EL-KAN: LEARNING SPARSE CONSERVATIVE DYNAMICS FROM POSITION-ONLY TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning interpretable particle dynamics from noisy position trajectories requires identifying both which particles interact and the laws governing their interactions. Uneven coverage of interaction distances and unobserved velocities complicate this task, while accurate trajectory fitting alone does not ensure reliable predictions under distribution shifts or interventions. We introduce the Edge-Law Kolmogorov–Arnold Network (EL-KAN), a framework for learning sparse interaction mechanisms in the generated position-only setting, without velocity, force, or graph supervision. EL-KAN represents pairwise interactions through learnable univariate radial potentials whose analytic derivatives produce conservative, equal-and-opposite forces. Weak-form regression transfers temporal differentiation from noisy position observations to known test functions, while sparse graph selection and regularized coefficient refitting identify the interaction structure and governing laws. To address limited radial resolution and discretization bias, EL-KAN combines structured potential candidates, including locally refined radial bases and compact polynomial–exponential dictionaries, with discretization-consistent refitting. Forecasting uses state estimation from an observed trajectory prefix while keeping the learned force field fixed; independent validation selects the potential representation and initialization scheme based on both identification quality and rollout behavior. Experiments on spring, Duffing, and Morse systems under noisy observations and varying sampling resolutions demonstrate effective recovery of sparse interaction structures and nonlinear interaction laws. Compared with the weak-form and anchor controls under the same validation budget, EL-KAN improves long-horizon prediction within and beyond the training distribution and under edge-removal interventions, while retaining explicit, directly editable interaction mechanisms.

1 INTRODUCTION

Learning interaction mechanisms from motion requires identifying both which particles interact and the laws governing their interactions. We study this problem using noisy, discretely sampled positions in conservative systems with known masses and pairwise radial interactions, without velocity, force, or graph supervision. The goal is to recover explicit mechanisms that support scientific interpretation and prediction when initial conditions or interactions change.

Recent graph-based methods learn interaction rules and latent heterogeneity (Allier et al., 2024) or efficiently fit invariant Hamiltonian dynamics (Rahma et al., 2026). Kolmogorov–Arnold networks represent relationships through learnable univariate edge functions (Liu et al., 2025), with applications to interatomic potentials (Nagai & Okumura, 2024). Weak-form methods complement these representations by avoiding explicit differentiation of noisy observations, including structure-preserving discovery of reduced Hamiltonian systems (Messenger et al., 2024) and spectral test-function selection for robust equation learning (Chen et al., 2026).

Yet predictive accuracy alone does not establish whether recovered interactions remain reliable under changed initial conditions or graph edits. In supervised interatomic potential learning, lower held-out errors can likewise fail to improve downstream physical predictions (Fu et al., 2025). For position-only identification, this distinction motivates evaluating interaction laws separately from the states used to initialize forecasts.

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.

Two coupled challenges arise. First, finite and uneven distance coverage creates a tension between resolving nonlinear forces and constraining behavior in poorly observed regions. Second, unobserved velocities couple forecast initialization to the learned force field: initial-state adjustments may compensate for force-field errors over an observed prefix without improving subsequent dynamics. Neither greater functional flexibility nor a closer prefix fit therefore suffices to establish better mechanism recovery.

We align learnable functions with physical edges and separate law fitting from latent-state estimation, then use held-out dynamics to assess their combination. We implement this principle in the **Edge-Law Kolmogorov–Arnold Network (EL-KAN)**.

EL-KAN has four stages. First, radial edge potentials define a conservative energy whose analytic derivatives yield equal-and-opposite forces. Second, weak-form regression transfers temporal differentiation to known test functions, combining sparse graph selection with regularized, discretization-aware refitting. Candidates span global radial functions, quantile-based local refinements, and compact polynomial or hybrid dictionaries. Third, prefix-based estimators, including shooting, infer forecast states while keeping each force field fixed. Finally, held-out dynamics guide model selection subject to identification, sparsity, and multi-horizon prediction criteria; independent confirmation accepts the selected candidate or triggers fallback. The configuration is frozen before test evaluation and forecasting uses Velocity-Verlet integration (Swope et al., 1982).

