\widehat{B} - B_\star + \Pi_\theta\|_F \le a_h, \qquad (\|\Pi_\theta\|_F - a_h)_+ \le \|\widehat{B} - B_\star\|_F \le \|\Pi_\theta\|_F + a_h. \tag{33}$$

For a fixed lift and initial design, if $a_h \to 0$, then $\widehat{B} - B_\star \to -\Pi_\theta$. A nonvanishing limiting bias therefore requires $\Pi_\theta \neq 0$; a nonzero active closure coefficient alone is insufficient.

Proof. The Taylor and logarithm arguments used for Theorem 3.3 bound $\|\widehat{B} - (D_n)_{:,1:d}\|_F$ by a_h without requiring equation-space adequacy. Equation (32) gives $(D_n)_{:,1:d} = B_\star - \Pi_\theta$. Substitution proves the first inequality; the triangle and reverse-triangle inequalities give the second. □

Proposition G.6 (Consistency of equation readout for a fixed lift)**.** *Fix a smooth observable $\mathbf{Z} = [\mathbf{U}, \mathbf{T}]$ with the state coordinates first, and sampled initial states $x_1, \dots, x_n$. Let the design $Y(t) = [\mathbf{Z}(\Phi^t(x_r))]_{r=1}^n$ satisfy $Y(0)$ full column rank and $\sup_{0\le t\le h_0} \|Y''(t)\|_F \le M$. Assume $\mathbf{f}(x) = \mathbf{U}(x)B_\star$ on these trajectories. Define $K_h = Y(0)^\dagger Y(h)$, $L_h = h^{-1}\log K_h$, $H_h = K_h - I$, and $D = Y(0)^\dagger Y'(0)$. Whenever the principal logarithm exists, let $\delta_{\log,h} = h^{-1}\|[\log K_h - H_h]_{:,1:d}\|_F$. Then*

$$\left\| (L_h)_{:,1:d} - \begin{bmatrix} B_\star \\ 0 \end{bmatrix} \right\|_F \le \frac{hM}{2\sigma_{\min}(Y(0))} + \delta_{\log,h}. \tag{34}$$

For the fixed design and observable, the right-hand side is $O(h)$ as $h \to 0$.

Proof. Taylor's formula and the derivative bound give $Y(h) = Y(0) + hY'(0) + R_h$ with $\|R_h\|_F \le h^2 M/2$. Since $Y(0)^\dagger Y(0) = I$, we have $K_h = I + hD + Y(0)^\dagger R_h$ and therefore $\|H_h/h - D\|_F \le hM/[2\sigma_{\min}(Y(0))]$. The first d columns of $Y'(0)$ equal the sampled vector field, namely $Y(0)[B_\star^\top, 0]^\top$, so $D_{:,1:d} = [B_\star^\top, 0]^\top$. Combining the two bounds proves Eq. (34); $H_h = hD + O(h^2)$ and analyticity of $\log(I + H)$ near $H = 0$ give $\delta_{\log,h} = O(h)$ and hence the stated rate. □

Corollary G.7 (Conditional consistency across post-training lifts)**.** *Let $h_n \to 0$ and let $Z_n = [U, T_n]$ be any sequence of post-training lifts with the same named head U and coefficient matrix $B_\star$. For each n, assume the hypotheses of Theorem 3.3, and index its quantities by n. Suppose $\inf_n s_{Y,n} > 0$; $C_{\Gamma,n}$, $b_{W,n}$, $M_{2,n}$, and the feature/state rescaling multiplier are uniformly bounded; $\epsilon_{W,n}/h_n \to 0$; and $\delta_{\log,h_n} \to 0$. Then the named-column error, after undoing scaling, converges to zero.*

Proof. Apply Eq. (9) to each n. Its three terms are, respectively, $o(1)$, $O(h_n)$, and $o(1)$; the uniformly bounded rescaling multiplier preserves convergence. □

The lifts may vary with n, but every condition above is imposed after training. The corollary is therefore a consistency criterion, not a guarantee that alternating optimization realizes such a sequence. It does not establish consistency at a fixed FSST resolution, with observation noise, or under arbitrary neural training. Full column rank also excludes exact head–tail redundancy, and poor conditioning can make the finite-sample bound uninformative.

H Recovered-Flow Truncation Beyond the Observation Horizon

Assumption H.1 (Well-posed flows)**.** $\Omega \subset \mathbb{R}^d$ is pre-compact and forward invariant under both $\mathbf{f}$ and $\hat{\mathbf{f}}$; on an open neighborhood of $\overline{\Omega}$, both fields are locally Lipschitz and their solutions from Ω exist on $[0, \infty)$.

Write $\mathcal{K}^t g = g \circ \Phi^t$ and $\mathcal{K}_{\hat{f}}^t g = g \circ \hat{\Phi}^t$, and define the missing-horizon recovered-flow discrepancy

$$\mathcal{T}_{\mu,\tau_s}^{\mathrm{readout}} g := \mu^2 \int_{\tau_s}^{\infty} e^{-\mu t}(\mathcal{K}^t - \mathcal{K}_{\hat{f}}^t) g \, dt.$$

Theorem H.2 (Conditional recovered-flow truncation bound)**.** *Under Assumption H.1, denote $\varepsilon := \|\hat{\mathbf{f}} - \mathbf{f}\|_\infty$ and the one-sided Lipschitz constant*

$$\alpha := \sup_{\mathbf{x}\ne\mathbf{y}\in\Omega} \frac{\langle \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle}{\|\mathbf{x} - \mathbf{y}\|^2}.$$

In the induced norm from Lipschitz observables to continuous functions, for $\mu > \max\{\alpha, 0\}$,

$$\|\mathcal{T}_{\mu,\tau_s}^{\mathrm{readout}}\| \le \varepsilon \begin{cases} \dfrac{\mu^2}{\alpha}\left[\dfrac{e^{-(\mu-\alpha)\tau_s}}{\mu-\alpha} - \dfrac{e^{-\mu\tau_s}}{\mu}\right], & \alpha \ne 0, \\ (\mu\tau_s + 1)e^{-\mu\tau_s}, & \alpha = 0. \end{cases} \tag{35}$$

For $\alpha \le 0$, the bound is at most $\varepsilon(\mu\tau_s + 1)e^{-\mu\tau_s}$. The result assumes rather than estimates ε.

Proof. Denote $\Delta\mathcal{K}^t := \mathcal{K}^t - \mathcal{K}^t_{\hat{f}}$. For functions on Ω, write $\|h\|_\infty := \sup_{\mathbf{x}\in\Omega} |h(\mathbf{x})|$ and $\|g\|_{\mathrm{Lip}} := \|g\|_\infty + \mathrm{Lip}(g)$, where $\mathrm{Lip}(g) := \sup_{x\neq y} |g(x) - g(y)|/\|x - y\|$. The induced operator norm from $\mathrm{Lip}(\Omega)$ to $\mathcal{C}(\Omega)$ is

$$\|\Delta\mathcal{K}^t\|_{\mathrm{Lip}\to\mathcal{C}} := \sup_{g\neq 0} \frac{\|\Delta\mathcal{K}^t g\|_\infty}{\|g\|_{\mathrm{Lip}}}.$$

This choice does not require Ω to be convex. It also applies to $\mathcal{C}^1$ observables on a convex neighborhood, whose Lipschitz seminorm is bounded by the supremum gradient norm on that neighborhood.