Our contributions are threefold:

1. **Sparse mechanism identification.** We introduce a position-only framework for estimating interaction graphs and explicit edge laws, with conservative forces and action–reaction symmetry enforced by construction.
2. **Structured law selection.** We develop a common potential representation spanning flexible radial bases and compact analytic dictionaries, with refitting and validation to select their effective complexity.
3. **Mechanism and initialization assessment.** We separate fixed-field state estimation from law fitting, then assess the combination on held-out dynamics. On nine generated settings this reduces $T{=}5$ test NMSE by 63.4% against a control with the same validation budget but no extra candidates, and the improvement persists under distribution shift and edge removal.

2 Related Work

EL-KAN belongs to position-only mechanism discovery for interacting dynamical systems: the objective is to infer both a sparse physical graph and explicit edge laws from noisy trajectories, without velocity, force, or graph supervision. Prior work has developed expressive physical representations, latent interaction learning, weak-form identification, and robust dynamical evaluation, and the remaining gap is to combine these ingredients under the same information assumptions.

Physical representations and conservative edge laws. Learned dynamics models take a data-driven route to the same quantities: neural ordinary differential equations represent continuous-time motion with a learned vector field (Chen et al., 2018), and Hamiltonian and Lagrangian networks enforce conservation by differentiating a learned energy (Greydanus et al., 2019; Cranmer et al., 2020). Kolmogorov–Arnold networks parameterize computational edges with learnable univariate functions (Liu et al., 2025), with applications to interatomic potentials (Nagai & Okumura, 2024), and Hamiltonian graph networks train invariant Hamiltonian forms efficiently with random features (Rahma et al., 2026). These approaches clarify how to represent physical relationships but do not by themselves identify which physical pairs are active from position-only observations. EL-KAN uses the edge-function principle at the level of candidate particle pairs, assigning each pair a radial conservative potential so that the recovered laws are both differentiable and individually editable.

Sparse graph and law recovery. Neural Relational Inference learns latent interaction graphs jointly with neural dynamics and without graph labels (Kipf et al., 2018), collective variants extend this to heterogeneous interaction types, though with positions, velocities and masses as node features (Han et al., 2024), and graph networks infer latent structure while predicting multi-object motion or latent heterogeneity across systems (Battaglia et al., 2016; Garcia Satorras et al., 2021; Allier et al., 2024). Three recent lines extend this family in different directions: decoupling the forward and

feedback flows of relational inference with dual attention removes the assumption that a pair plays one symmetric role (Zhang et al., 2026); alternating a pretrained structural estimator with a variational joint inference module bootstraps structural priors and refines them from pseudolabels (Gong et al., 2025); and inserting a layer of latent category preferences between observed behaviour and pairwise interaction replaces the categorical relation variable with a preference integrated into opinion dynamics (Yang et al., 2025). All three infer *which* pairs interact, but none exports a per-edge conservative potential: their outputs are attention weights, posterior edge probabilities or preference scores, so they cannot be differentiated into forces, summed into an energy or edited edge by edge. Kernel methods for interacting particles instead recover weighted networks and interaction kernels by alternating regression (Lang et al., 2026b;a), but organise the problem around first-order kernels, so neither family addresses second-order conservative radial potentials with unobserved velocities and per-edge law selection from positions alone. EL-KAN therefore treats graph recovery as sparse selection over physical radial potentials, fitting an independent potential on each retained edge and preserving action–reaction symmetry by construction.