Define the flow deviation

$$\delta(t) := \sup_{\mathbf{x}\in\Omega} \|\mathbf{\Phi}^t(\mathbf{x}) - \hat{\mathbf{\Phi}}^t(\mathbf{x})\|.$$

Fix any $g \in \mathrm{Lip}(\Omega)$ and any $\mathbf{x} \in \Omega$. By Lipschitz continuity,

$$|(\Delta\mathcal{K}^t g)(\mathbf{x})| = \left|g(\mathbf{\Phi}^t(\mathbf{x})) - g(\hat{\mathbf{\Phi}}^t(\mathbf{x}))\right| \leqslant \mathrm{Lip}(g)\, \|\mathbf{\Phi}^t(\mathbf{x}) - \hat{\mathbf{\Phi}}^t(\mathbf{x})\| \leqslant \mathrm{Lip}(g)\, \delta(t).$$

Taking the supremum over $\mathbf{x} \in \Omega$ yields

$$\|\Delta\mathcal{K}^t g\|_\infty \leqslant \mathrm{Lip}(g)\, \delta(t) \leqslant \|g\|_{\mathrm{Lip}}\, \delta(t),$$

where we used $\mathrm{Lip}(g) \leqslant \|g\|_{\mathrm{Lip}}$. Dividing both sides by $\|g\|_{\mathrm{Lip}}$ and taking $\sup_{g\neq 0}$ gives the operator-norm bound

$$\|\Delta\mathcal{K}^t\|_{\mathrm{Lip}\to\mathcal{C}} \leqslant \delta(t). \tag{36}$$

Now consider the tail term $\mathcal{T}^{\mathrm{readout}}_{\mu,\tau_s}$ viewed as a linear operator from $\mathrm{Lip}(\Omega)$ to $\mathcal{C}(\Omega)$ with induced norm $\|\mathcal{T}^{\mathrm{readout}}_{\mu,\tau_s}\|_{\mathrm{Lip}\to\mathcal{C}} := \sup_{g\neq 0} \frac{\|\mathcal{T}^{\mathrm{readout}}_{\mu,\tau_s} g\|_\infty}{\|g\|_{\mathrm{Lip}}}$. For any $g \in \mathrm{Lip}(\Omega)$ with $g \neq 0$, by the triangle inequality for $\|\cdot\|_\infty$,

$$\|\mathcal{T}^{\mathrm{readout}}_{\mu,\tau_s} g\|_\infty = \left\|\mu^2 \int_{\tau_s}^\infty e^{-\mu t} \Delta\mathcal{K}^t g\, dt\right\|_\infty \leqslant \mu^2 \int_{\tau_s}^\infty e^{-\mu t} \|\Delta\mathcal{K}^t g\|_\infty\, dt,$$

hence

$$\frac{\|\mathcal{T}^{\mathrm{readout}}_{\mu,\tau_s} g\|_\infty}{\|g\|_{\mathrm{Lip}}} \leqslant \mu^2 \int_{\tau_s}^\infty e^{-\mu t} \frac{\|\Delta\mathcal{K}^t g\|_\infty}{\|g\|_{\mathrm{Lip}}}\, dt \leqslant \mu^2 \int_{\tau_s}^\infty e^{-\mu t} \|\Delta\mathcal{K}^t\|_{\mathrm{Lip}\to\mathcal{C}}\, dt.$$

Taking $\sup_{g\neq 0}$ on the left yields

$$\|\mathcal{T}^{\mathrm{readout}}_{\mu,\tau_s}\|_{\mathrm{Lip}\to\mathcal{C}} \leqslant \mu^2 \int_{\tau_s}^\infty e^{-\mu t} \|\Delta\mathcal{K}^t\|_{\mathrm{Lip}\to\mathcal{C}}\, dt. \tag{37}$$

Combining (37) with (36) gives

$$\|\mathcal{T}^{\mathrm{readout}}_{\mu,\tau_s}\|_{\mathrm{Lip}\to\mathcal{C}} \leqslant \mu^2 \int_{\tau_s}^\infty e^{-\mu t} \delta(t)\, dt. \tag{38}$$

It remains to bound $\delta(t)$ in terms of the model mismatch $\varepsilon := \|\hat{\mathbf{f}} - \mathbf{f}\|_\infty$ and the one-sided Lipschitz constant

$$\alpha := \sup_{\mathbf{x}\neq\mathbf{y}\in\Omega} \frac{\langle \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y}),\, \mathbf{x} - \mathbf{y}\rangle}{\|\mathbf{x} - \mathbf{y}\|^2}.$$

Fix any $\mathbf{x} \in \Omega$ and define the error trajectory $\mathbf{e}(t) := \mathbf{\Phi}^t(\mathbf{x}) - \hat{\mathbf{\Phi}}^t(\mathbf{x})$ and $r(t) := \|\mathbf{e}(t)\|$. Using $\dot{\mathbf{\Phi}}^t(\mathbf{x}) = \mathbf{f}(\mathbf{\Phi}^t(\mathbf{x}))$ and $\dot{\hat{\mathbf{\Phi}}}^t(\mathbf{x}) = \hat{\mathbf{f}}(\hat{\mathbf{\Phi}}^t(\mathbf{x}))$, we have

$$\dot{\mathbf{e}}(t) = \mathbf{f}(\mathbf{\Phi}^t(\mathbf{x})) - \hat{\mathbf{f}}(\hat{\mathbf{\Phi}}^t(\mathbf{x})) = \big(\mathbf{f}(\hat{\mathbf{\Phi}}^t(\mathbf{x})) - \hat{\mathbf{f}}(\hat{\mathbf{\Phi}}^t(\mathbf{x}))\big) + \big(\mathbf{f}(\mathbf{\Phi}^t(\mathbf{x})) - \mathbf{f}(\hat{\mathbf{\Phi}}^t(\mathbf{x}))\big).$$

Let $V(t) := \frac{1}{2}\|\mathbf{e}(t)\|^2 = \frac{1}{2} r(t)^2$. Then

$$\dot{V}(t) = \langle \mathbf{e}(t), \dot{\mathbf{e}}(t)\rangle = \langle \mathbf{e}(t), \mathbf{f}(\hat{\mathbf{\Phi}}^t(\mathbf{x})) - \hat{\mathbf{f}}(\hat{\mathbf{\Phi}}^t(\mathbf{x}))\rangle + \langle \mathbf{e}(t), \mathbf{f}(\mathbf{\Phi}^t(\mathbf{x})) - \mathbf{f}(\hat{\mathbf{\Phi}}^t(\mathbf{x}))\rangle.$$

By the definition of α,

$$\langle \mathbf{e}(t), \mathbf{f}(\mathbf{\Phi}^t(\mathbf{x})) - \mathbf{f}(\hat{\mathbf{\Phi}}^t(\mathbf{x}))\rangle \leqslant \alpha \|\mathbf{e}(t)\|^2 = \alpha r(t)^2,$$

and by Cauchy–Schwarz together with $\varepsilon = \|\hat{\mathbf{f}} - \mathbf{f}\|_\infty$,

$$\langle \mathbf{e}(t), \mathbf{f}(\hat{\mathbf{\Phi}}^t(\mathbf{x})) - \hat{\mathbf{f}}(\hat{\mathbf{\Phi}}^t(\mathbf{x}))\rangle \leqslant \|\mathbf{e}(t)\|\, \|\mathbf{f}(\hat{\mathbf{\Phi}}^t(\mathbf{x})) - \hat{\mathbf{f}}(\hat{\mathbf{\Phi}}^t(\mathbf{x}))\| \leqslant r(t)\,\varepsilon.$$

Therefore $\dot{V}(t) \leqslant \alpha r(t)^2 + \varepsilon r(t)$. Whenever $r(t) > 0$, using $\dot{V}(t) = r(t)\dot{r}(t)$ yields

$$\dot{r}(t) \leqslant \alpha r(t) + \varepsilon, \qquad r(0) = 0. \tag{39}$$

At zeros of r, the same inequality holds for the upper right Dini derivative, so scalar comparison applies throughout. Solving (39) gives

$$r(t) \leqslant \varepsilon\, \psi_\alpha(t), \qquad \psi_\alpha(t) := \begin{cases} \dfrac{e^{\alpha t} - 1}{\alpha}, & \alpha \neq 0, \\ t, & \alpha = 0, \end{cases}$$

and taking the supremum over $\mathbf{x} \in \Omega$ implies

$$\delta(t) \leqslant \varepsilon\, \psi_\alpha(t). \tag{40}$$

Substituting (40) into (38) yields

$$\|\mathcal{T}^{\mathrm{readout}}_{\mu,\tau_s}\|_{\mathrm{Lip}\to\mathcal{C}} \leqslant \varepsilon\, \mu^2 \int_{\tau_s}^{\infty} e^{-\mu t}\psi_\alpha(t)\, dt. \tag{41}$$

If $\alpha \neq 0$, then $\psi_\alpha(t) = \frac{e^{\alpha t}-1}{\alpha}$; for $\mu > \max\{\alpha, 0\}$ we compute

$$\int_{\tau_s}^{\infty} e^{-\mu t}\psi_\alpha(t)\, dt = \frac{1}{\alpha}\int_{\tau_s}^{\infty} \big(e^{-(\mu-\alpha)t} - e^{-\mu t}\big)\, dt = \frac{1}{\alpha}\Big[\frac{e^{-(\mu-\alpha)\tau_s}}{\mu-\alpha} - \frac{e^{-\mu\tau_s}}{\mu}\Big].$$

Plugging this into (41) gives

$$\boxed{\|\mathcal{T}^{\mathrm{readout}}_{\mu,\tau_s}\|_{\mathrm{Lip}\to\mathcal{C}} \leqslant \varepsilon\, \frac{\mu^2}{\alpha}\Big[\frac{e^{-(\mu-\alpha)\tau_s}}{\mu-\alpha} - \frac{e^{-\mu\tau_s}}{\mu}\Big].}$$

If $\alpha = 0$, then $\psi_0(t) = t$ and

$$\int_{\tau_s}^{\infty} te^{-\mu t}\, dt = e^{-\mu\tau_s}\Big(\frac{\tau_s}{\mu} + \frac{1}{\mu^2}\Big),$$

so (41) yields

$$\boxed{\|\mathcal{T}^{\mathrm{readout}}_{\mu,\tau_s}\|_{\mathrm{Lip}\to\mathcal{C}} \leqslant \varepsilon\, \mu^2\, e^{-\mu\tau_s}\Big(\frac{\tau_s}{\mu} + \frac{1}{\mu^2}\Big) = \varepsilon(\mu\tau_s + 1)e^{-\mu\tau_s}.}$$

This completes the proof. □

I THE FOH INTEGRATION AND THE CHOICE OF ω_μ

This section derives a closed-form FOH (First-Order Hold) discretization of the μ-weighted one-step-residual integral appearing in (7), and specifies a practical normalization $\omega_\mu = (A_\mu^2 + B_\mu^2 + 10^{-24}\mu^4)^{-1/2}$ to balance numerical scales across different μ.

Let t denote continuous time, and assume τ_s is an integer multiple of T_s, and define $J := \tau_s/T_s \in \mathbb{N}$ together with sampling instants $t_k := kT_s$ for $k = 0, 1, \ldots, J$. Let $\mu > 0$ be the exponential-kernel parameter. Consider the FOH interpolation $\widetilde{\mathbf{E}}(t)$ of the indexed one-step residuals $\mathbf{E}_k = \mathbf{Z}_{k+1} - \mathbf{Z}_k\widehat{\mathbf{K}}$ for $k = 0, \ldots, J-1$. Each $\mathbf{E}_k$ is associated with transition interval $[t_k, t_{k+1}]$, not a uniquely defined physical sample time. For quadrature, the index k is assigned to node t_k as a numerical convention. The implementation closes the final FOH interval with the numerical endpoint convention $\mathbf{E}_J = 0$, and defines

$$\mathbf{I}_\mu \;:=\; \int_0^{\tau_s} \mu^2 e^{-\mu t}\, \widetilde{\mathbf{E}}(t)\, dt. \tag{42}$$

By definition, the FOH interpolant varies linearly within each sampling interval. For any $t \in [t_k, t_{k+1}]$, introduce the dimensionless local parameter $\theta := (t - t_k)/T_s \in [0, 1]$ and write

$$\widetilde{\mathbf{E}}(t) = (1 - \theta)\, \mathbf{E}_k + \theta\, \mathbf{E}_{k+1}. \tag{43}$$

Substituting (43) into (42) and decomposing the integral over sub-intervals gives

$$\mathbf{I}_\mu = \sum_{k=0}^{J-1} \int_{t_k}^{t_{k+1}} \mu^2 e^{-\mu t} \Big((1-\theta)\mathbf{E}_k + \theta \mathbf{E}_{k+1} \Big)\, dt. \tag{44}$$

For each interval, apply the change of variables $s := t - t_k$, so that $s \in [0, T_s]$, $t = t_k + s$, $dt = ds$, $\theta = s/T_s$, and $e^{-\mu t} = e^{-\mu t_k} e^{-\mu s}$. Then the k-th term in (44) becomes

$$e^{-\mu t_k} \int_0^{T_s} \mu^2 e^{-\mu s} \Big(\Big(1 - \frac{s}{T_s}\Big)\mathbf{E}_k + \frac{s}{T_s}\mathbf{E}_{k+1} \Big)\, ds. \tag{45}$$

Define the two scalar endpoint coefficients

$$A_\mu \;:=\; \int_0^{T_s} \mu^2 e^{-\mu s} \Big(1 - \frac{s}{T_s}\Big)\, ds, \qquad B_\mu \;:=\; \int_0^{T_s} \mu^2 e^{-\mu s} \Big(\frac{s}{T_s}\Big)\, ds, \tag{46}$$

so that (45) equals $e^{-\mu t_k} \big(A_\mu \mathbf{E}_k + B_\mu \mathbf{E}_{k+1} \big)$.

To obtain a closed form, we use

$$\int_0^{T_s} e^{-\mu s}\, ds = \frac{1 - e^{-\mu T_s}}{\mu}, \qquad \int_0^{T_s} s e^{-\mu s}\, ds = \frac{1 - e^{-\mu T_s}(1 + \mu T_s)}{\mu^2},$$

where the second identity follows from the integration-by-parts antiderivative $\int s e^{-\mu s} ds = -\frac{s}{\mu} e^{-\mu s} - \frac{1}{\mu^2} e^{-\mu s} + C$ evaluated on $[0, T_s]$. Substituting these two integrals into (46) and simplifying yields the FOH coefficients

$$\boxed{A_\mu = \mu\big(1 - e^{-\mu T_s}\big) - \frac{1 - e^{-\mu T_s}(1 + \mu T_s)}{T_s}, \qquad B_\mu = \frac{1 - e^{-\mu T_s}(1 + \mu T_s)}{T_s}.}$$

Summing (45) over $k = 0, \ldots, J-1$ and using $t_k = kT_s$ gives the FOH closed-form discretization of (42):

$$\boxed{I_\mu = \sum_{k=0}^{J-1} e^{-\mu k T_s} \Big(A_\mu\, \mathbf{E}_k + B_\mu\, \mathbf{E}_{k+1} \Big), \qquad J = \tau_s / T_s.} \tag{47}$$

With $\rho_\mu = e^{-\mu T_s}$ and $S_\mu = \sum_{k=0}^{J-1} \rho_\mu^k E_k$, the same implemented moment satisfies the exact identity

$$I_\mu = (A_\mu + B_\mu/\rho_\mu) S_\mu - (B_\mu/\rho_\mu) E_0, \tag{48}$$

which makes precise how the FOH loss relates to the discounted sum in Eq. (11).

Since A_μ and B_μ depend on (μ, T_s) and may vary significantly in magnitude across different μ, directly aggregating multiple μ terms can introduce an unintended scale bias. To balance the endpoint-weight scale under FOH, we choose

$$\boxed{\omega_\mu \;:=\; \frac{1}{\sqrt{A_\mu^2 + B_\mu^2 + 10^{-24}\mu^4}}.}$$

The 10^{-24} term is the implemented floating-point stabilizer; absent it, the endpoint-weight norm is exactly one. The normalized weights therefore mitigate μ-dependent scale disparities in the discretized residual integral (47) when combining multiple μ values in the loss.