Weak identification and structured refitting. Sparse identification of nonlinear dynamics recovers interpretable equations when the governing terms are already in a candidate library (Brunton et al., 2016), and symbolic regression can widen that search (Udrescu & Tegmark, 2020); because library methods depend on derivative estimates, which are ill-posed for noisy observations (Chartrand, 2011), integral and weak formulations replace pointwise derivatives with test functions (Schaeffer & McCalla, 2017; Messenger & Bortz, 2021), and recent work preserves Hamiltonian structure in coarse-grained systems (Messenger et al., 2024) or selects spectral test functions (Chen et al., 2026). These methods address temporal differentiation and test-function design, but they do not resolve the edge-law model-selection problem caused by uneven distance coverage in position-only particle trajectories. EL-KAN uses weak constraints for fitting while validating a structured candidate pool spanning global radial bases, training-range local refinements, and compact analytic dictionaries, so that functional flexibility is selected by held-out dynamics rather than assumed to be beneficial.

3 METHOD

3.1 PROBLEM SETTING AND EDGE-WISE POTENTIALS

We consider N particles with unit masses in $\mathbb{R}^d$ whose interaction is conservative and pairwise, so one scalar potential describes the system. The coefficient vectors $c_{ij} \in \mathbb{R}^K$ are free parameters here and are determined later from the weak-form equations equation 7–equation 8; equation 1–equation 3 define the model family as a function of them, and are evaluated only once they are fixed.

$$U(\mathbf{q}) \;=\; \sum_{i<j} z_{ij}\, U_{ij}(r_{ij}), \qquad U_{ij}(r) \;=\; \sum_{m=1}^{K} c_{ijm}\, \psi_m(r), \qquad m_i \ddot{q}_i \;=\; -\nabla_{q_i} U(\mathbf{q}), \quad (1)$$

with $r_{ij} = \|q_j - q_i\|$, $\mathbf{q}$ collecting positions of all particles, and $\{\psi_m\}_{m=1}^{K}$ a radial dictionary shared by all edges. We keep the two roles distinct in notation: ψ_m is a shared basis function, whereas ϕ is reserved for the per-edge force law $\phi_{ij} = U'_{ij}$ introduced below. The structure variable $z_{ij} \in \{0, 1\}$ says *which* pairs interact and the coefficients c_{ijm} say *how* a retained pair interacts. With $u_{ij} = (q_j - q_i)/r_{ij}$ the unit vector pointing from i to j, the pair force of a retained edge and the acceleration it induces are

$$f_{ij} \;=\; \phi_{ij}(r_{ij})\, u_{ij}, \qquad m_i \ddot{q}_i \;=\; \sum_{j \neq i} \bar{z}_{ij}\, f_{ij}, \qquad \phi_{ij} \;=\; U'_{ij}, \quad (2)$$

where $\bar{z}_{ij}$ is the retained value of the structure variable. In the dense fit $\bar{z}_{ij} = 1$ for every candidate edge; after pruning it is the fixed binary mask used in the refit and rollout. Because every interaction is pairwise and acts along the line joining the two particles, Newton's third law holds edge-wise, $f_{ji} = -f_{ij}$, so internal forces cancel and linear momentum is conserved. We see positions only, $y_i^{(t)} = q_i(t) + \epsilon_i^{(t)}$, on a regular grid; velocities, forces and the graph are hidden, while masses are known and set to unity in the generated systems (the molecular benchmarks record atomic masses as metadata only and regress the tabulated forces directly). The task is to recover a sparse edge set with an explicit law on each retained edge and then to forecast beyond an observed prefix. Candidate edges form the complete graph $\mathcal{E} = \{(i, j) : i < j\}$ and sparsity is decided after fitting.