The auxiliary value $\mathbf{E}_J = 0$ closes the final FOH interval; every residual except E_0 receives its left-endpoint contribution and the right-endpoint contribution from the preceding interval. A zero-order hold would define different exact cell weights and hence a different objective requiring its own validation-selected optimization settings; no fixed-hyperparameter comparison is used.

For clarity about cancellation, collect the normalized FOH coefficients into $W \in \mathbb{R}^{|\mathcal{M}|\times J}$, where

$$W_{\ell 0} = \omega_{\mu_\ell} A_{\mu_\ell}, \qquad W_{\ell k} = \omega_{\mu_\ell}\left(\rho_{\mu_\ell}^{k} A_{\mu_\ell} + \rho_{\mu_\ell}^{k-1} B_{\mu_\ell}\right), \quad 1 \le k < J.$$

Equation (7) therefore penalizes temporal moments $W\mathbf{E}$, not the pointwise energy $\sum_k \|\mathbf{E}_k\|^2$. Oppositely signed defects can cancel within one moment. Full column rank of W excludes exact cancellation across the complete family, while its smallest singular value controls the bound on residual energy in Eq. (8). This does not by itself guarantee multi-step state prediction. Appendix D reports the singular spectrum and conditioning of the implemented μ grid.

Empirical propagation bound. For a fitted trajectory, repeated substitution of $\mathbf{Z}_{k+1} = \mathbf{Z}_k\widehat{\mathbf{K}} + \mathbf{E}_k$ gives

$$\mathbf{Z}_j - \mathbf{Z}_0\widehat{\mathbf{K}}^j = \sum_{k=0}^{j-1} \mathbf{E}_k \widehat{\mathbf{K}}^{j-1-k}.$$

Cauchy–Schwarz over the time blocks, followed by Eq. (8), yields

$$\|\mathbf{Z}_j - \mathbf{Z}_0\widehat{\mathbf{K}}^j\|_F \le \left(\sum_{s=0}^{j-1} \|\widehat{\mathbf{K}}\|_2^{2s}\right)^{1/2} \|\mathbf{E}\|_F \le \frac{\gamma_j(\widehat{\mathbf{K}})}{\sigma_{\min}(W)} \|W\mathbf{E}\|_F.$$

The identical inequality holds after stacking trajectories. Because observed states are fixed coordinates of $\mathbf{Z}$, their fitted-trajectory discrepancies are no larger than the full lifted discrepancy. This is a deterministic empirical bound; it neither controls validation trajectories nor connects recursive linear prediction directly to the nonlinear vector-field rollout used for equation evaluation.

J Selection of Analytic Basis Functions

This section specifies the analytic functions used to construct the nested equation and closure spaces. We consider (i) a monomial basis in d dimensions and (ii) a trigonometric basis with single-variable harmonics and pairwise-difference harmonics. Throughout, let $\mathbf{x} = [x_1, \ldots, x_d]^\top \in \mathbb{R}^d$ denote the system state, and let (1) denote an optional constant basis element, included only if desired.

J.1 Monomial basis in d dimensions

Let $p \in \mathbb{N}$ denote the maximum polynomial degree. We use the standard multi-index notation. For a multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ (dimensionless entries), define its total degree as $|\boldsymbol{\alpha}| := \sum_{i=1}^d \alpha_i$ and the monomial

$$\mathbf{x}^{\boldsymbol{\alpha}} := \prod_{i=1}^{d} x_i^{\alpha_i}.$$

The d-dimensional monomial dictionary up to degree p is then

$$\mathcal{B}_{\text{poly}}^{(p)} := \left\{ \mathbf{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{N}_0^d,\ |\boldsymbol{\alpha}| \leqslant p \right\}. \tag{49}$$

If the constant term is excluded, we simply remove (1) from (49). Accordingly, the cardinality is $\binom{d+p}{p}$ with the constant and $\binom{d+p}{p} - 1$ without it. The default polynomial and scaling experiments omit the constant (hence 19 and 83 terms for 3D and 6D cubics); Reaction explicitly uses the constant-inclusive dictionary.

The implementation first lists the coordinate states and then increasing total degree; within each degree it enumerates nondecreasing variable-index tuples lexicographically, as produced by `combinations_with_replacement(range(d), degree)`. If present, the constant follows the coordinates and precedes degree-two terms. This order defines nested, named polynomial sets. For selected $q_{\text{tail}} =$ `tail_n`, the preceding functions define the equation space U and the final q_{tail} functions provide closure references V; coordinated permutations within either set leave the estimator unchanged. In the degree-aligned case this construction has the transparent form

$$\mathcal{F}_{\text{eq}} = \text{span}\{x^\alpha : |\alpha| \le p_{\text{eq}}\} \subset \mathcal{F}_{K,\theta} = \text{span}[U, V + r_\theta],$$

where higher-degree observables help close the finite Koopman approximation but are not reported as governing terms. The implementation also permits a predeclared named boundary within one degree; function membership, rather than degree alone, defines the scientific class.

J.2 Trigonometric basis with harmonics and pairwise differences

Let $K \in \mathbb{N}$ denote the maximum harmonic index (dimensionless), and let $k \in \{1, 2, \ldots, K\}$ be a harmonic order. The trigonometric dictionary is constructed from (i) single-variable harmonics $\sin(kx_i), \cos(kx_i)$ for each state component x_i, and (ii) pairwise-difference harmonics $\sin(k(x_i - x_j)), \cos(k(x_i - x_j))$ for each pair (i, j) with $1 \leqslant i < j \leqslant d$:

$$\begin{aligned}\mathcal{B}^{(K)}_{\mathrm{trig}} \;:=\; \{(1)\} \;\cup\; \big\{ \sin(kx_i),\; \cos(kx_i) \;:\; k = 1, \ldots, K,\; i = 1, \ldots, d \big\} \\ \cup\; \big\{ \sin(k(x_i - x_j)),\; \cos(k(x_i - x_j)) \;:\; k = 1, \ldots, K,\; 1 \leqslant i < j \leqslant d \big\}.\end{aligned} \tag{50}$$

As in the polynomial case, the constant element (1) is optional. The implemented order is: coordinate states; then, for each $k = 1, \ldots, K$, $\sin(kx_i), \cos(kx_i)$ for increasing i; then $\sin(k(x_i - x_j)), \cos(k(x_i - x_j))$ for lexicographically ordered $i < j$. This ordering defines nested, named harmonic sets; the final q functions provide V and the preceding functions define U. Membership in the equation or closure set, rather than a stored column number, determines the estimator: coordinated permutations within either set only relabel the same fit and readout. Only U remains candidate in the reported equation; $V + r_\theta$ is auxiliary Koopman closure. A scientifically admitted higher harmonic must therefore be placed in U before fitting. Selecting q changes the declared equation class but not the full analytic initialization, which is why nested selection uses only short validation and may abstain when adequacy is doubtful.

K Comparison with Other Methods

We compare against SINDy (Brunton et al., 2016), E-SINDy (Fasel et al., 2022), PySR (Cranmer, 2023), and NeSymReS (Biggio et al., 2021) under the dense-grid protocol (30 observations for Van der Pol and 31 otherwise). These methods fit an explicit right-hand side from pointwise derivative targets, KLM reads from the generator, WSINDy uses weak-form regression, RTM estimates through a resolvent, and ReTRN trains a residual lift with ordered moments before principal-log readout. Seven-point PySR results are reported in Appendix A.1.

Learned-dictionary EDMD alternates closed-form operator fitting and neural feature optimization and may retain fixed observables (Li et al., 2017); Koopman autoencoders learn latent coordinates for reconstruction or prediction (Takeishi et al., 2017; Lusch et al., 2018; Azencot et al., 2020). These formulations optimize a predictive representation but do not specify a complete equation space, an admissible reparameterization group, and a head-only readout as one estimator. A fixed analytic block protects its own rows under tail-only basis changes, yet learned coordinates can still absorb equation directions, and post-hoc projection defines a second estimator. ReTRN instead embeds $\mathcal{F}_{\mathrm{eq}}$ pointwise inside a richer closure space, residualizes only designated auxiliaries, and trains every operator refit in the equation chart used for readout. Distributed and appended controls test this distinction under a common equation-space projection and show uniformly more accurate ReTRN coefficients; WSINDy supplies the prescribed-library weak-form comparison.

Table 24 reports the dense-grid results. Coefficient NRMSE is omitted for PySR and NeSymReS because their expressions may leave the prescribed library. The symbolic searches use a 72 h compute budget; the validation-selected PySR expressions for Duffing and Lorenz yield rollout MSE .00230 and 52.9, respectively. Deterministic methods are evaluated once on the fixed dataset, whereas ReTRN is summarized over five optimization seeds. Matched-subsample results are given in Tables 1 and 7.

L Data preparation

The fixed six-system simulation pool is generated with MATLAB's `ode45` solver (explicit Runge–Kutta (4,5)). For each trajectory, we integrate over a finite horizon τ_s and sample uniformly at interval T_s, yielding snapshots $\{\mathbf{x}_k\}_{k=0}^{J}$ at $t_k = kT_s$, where $J = \tau_s/T_s \in \mathbb{N}$. The independent-replicate study instead uses full-precision RK4 simulation, as specified in Appendix A.4; the seven-observation subsampling and Van der Pol endpoint completion are specified in Appendix A.1.

To generate diverse trajectories, we draw random initial conditions $\mathbf{x}_0 = \mathbf{x}(0)$ from a bounded initialization set $\mathcal{X}_0 \subset \mathbb{R}^d$. In practice, we use an axis-aligned box $\mathcal{X}_0 = [x_1^{\min}, x_1^{\max}] \times \cdots \times$

Method	Van der Pol	Duffing	Lorenz	Competition (sheep-rabbits)	Autocatalytic reaction	Magnet interaction
WSINDy	$\mathbf{6.217 \times 10^{-6}}$ $\mathbf{7.037 \times 10^{-10}}$	9.959×10^{-6} 4.510×10^{-8}	4.599×10^{-6} 3.253×10^{-5}	1.063×10^{-4} 1.120×10^{-7}	6.383×10^{-4} 7.727×10^{-7}	5.257×10^{-3} 7.399×10^{-7}
KLM	3.148×10^{-4} 2.921×10^{-7}	4.697×10^{-6} 4.247×10^{-8}	6.156×10^{-6} 9.371×10^{-5}	8.142×10^{-5} 2.751×10^{-7}	5.216×10^{-4} 9.782×10^{-7}	3.556×10^{-3} 2.947×10^{-6}
RTM	1.164×10^{-3} 1.197×10^{-6}	6.837×10^{-4} 3.867×10^{-5}	4.938×10^{-4} 9.237×10^{-1}	2.069×10^{-3} 4.943×10^{-5}	2.412×10^{-3} 4.744×10^{-5}	2.839×10^{-2} 2.467×10^{-5}
SINDy	4.473×10^{-3} 2.635×10^{-5}	2.039×10^{-3} 3.872×10^{-3}	3.367×10^{-3} 1.586×10^{1}	1.998×10^{-1} 4.304×10^{-1}	7.314×10^{-3} 3.087×10^{-5}	1.952×10^{-2} 4.755×10^{-6}
E-SINDy	2.648×10^{-2} 2.572×10^{-5}	2.039×10^{-3} 1.354×10^{-2}	3.367×10^{-3} 1.028×10^{-1}	6.738×10^{-1} 5.548×10^{-4}	2.791 9.631×10^{-6}	1.437 4.462×10^{-5}
PySR	– 3.153×10^{-7}	– 2.30×10^{-3}	– 5.29×10^{1}	– 4.675×10^{-3}	– 3.050×10^{-7}	– $\mathbf{1.637 \times 10^{-10}}$
NeSymReS	– 1.104×10^{-5}	– 6.067	– 2.445×10^{2}	– 4.514×10^{-1}	– 1.974×10^{-3}	– 1.964×10^{-3}
ReTRN	$(2.349 \pm 0.750) \times 10^{-4}$ $(1.905 \pm 0.501) \times 10^{-7}$	$(\mathbf{2.664 \pm 0.354}) \times \mathbf{10^{-6}}$ $(\mathbf{4.217 \pm 0.008}) \times \mathbf{10^{-8}}$	$(\mathbf{3.025 \pm 1.005}) \times \mathbf{10^{-6}}$ $(\mathbf{1.924 \pm 0.488}) \times \mathbf{10^{-5}}$	$(\mathbf{4.546 \pm 0.788}) \times \mathbf{10^{-5}}$ $(\mathbf{2.529 \pm 1.179}) \times \mathbf{10^{-8}}$	$(\mathbf{4.309 \pm 0.298}) \times \mathbf{10^{-4}}$ $(\mathbf{9.153 \pm 2.502}) \times \mathbf{10^{-8}}$	$(\mathbf{4.087 \pm 1.676}) \times \mathbf{10^{-4}}$ $(1.726 \pm 0.545) \times 10^{-8}$

Table 24: **Comparison on the dense-grid development benchmark.** Van der Pol has 30 observations and the remaining systems have 31. Each cell reports coefficient NRMSE (top) and long-horizon rollout MSE (bottom); coefficient NRMSE is omitted when a symbolic expression lies outside the prescribed library. ReTRN reports mean ± SD over five optimization seeds, while deterministic methods are single fits on the same fixed dataset. Search spaces and compute are method-specific; bold marks only the numerical minimum and does not denote a compute-matched ranking.

$[x_d^{\min}, x_d^{\max}]$ and sample $x_i(0) \sim \mathrm{Unif}(x_i^{\min}, x_i^{\max})$ independently for $i = 1, \ldots, d$. The bounds are chosen per system to avoid numerical instabilities while ensuring sufficient excitation; the exact ranges used for each benchmark are reported together with the experiment settings.

For numerical accuracy, we set `RelTol` and `AbsTol` to 10^{-12}, and optionally limit the solver's internal step size via `MaxStep` to prevent overly large steps compared to T_s. We generate N_{traj} independent trajectories using independent draws of $\mathbf{x}_0$, and split trajectories into training and validation sets to avoid temporal leakage. The test set is generated independently, and its length is approximately ten times that of the training and validation sets.

For the real-world experiment, the data can be obtained directly from Kramer et al. (2021). We use its most strongly nonlinear condition, which contains four measured heave-decay trajectories. The first three trajectories are segmented for training, while the fourth is reserved intact as the long test. To construct the state $\mathbf{x} = [x_3, \dot{x}_3]^{\mathsf{T}}$, we estimate $\dot{x}_3$ from the measured displacement x_3 using a five-point central finite-difference stencil on the training segments. This derivative estimate serves only as an observed state coordinate, never as a pointwise vector-field target. On the test record, the stencil supplies $\dot{x}_3(t_0)$ for rollout initialization; model integration thereafter uses no estimated derivative. Thus this is a common five-sample state-initialization protocol, not a zero-lookahead displacement-only forecast. The separately reported derived-state evaluation applies the same stencil throughout its centered valid region only to construct the reference velocity against which the frozen rollout is scored.

M Parameters and Settings

We summarize the hyperparameters used by each method and explain how we choose their values in our experiments. Unless otherwise stated, we select hyperparameters based on validation performance and then keep the chosen values fixed for all reported runs. The test set is used only for the final evaluation.