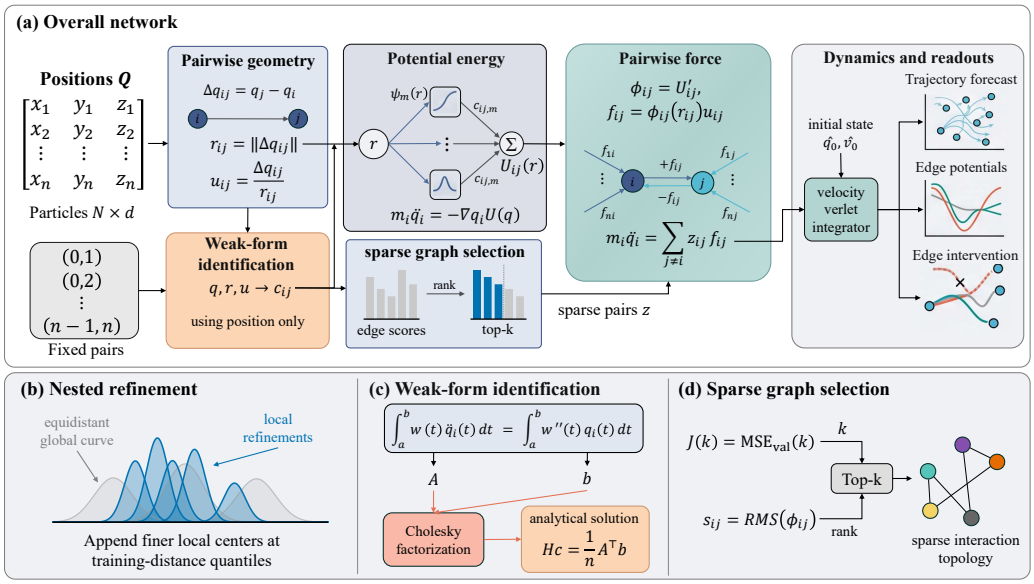

Figure 1: Overview of EL-KAN. (a) Positions feed pairwise geometry, a radial dictionary and a conservative edge potential; the analytic gradient produces equal-and-opposite pair forces, which are assembled and rolled out with Velocity-Verlet. (b) Nested refinement appends finer local centres at training-distance quantiles while retaining the equidistant global basis. (c) Weak-form identification moves the time derivative onto fixed test functions, giving a linear system in the coefficients that is solved by Cholesky factorisation. (d) Edges are scored, ranked and truncated to a sparse graph, which is then refitted.

Three questions organise the method: which pairs interact (z_{ij}, Section 3.3), what law acts on a retained edge (U'_{ij}, Sections 3.2 and 3.3), and whether the mechanism survives a change of state or of graph (Sections 3.4 and 3.4). The generated systems instantiate this with $N = 5$ particles in two dimensions, ten candidate edges present independently with probability 0.45 and weight in $[0.7, 1.3]$, and, with $x = r - 0.85$, the laws Spring $\phi(r) = x$, Duffing $\phi(r) = 0.6x + 0.4x^3$ and Morse $\phi(r) = 1.8(e^{-1.5x} - e^{-3x})$; Morse is non-polynomial and its exponents are never passed as labels to the fitter. The predeclared default slope grid nevertheless contains 1.5 and 3, the two Morse rates, so it is a declared functional prior rather than a law-free basis.

As in a KAN the learnable nonlinearity sits on the edges and a node only sums what arrives; here an edge function reads a pairwise distance, and we use a radial dictionary with closed-form derivatives so a potential yields a force analytically. Our default is a Gaussian radial basis,

$$U_{ij}(r) = \sum_{m=1}^{K} c_{ijm}\,\psi_m(r), \qquad \psi_m(r) = \exp\left(-\frac{(r-\mu_m)^2}{2\sigma^2}\right), \qquad \phi_{ij}(r) = U'_{ij}(r), \quad (3)$$

with fixed centers μ_m and width σ. Because the force depends only on the scalar distance and the relative direction, the field is conservative by construction, the two endpoints receive equal and opposite updates, and translation invariance, rotation/reflection equivariance and internal force balance hold exactly (Appendix B). Equations equation 1 and equation 2 state the final EL-KAN model, in which the structure variable is binary. During fitting the gate is relaxed to a continuous value in $[0, 1]$ through a Hard-Concrete distribution (Louizos et al., 2018), and the same relaxation multiplies the pair force; hardening it recovers the binary form. Because z_{ij} is explicit, the model also has an editing interface: setting $z_{ij} = 0$ removes that pair from equation 1 and equation 2, and re-integrating the frozen field yields the counterfactual dynamics of the edited system, with the pre-intervention state estimated from the unedited model first (Section 4.4).