The dense-grid ReTRN configuration includes the analytic dictionary order, tail dimension, residual-block count, logarithmically spaced μ grid, learning rate, and training budget. For a d-dimensional polynomial state with maximum degree p, the library has $\binom{d+p}{p} - 1$ monomials when the constant is omitted and $\binom{d+p}{p}$ when it is included; Reaction uses the latter, whereas the other polynomial benchmarks and scaling study use the former. The trigonometric order instead denotes the maximum harmonic. Table 25 reports the dense-grid settings and data domains; the independently selected seven-observation configurations appear in Table 6. We use Optuna (Akiba et al., 2019) with short-validation rollout MSE as the selection objective.

Dynamical system	$polyorder$	$tail_n$	res_blocks	mu_min	mu_max	mu_n	lr	$epochs$	N_{traj}	τ_s	T_s	$\mathcal{X}_0$
a) Van der Pol	4	1	10	10	10^4	30	2.308×10^{-4}	300	1000	0.29	0.01	$[-5,5]^2$
b) Duffing	4	1	20	10	10^4	30	2.068×10^{-4}	300	1000	0.3	0.01	$[-5,5]^2$
c) Lorenz	3	6	12	10	10^4	30	1.933×10^{-3}	300	1000	0.3	0.01	$[-20,20] \times [-30,30] \times [0,50]$
d) Competition	4	7	14	10	10^4	30	1.702×10^{-4}	300	1000	0.3	0.01	$[2,10]^2$
e) Reaction	4	3	8	10	10^4	30	5.401×10^{-4}	300	1000	0.3	0.01	$[0,2]^2$
f) Magnet	3	10	4	10	10^4	30	2.802×10^{-3}	300	1000	0.3	0.01	$[-1.5,1.5]^2$
g) Heave decay	4	2	4	10	10^4	30	1.000×10^{-4}	300	906	0.018	0.002	–

Table 25: ReTRN settings for the dense-grid benchmark and heave-decay case study. For heave decay, $N_{\mathrm{traj}} = 906$ denotes the nonoverlapping ten-sample segments obtained from the first three measured records.

Within the dense-grid comparison of Table 24, every method receives the same fixed state observations for a system, and the prescribed-library estimators share the same basis; PySR and NeSymReS instead use their native expression spaces. We therefore state the dense-grid dataset and polynomial order once. KLM has no operator-estimation hyperparameter in this fixed-dataset setting. In the matched seven-observation experiments, however, its sparse-readout threshold is selected on the common short validation data exactly as specified in Appendix A.1; the distinct data pools disclosed there are not covered by this dense-grid statement.

For RTM, it computes the resolvent using a small μ, and then uses the first resolvent identity to estimate the Yosida approximation at a large λ as the generator. Therefore, it involves two parameters, μ and λ. We set λ to the fixed value recommended in Meng et al. (2024), and choose μ by grid search over a predefined candidate set, stopping once the validation error becomes essentially insensitive to further changes in μ. The resulting settings are reported in Table 26.

Hyperparameter	Van der Pol	Duffing	Lorenz	Competition	Reaction	Magnet
μ	35	35	30	50	35	30
λ	10^8	10^8	10^8	10^8	10^8	10^8

Table 26: Detailed setup of RTM for dynamical systems

Both WSINDy-SR3 and SINDy-SR3 use the same SR3 sparse optimizer and therefore share the same three SR3 hyperparameters, namely $threshold$, nu, and max_iter. In our implementation, threshold controls the sparsity level of the identified model. Concretely, it is converted into an L0-type regularization weight through `SR3.calculate_l0_weight`, so the resulting sparsity prior is an L0-style penalty that promotes hard sparsity in the coefficient vector rather than an L1 penalty. The parameter nu is the SR3 relaxation coefficient, which governs the strength of the variable-splitting relaxation and affects convergence behavior. The parameter max_iter specifies the maximum number of SR3 iterations used in the sparse regression stage. In addition, WSINDy introduces one extra method-specific hyperparameter K, which controls the number of weak-form test functions in the weak library and only appears in the WSINDy construction, not in the finite-difference SINDy baseline. Both methods are tuned by `Optuna`. The settings are shown in Table 27 and 28.

For PySR, the operator class, parsimony, maximum expression size, populations, and search budget are selected on short validation; the seven-point settings are then locked across its separate 20-subsample pool. NeSymReS uses the largest pretrained model released by its authors.

Hyperparameter	Van der Pol	Duffing	Lorenz	Competition	Reaction	Magnet
$threshold$	3.735×10^{-4}	2.372×10^{-3}	7.226×10^{-2}	9.496×10^{-3}	8.869×10^{-3}	6.036×10^{-4}
nu	5.275×10^{-4}	3.615×10^{-6}	1.005×10^{-6}	1.209×10^{-6}	1.747×10^{-6}	1.006×10^{-6}
max_iter	300	300	300	300	300	300
K	900	1000	800	100	700	200

Table 27: Detailed setup of WSINDy for dynamical systems

Hyperparameter	Van der Pol	Duffing	Lorenz	Competition	Reaction	Magnet
$threshold$	2.005×10^{-2}	6.719×10^{-2}	3.259×10^{-4}	7.912×10^{-5}	9.535×10^{-2}	2.395×10^{-2}
nu	1.016×10^{-6}	2.259×10^{-6}	1.615×10^{-5}	1.083×10^{-6}	8.713×10^{-3}	1.062×10^{-2}
max_iter	300	300	300	300	300	300

Table 28: Detailed setup of SINDy for dynamical systems

For methods whose configurations are selected in this study, hyperparameter selection uses validation rollout MSE and does not access the long test. ReTRN selects tail size, block count, learning rate, training budget, batch size, and temporal-weight grid; SINDy selects its threshold and regularization; WSINDy additionally selects K. The dense-grid development study uses system-specific Optuna searches, whereas the matched seven-observation experiments use five fixed tuning subsamples and disclose their inferential role separately. PySR uses a 72 h search budget per system; NeSymReS uses the released pretrained model rather than a study-specific search. The dense-grid Torch experiments use PyTorch 2.4.0 and CUDA 12.4 on one NVIDIA H100; the seven-observation and scaling experiments use an RTX 3500 Ada GPU, with final-fit times reported alongside the results.

N Detailed Results of Diverse Dynamics

This section provides system-by-system descriptions and the explicit governing equations identified by ReTRN for the six dynamical systems in the main text. For each system, we first describe the dynamical meaning of the state and parameters, summarize qualitative properties (equilibria, attractors, stability/chaos), then present the ground-truth dynamics and the identified equations.

N.1 Van der Pol oscillator

System description. The Van der Pol oscillator is a second-order nonlinear oscillator with state-dependent damping. We use $\mathbf{x} = [x_1, x_2]^\top$, where x_1 is a position-like state and x_2 is a velocity-like state. The scalar parameter μ controls the strength of the nonlinearity and the shape of the limit cycle.

Dynamics properties. For $\mu > 0$ (here $\mu = 1$), the origin is an unstable equilibrium, and the system admits a unique stable limit cycle attracting almost all trajectories. The combined term $\mu(1 - x_1^2)x_2$ gives *state-dependent damping*: negative damping for $|x_1| < 1$ and positive damping for $|x_1| > 1$, which self-regulates the oscillation amplitude. As μ increases, the dynamics transition from near-sinusoidal motion to *relaxation oscillations* with fast–slow time scales; with $\mu = 1$ the oscillation is nonlinear but not strongly stiff.

Ground-truth dynamics. With $\mu = 1$,

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -x_1 + \mu x_2 - \mu x_1^2 x_2 \;=\; -x_1 + x_2 - x_1^2 x_2. \end{aligned}$$

Identified dynamics (ReTRN).

$$\dot{x}_1 = 1.00002\,x_2 + 2.08757 \times 10^{-5}\,x_1x_2 + 1.72936 \times 10^{-5}\,x_1 - 2.16669 \times 10^{-6}\,x_1^2,$$
$$\dot{x}_2 = -0.999801\,x_1^2x_2 - 0.999706\,x_1 + 0.999346\,x_2 + 0.000307531\,x_1x_2.$$

N.2 DUFFING OSCILLATOR

System description. The Duffing oscillator is a nonlinear second-order system with cubic stiffness and linear damping. We use $\mathbf{x} = [x_1, x_2]^\top$, where x_1 is displacement-like and x_2 is velocity-like. Parameters δ (damping), α (linear stiffness), and β (cubic stiffness) are dimensionless.

Dynamics properties. With $\alpha = -1$ and $\beta = 3$, the restoring force corresponds to a *double-well potential* $V(x_1) = -\frac{1}{2}x_1^2 + \frac{3}{4}x_1^4$. The damping $\delta > 0$ (here $\delta = 0.5$) renders the system *dissipative*, so trajectories relax toward equilibria. Equilibria satisfy $-\alpha x_1 - \beta x_1^3 = 0$, i.e., $x_1^\star \in \{0, \pm 1/\sqrt{3}\}$ with $x_2^\star = 0$. The origin is an unstable saddle (local maximum of V), while $x_1^\star = \pm 1/\sqrt{3}$ correspond to stable minima; hence the system is typically *bistable* with two attraction basins separated by a separatrix.

Ground-truth dynamics. With $\delta = 0.5$, $\alpha = -1$, $\beta = 3$,

$$\dot{x}_1 = x_2,$$
$$\dot{x}_2 = -\delta x_2 - \alpha x_1 - \beta x_1^3 \;=\; -0.5\,x_2 + x_1 - 3\,x_1^3.$$

Identified dynamics (ReTRN).

$$\dot{x}_1 = 0.999999\,x_2 - 4.53847 \times 10^{-6}\,x_1 - 3.70968 \times 10^{-6}\,x_1^2 - 1.1401 \times 10^{-6}\,x_1x_2,$$
$$\dot{x}_2 = -3\,x_1^3 + 0.999971\,x_1 - 0.499997\,x_2 - 1.00104 \times 10^{-5}\,x_1^2.$$

N.3 LORENZ SYSTEM

System description. The Lorenz system is a three-dimensional nonlinear flow exhibiting chaotic behavior for a classical parameter regime. We use $\mathbf{x} = [x_1, x_2, x_3]^\top$ and parameters (σ, ρ, β).

Dynamics properties. For $(\sigma, \rho, \beta) = (10, 28, 8/3)$, the Lorenz system is a canonical *dissipative chaotic* system with a strange attractor and strong sensitivity to initial conditions. It has three equilibria: the origin and the two symmetric points $(\pm\sqrt{\beta(\rho-1)}, \pm\sqrt{\beta(\rho-1)}, \rho-1)$; at $\rho = 28$ these nontrivial equilibria are unstable and trajectories are attracted to the butterfly-shaped chaotic set. The flow is bounded and volume-contracting (negative divergence), while generic chaotic trajectories switch aperiodically between two lobes.

Ground-truth dynamics. With $\sigma = 10$, $\rho = 28$, $\beta = 8/3$,

$$\dot{x}_1 = \sigma(x_2 - x_1) \;=\; 10(x_2 - x_1),$$
$$\dot{x}_2 = x_1(\rho - x_3) - x_2 \;=\; 28x_1 - x_1x_3 - x_2,$$
$$\dot{x}_3 = x_1x_2 - \beta x_3 \;=\; x_1x_2 - \tfrac{8}{3}x_3.$$

Identified dynamics (ReTRN).

$$\dot{x}_1 = 10\,x_2 - 9.99992\,x_1 + 1.95549 \times 10^{-5}\,x_3 - 3.63754 \times 10^{-6}\,x_1x_3,$$
$$\dot{x}_2 = 28\,x_1 - 1\,x_2 - 0.999998\,x_1x_3 - 7.75133 \times 10^{-6}\,x_3,$$
$$\dot{x}_3 = -2.66668\,x_3 + 1.00001\,x_1x_2 - 8.30149 \times 10^{-5}\,x_1 - 1.30998 \times 10^{-5}\,x_2.$$

N.4 LOTKA–VOLTERRA COMPETITION MODEL (WITH SHEEP AND RABBITS)

System description. This two-species competition model captures logistic self-limitation and inter-species inhibition. We use $\mathbf{x} = [x_1, x_2]^\top$, where x_1 and x_2 are population densities. Parameters c_0, c_1, c_2 encode intrinsic growth and competition strengths.

Dynamics properties. The vector field preserves the positive quadrant under standard conditions and typically exhibits convergence to equilibria (no sustained oscillations for the parameter regime here). Equilibria include extinction states and a positive coexistence equilibrium obtained by solving

$$c_0 - c_1 x_2 - x_1 = 0, \qquad c_2 - x_1 - x_2 = 0.$$

With $c_0 = 3, c_1 = 2, c_2 = 2$, this yields $(x_1^\star, x_2^\star) = (1, 1)$, an unstable coexistence equilibrium with eigenvalues $-1 \pm \sqrt{2}$. The boundary equilibria $(3, 0)$ and $(0, 2)$ are locally stable, giving two competitive-exclusion outcomes. The quadratic terms $-x_1^2$ and $-x_2^2$ encode intra-species competition (logistic saturation), while the bilinear terms $-2x_1x_2$ and $-x_1x_2$ encode asymmetric cross-inhibition.

Ground-truth dynamics. With $c_0 = 3$, $c_1 = 2$, $c_2 = 2$,

$$\begin{aligned}\dot{x}_1 &= x_1(c_0 - c_1x_2 - x_1) \;=\; 3x_1 - 2x_1x_2 - x_1^2,\\ \dot{x}_2 &= x_2(c_2 - x_1 - x_2) \;=\; 2x_2 - x_1x_2 - x_2^2.\end{aligned}$$

Identified dynamics (ReTRN).

$$\begin{aligned}\dot{x}_1 &= 3\,x_1 - 2\,x_1x_2 - 0.999937\,x_1^2 - 0.000101429\,x_2,\\ \dot{x}_2 &= 2.00015\,x_2 - 1.00013\,x_2^2 - 0.999821\,x_1x_2 - 0.000194963\,x_1.\end{aligned}$$

N.5 Isothermal autocatalytic reaction model

System description. This two-state kinetic model combines a linear feed/removal mechanism and an autocatalytic nonlinear pathway. We use $\mathbf{x} = [x_1, x_2]^\top$, where x_1 and x_2 are concentrations in dimensionless units. Parameters c_0 and c_1 act as feed/removal-like rates.

Dynamics properties. The nonlinear term $x_1x_2^2$ introduces autocatalysis: x_2 can grow rapidly when both x_1 and x_2 are present, followed by relaxation due to the linear decay $-c_1x_2$. The system admits at least one evident equilibrium $(x_1^\star, x_2^\star) = (1, 0)$. Additional positive equilibria satisfy $\dot{x}_2 = 0 \Rightarrow x_1x_2 = c_1$ and $\dot{x}_1 = 0 \Rightarrow c_0(1 - x_1) = x_1x_2^2$, which can yield multiple steady states depending on the admissible state range. Dynamically, this model often exhibits sharp transients and multi-time-scale relaxation to a steady state for the parameter regime considered.