Figure 1 summarises the full pipeline; each stage is detailed in the subsections below.

3.2 WEAK-FORM IDENTIFICATION AND DISCRETIZATION CONSISTENCY

Multiplying the equations of motion by a smooth weight w supported on a short window and integrating by parts moves the time derivative off the data: whenever w and w' vanish at both ends,

$$\int_a^b w(t)\,\ddot{q}_i(t)\,dt \;=\; \int_a^b w''(t)\,q_i(t)\,dt. \tag{4}$$

Substituting the edge law gives one linear equation per node, dimension and window,

$$\sum_{j\neq i}\sum_{m=1}^{K} c_{ijm}\int_a^b w(t)\,\psi'_m(r_{ij}(t))\,u_{ij}(t)\,dt \;=\; \int_a^b w''(t)\,q_i(t)\,dt, \tag{5}$$

so the coefficients enter linearly. Rows are indexed by $r=(w,i,\alpha)$, a window, node and dimension, and columns by $p=(ij,m)$, an edge and atom:

$$A_{r,p}=\iota_{r,(ij)}\int_{t_a}^{t_b} w(t)\,\psi'_m\big(r_{ij}(t)\big)\big[u_{ij}(t)\big]_\alpha\,dt, \qquad b_r=\int_{t_a}^{t_b} w''(t)\,q_{i,\alpha}(t)\,dt, \tag{6}$$

where $\iota_{r,(ij)}=+1$ if row r belongs to the source endpoint of (i,j), -1 if it belongs to the sink endpoint, and 0 otherwise, so every edge contributes equal and opposite rows and Ac accumulates edge forces. Both A and b are built from the observed positions alone, so A and b are fixed before any fitting. Collecting rows gives

$$\min_c\;\frac{1}{n}\|Ac-b\|_2^2 \;+\; \lambda\sum_{(i,j)\in\mathcal{E}}\|c_{ij}\|_2^2 \;+\; \nu\sum_{(i,j)\in\mathcal{E}}\|Dc_{ij}\|_2^2, \tag{7}$$

with D a second-difference operator and (λ,ν) chosen on held-out trajectories; equation 7 is first solved on the complete candidate set, so that all candidate edges receive a coefficient before any is discarded. Only retained edges carry a free coefficient in the per-budget refits, so writing $B=A_{:,\mathcal{E}_k}$ for the design restricted to the top-k edges of the edge-score ranking, each refit is solved by the normal equations

$$Hc=g, \qquad H=\tfrac{1}{n}B^\top B+\lambda I+\nu\,\mathrm{blkdiag}(D^\top D), \qquad g=\tfrac{1}{n}B^\top b, \tag{8}$$

with $c_{ij}=0$ for $(i,j)\notin\mathcal{E}_k$. The ridge term makes H symmetric positive definite, and its Cholesky factor L, $H=LL^\top$, gives $c=L^{-\top}L^{-1}g$ without forming an inverse. With c fixed, equation 1–equation 3 are then evaluated directly to obtain U_{ij} and ϕ_{ij}, and no solve is repeated during evaluation or rollout. We take $w(x)=(x(1-x))^3$ and solve in double precision, while the finite-difference control replaces the weak design and target with instantaneous rows and raw second differences. At coarse sampling the weak identity itself acquires a bias: pairing the centered position difference with the point acceleration is only $\mathcal{O}(\Delta t^2)$ accurate, because the former carries a $\Delta t^2 q''''/12$ term the latter does not. The corrected candidate matches the left side to the three-point average $\tilde{a}_t=\frac{1}{12}(a_{t-1}+10a_t+a_{t+1})$, leaving $\mathcal{O}(\Delta t^4)$; it changes only the fitted equation, and Section 4.4 quantifies its effect.