Ground-truth dynamics. With $c_0 = 0.5$, $c_1 = 0.02$,

$$\begin{aligned}\dot{x}_1 &= c_0(1 - x_1) - x_1x_2^2 \;=\; 0.5 - 0.5\,x_1 - x_1x_2^2,\\ \dot{x}_2 &= -c_1x_2 + x_1x_2^2 \;=\; -0.02\,x_2 + x_1x_2^2.\end{aligned}$$

Identified dynamics (ReTRN).

$$\begin{aligned}\dot{x}_1 &= -1.00025\,x_1x_2^2 - 0.50031\,x_1 + 0.500071 + 0.000505095\,x_1x_2,\\ \dot{x}_2 &= 1.00026\,x_1x_2^2 - 0.0200679\,x_2 - 0.000429862\,x_1x_2 + 0.000182028\,x_1.\end{aligned}$$

N.6 Magnet interaction model

System description. This model describes interactions between two coupled angular states. We use $\mathbf{x} = [x_1, x_2]^\top$, where x_1 and x_2 are angles, and all sine arguments follow the radian convention. The coupling strength c_0 is dimensionless.

Dynamics properties. The terms $-\sin(x_1)$ and $-\sin(x_2)$ act as individual restoring torques that pull each angle toward stable configurations (e.g., neighborhoods of multiples of 2π). The coupling terms $\pm c_0 \sin(x_1 - x_2)$ depend on the relative phase and oppose alignment near the origin: a small positive phase difference increases the first component and decreases the second. Individual restoring terms dominate for $c_0 = 0.33$. The linearized common and difference modes have eigenvalues -1 and $2c_0 - 1 = -0.34$, respectively, so the origin remains locally asymptotically stable.

Ground-truth dynamics. With $c_0 = 0.33$,

$$\begin{aligned}\dot{x}_1 &= c_0 \sin(x_1 - x_2) - \sin(x_1),\\ \dot{x}_2 &= -c_0 \sin(x_1 - x_2) - \sin(x_2).\end{aligned}$$

Identified dynamics (ReTRN).

$$\dot{x}_1 = -0.999171\,\sin(x_1) + 0.329836\,\sin\big(x_1 - x_2\big) - 0.000417684\,x_1 - 0.000231656\,\sin(2x_1),$$
$$\dot{x}_2 = -0.999957\,\sin(x_2) - 0.330052\,\sin\big(x_1 - x_2\big) - 0.000220674\,\sin(x_1) + 0.000147709\,x_1.$$

N.7 Real-World Heave Decay Tests

System description. We evaluate ReTRN on real-world heave decay measurements of a semi-submerged spherical model interacting with water Kramer et al. (2021). The measured signal is the vertical displacement from equilibrium, denoted by $x_3(t)$. We adopt the standard heave state choice $\mathbf{x} = [x_1, x_2]^\top$ with $x_1 = x_3$ (displacement) and $x_2 = \dot{x}_3$ (velocity); the unobserved velocity is constructed by a five-point finite difference, not used as a pointwise vector-field regression target. After the release instant, each of the first three records contributes 3,020 samples, partitioned chronologically into 302 nonoverlapping ten-sample segments. The fixed seed-3407 split assigns 181 of the resulting 906 segments to within-record validation and the remainder to fitting; the fourth record is never segmented for fitting or selection and remains the long-rollout test.

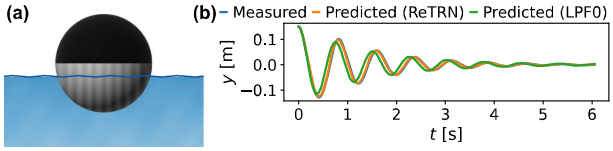


Figure 4: **Measured heave decay.** Experimental setup and held-out rollout.

Dynamics properties. Heave decay is a free-response experiment initialized by a prescribed vertical offset and instantaneous release. The resulting dynamics combine hydrostatic restoring, radiation effects that are nonlocal in time in potential-flow theory, and viscous dissipation. While Kramer et al. (2021) provides potential-flow models (LPF0–LPF4) as theoretical references, real measurements can exhibit amplitude-dependent restoring and damping. We therefore test whether a phenomenological explicit ODE recovers the LPF0 linear time scale while providing a compact closure for the observed nonlinear response.

Reference linear heave backbone (LPF0). Theoretical LPF0 in Kramer et al. (2021) models the heave decay as a linear mass–spring–damper system

$$(m + A_{33}(\omega_0))\ddot{x}_3 + B_{33}(\omega_0)\dot{x}_3 + C_{33}x_3 = 0,$$

where (at the selected ω_0) $m = 7.056$, $A_{33}(\omega_0) = 2.97$, $B_{33}(\omega_0) = 13.95$, and $C_{33} = 692.89$. Under $x_1 = x_3$ and $x_2 = \dot{x}_3$, this becomes

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = -k_{\mathrm{LPF0}}x_1 - c_{\mathrm{LPF0}}x_2,$$

with

$$k_{\mathrm{LPF0}} = \frac{C_{33}}{m + A_{33}(\omega_0)} = \frac{692.89}{7.056 + 2.97} = 69.1093,$$
$$c_{\mathrm{LPF0}} = \frac{B_{33}(\omega_0)}{m + A_{33}(\omega_0)} = \frac{13.95}{7.056 + 2.97} = 1.39138.$$

Identified dynamics (ReTRN). The degree-four nonconstant dictionary contains 14 monomials, with 12 prescribed analytic coordinates and two neural-tail coordinates that are deleted at readout. We compute each analytic term's trajectory-aggregated absolute rollout contribution on held-out 20% splits from the first three measured records. The kinematic equation retains its dominant term. For the dynamic equation, we evaluate contribution-ranked supports of size $k = 1, \ldots, 12$ and choose the smallest support within 10% of the minimum short-validation MSE. The validation minimum is $1.108e{-}5$ at $k = 8$; $k = 7$ attains $1.216e{-}5$ and is selected. The fourth measured record is used only for final long-rollout evaluation. This procedure yields

$$\dot{x}_1 = 0.999963\,x_2,$$
$$\dot{x}_2 = -68.871\,x_1 - 1.66979\,x_2 - 16.8541\,x_1x_2 - 87.6603\,x_1^2x_2 + 1.2287\,x_2^3 + 533.214\,x_1^3 - 4732.88\,x_1^4.$$

For interpretation, we group the terms phenomenologically:

$$\dot{x}_2 = \underbrace{-k_{\mathrm{id}}x_1 - c_{\mathrm{id}}x_2}_{\textbf{linear backbone}} - \underbrace{(\alpha_1 x_1 + \alpha_2 x_1^2)x_2}_{\textbf{state-dependent damping}} + \underbrace{\beta x_2^3}_{\textbf{velocity-dependent closure}} + \underbrace{\gamma_3 x_1^3 + \gamma_4 x_1^4}_{\textbf{displacement closure}} .$$

Here $k_{\mathrm{id}} = 68.871$, $c_{\mathrm{id}} = 1.66979$, $\alpha_1 = 16.8541$, $\alpha_2 = 87.6603$, $\beta = 1.2287$, $\gamma_3 = 533.214$, and $\gamma_4 = -4732.88$.

Phenomenological term grouping. *State consistency:* the kinematic coefficient 0.999963 differs from unity by $3.7e{-}5$, consistent with $(x_1, x_2) = (x_3, \dot{x}_3)$. *Linear backbone:* $-k_{\mathrm{id}}x_1 - c_{\mathrm{id}}x_2$ matches the LPF0 mass–spring–damper form; the 0.3449% stiffness difference supports agreement of the linear time scale without identifying either coefficient's hydrodynamic origin. *Displacement closure:* $\gamma_3 x_1^3 + \gamma_4 x_1^4$ captures amplitude-dependent restoring behavior consistent with the nonlinear hydrostatic relation in Kramer et al. (2021), but is not uniquely attributable to hydrostatics. *State-dependent damping:* $-(\alpha_1 x_1 + \alpha_2 x_1^2)x_2$ can absorb radiation memory, draft dependence, and other omitted effects without identifying radiation physics. *Velocity closure:* because $\beta > 0$, βx_2^3 contributes $\beta x_2^4 > 0$ to the kinetic-energy rate and is not dissipative; decay arises from the combined dynamics, and these measurements do not uniquely identify this coefficient's physical source.