3.3 STRUCTURED DICTIONARIES AND SPARSE GRAPH SELECTION

A single global basis is too coarse wherever a law changes quickly, so we build several dictionaries inside the same edge-wise structure and let validation choose. Nested refinement appends $K_{\mathrm{loc}}=48$ functions at quantiles of the pooled training edge-distance distribution (widths 1.5 times the local spacing, coefficients starting at zero so the global solution stays exactly recoverable), and a variant places 96 functions on a *uniform* grid, separating added capacity from data-driven placement. Two compact analytic dictionaries are fitted as well, a truncated polynomial and its hybrid extension with paired exponentials,

$$U_{ij}^{\mathrm{hyb}}(r)=\sum_{k=1}^{P}\frac{c_{ijk}}{k}\Big(\frac{r}{\varsigma}\Big)^k-\sum_{s\in S}\Big[\frac{a_{ij,s}}{s}e^{-s(r-\kappa)}+\frac{b_{ij,s}}{2s}e^{-2s(r-\kappa)}\Big], \tag{9}$$

with independent coefficients for the two exponential terms, so each edge has $P+2|S|$ coefficients. The exponential atoms can supply the two decay rates of a Morse law, and because their coefficients

can take opposite signs a signed combination reproduces the attractive–repulsive form whenever the slope set makes those rates reachable. In the generated Morse setting this is an explicitly declared prior, not a claim that the dictionary is mechanism-agnostic: the exponentials carry the short-range branch and the truncating polynomial supplies the long-range one. A Morse-like mechanism is therefore representable without passing a law label to the fitter, but the dictionary is a functional prior rather than a family that is Morse-shaped by itself. The RMS radius ς and median radius κ come from training distances only, and the dictionary is parameterized by the polynomial count P and the slope set S. Sparsity comes after fitting, in two stages. First equation 7 is solved once on the full candidate set, which gives every edge a coefficient from which an RMS force score is read off,

$$s_{ij} \;=\; \Big(\frac{1}{T}\sum_{t=1}^{T}\phi_{ij}\big(r_{ij}(t)\big)^2\Big)^{1/2}, \tag{10}$$

Then the edges are ranked by this score, and a nested family of masks is formed by taking the top-k edges; the ranking is computed once and is not revisited, so the budget only sets where the ranking is truncated. For every budget k the top-k mask is refit by equation 7 with $B = A_{:,\mathcal{E}_k}$ and λ re-selected on validation within the refit, and the result is scored by

$$J(k) \;=\; \mathrm{WeakNMSE}_{\mathrm{val}}(k) \;+\; \eta\, k/|\mathcal{E}|, \qquad \eta = 10^{-3}, \tag{11}$$

so the laws settle before any edge is removed. For the molecular benchmarks, where the true graph is a proxy, we instead keep the *sparsest* mask within a validation tolerance $\tau = 10^{-2}$ of the best. A Hard-Concrete gate (Louizos et al., 2018) trained jointly with the weak objective followed by least-squares refitting agrees in edge count and F1 with score-ranked pruning in 43 of 45 runs; both controls remain in the candidate pool. We also record an observation-only extrapolation risk (the 95th force and slope percentiles just outside the observed radius range relative to inside); Section 4.6 shows validation error predicts held-out error better.

3.4 State estimation, candidate competition and guarded selection

A rollout needs a state at the prefix end, and the initial velocity is latent. The default estimator fits a degree-2 polynomial to the prefix; a second solves for the state by an independent damped Gauss-Newton least-squares fit per trajectory (eight iterations, damping 10^{-6}); and a third, shooting, optimizes an additive correction δ so that integrating the frozen field from the corrected state reproduces the prefix. Let Let L be the number of observed prefix frames, $\tau = (L-1)\Delta t$, and s_q the training-position RMS. Direct shooting starts from the quadratic state, scales position and velocity corrections by s_q and s_q/τ, and solves

$$\min_{\delta}\; \frac{1}{LNd}\sum_{\ell=1}^{L}\left\|\frac{\hat{q}_\ell(\theta,\delta)-y_\ell}{s_q}\right\|_2^2 + \frac{\rho}{2Nd}\|\delta\|_2^2, \qquad \rho = 10^{-4}, \tag{12}$$

with LBFGS, at most 20 iterations and a strong-Wolfe line search. The least-squares, direct-shooting, and adaptive-blend estimators are separate candidates. Only the adaptive candidates interpolate the quadratic and shooting states using support, prefix consistency, and a correction cap; a failed numerical estimate records and uses the quadratic fallback. The force field remains frozen throughout state estimation, and forecasting uses Velocity-Verlet with internal step $\Delta t/4$.

Per run we then build a 32-entry tuning pool from these fitted quantities – factorial configurations, α-interpolations between the global and refined potentials, β state blends, corrected and uncorrected refits, the uniform-grid refinement, compact polynomials and hybrid dictionaries (Appendix C) – and select on independent ID and stress validation streams. Two frozen reference anchors are first constructed; a candidate must be no worse than both anchors in weak error, edge count, and every ID/stress horizon mean, while its terminal 95th-percentile error must remain within the larger anchor envelope. Eligible candidates are ranked by the mean relative rollout error across anchors, domains, and horizons. Confirmation rechecks only the winner; failure deploys the frozen anchor fallback. Every decision is written before test evaluation; the complete procedure is given in Algorithm 1. Section 4.7 audits the failures.

4 EXPERIMENTS

4.1 SETUP

Datasets. Twelve datasets are used (Appendix Table 3): three public benchmarks and nine generated systems. Only the generated systems have ground truth for both graph and law, so they carry the mechanism claims while the benchmarks carry transfer; difficulty rises from clean to noisy to coarse sampling and Morse is the hardest law.

Generated systems and data. The generated systems of Section 3 are observed under three regimes: clean ($\Delta t = 0.02$), noisy (the same, with i.i.d. Gaussian noise at 1% of the training position RMS), and coarse ($\Delta t = 0.04$). Each run contains $32/8/32/16$ train, validation, test, and OOD trajectories. Training trajectories have 251 frames (126 in the coarse setting); validation, test, and OOD trajectories have 261 frames (131 coarse), with an 11-frame prefix (6 coarse). We evaluate $T = 1, 2, 5$, and OOD scales the initial distributions by 1.8. Fitting and selection see noisy observations where applicable; all errors use clean references.

Public data. rMD17/Aspirin and MD22/DHA are force-regression benchmarks. NRI/Springs-5 has 5 balls and 10 candidate edges; its conditioned model reads five context frames containing positions and velocities, predicts acceleration targets estimated from velocities, and uses graph labels for training and validation threshold selection. Molecular graph truth is a covalent-radius proxy rather than a measured bond table. Public NMSE divides target MSE by the squared RMS of the training target: force for the molecular tasks and estimated acceleration for NRI. The public results are separate applicability experiments, and no external baseline was run on them.

External baselines. On the generated systems we compare with three published relational-inference methods: DARI-P (Zhang et al., 2026), IPSI-P (Gong et al., 2025), and BINN-P (Yang et al., 2025). We re-fit each released model as a position-only adaptation by replacing velocity or force inputs with a five-frame position prefix and using the same generated splits; details of the adaptation and its evaluation budget are given in Appendix C. The suffix $-\mathrm{P}$ denotes our adaptation, and BINN-P outputs preferences rather than an adjacency, so its structural metrics are not applicable.

Metrics and protocol. We report position NMSE (mean squared future position error over the squared training-position RMS), initial-velocity MSE, acceleration NMSE, graph F1 and AUROC, exact-graph recovery, edge count, relative L_2 law error, OOD NMSE and intervention NMSE; the public suite reports the same NMSE on forces with F1, AUROC and top-k variants at the true edge count. The main grid is 3 laws $\times$ 3 regimes $\times$ 5 seeds $= 45$ runs, each reporting 32-entry tuning pools over shared force fields, with a coarse supplement at $T = 2, 4, 10$ (15 runs), a Morse dictionary sweep over 27 settings (135 runs) and the public suite (9 runs); the run inventory, the automated checks and the $13{,}860$ converged solver trials are tabulated in Appendix C. Means are over five seeds with sample standard deviations. The external comparison uses three paired seeds and is reported separately from the five-seed main result; its dependence and exploratory interval convention are described in Appendix C.

4.2 MAIN RESULTS AND GUARDED SELECTION

Table 1 reports the five-seed main result together with a three-seed EL-KAN subset paired with the external adaptations. In that paired comparison, EL-KAN has the lowest error at every horizon: BINN-P's $T = 5$ NMSE is $83\times$ the paired EL-KAN value (0.790 against 9.5×10^{-3}), while DARI-P and IPSI-P are $113\times$ and $156\times$ higher. Graph recovery separates the methods more sharply: EL-KAN recovers 24 of 27 graphs exactly, with mean SHD 0.148. DARI-P reaches F1 0.392 and IPSI-P 0.167 at SHD 4.37 and 4.07, and neither recovers a single graph exactly. BINN-P does not emit an adjacency. These results separate predictive accuracy from mechanism recovery: the strongest external predictor is not the one closest to the true graph, and the two methods that produce graphs remain close to guessing. Against our own generations the final EL-KAN beats the weak-form control by 95.7% at $T = 5$, the anchor control by 62.0%, and the equal-budget control by 63.4%. Figure 2 shows the per-setting reductions against these controls.

Per seed at $T = 5$ we are lower in 79 of the 81 out-of-distribution pairs at a median ratio of 1.2×10^{7} and never lose to DARI-P or IPSI-P. The two exceptions are BINN-P on noisy settings, where our NMSE is 10% higher on Spring and 4% higher on Duffing; both fall on the law we already flag as hard under noise, and we report them rather than dropping them.

Table 1: Generated systems, macro-averaged over the nine settings. Weak and Anchor are five-seed weak-form and anchor controls. EL-KAN (45) is the final five-seed main result; EL-KAN (paired) uses the three seeds paired with the external adaptations. DARI-P, IPSI-P, and BINN-P are fixed-protocol position-only adaptations using those three paired seeds. NMSE is test position error, OOD is the $1.8\times$ initial-condition split, and Law L_2 is relative edge-law error; lower is better except for F1 and the exact count. AUROC and SHD are unavailable for Weak and Anchor, which share one candidate's graph, and structural columns are not applicable for BINN-P, which predicts pairwise preferences rather than an adjacency. SHD is out of ten candidate edges, so 4.4 is near chance.

	Prediction				Mechanism			
Model	T=1	T=2	T=5	OOD	Law L_2	F1	AUROC	SHD
Weak	0.00448	0.0163	0.124	1.72	0.145	0.974	–	–
Anchor	0.000931	0.00312	0.0174	1.12	0.0429	0.991	–	–
EL-KAN (45)	**0.000382**	**0.00127**	**0.0105**	**0.221**	**0.0302**	**0.991**	**1.000**	**0.089**
EL-KAN (paired)	0.000338	0.00110	0.00947	0.188	0.0331	0.985	1.000	0.148
DARI-P	0.0296	0.147	1.07	3.00	–	0.392	0.663	4.37
IPSI-P	0.0523	0.249	1.48	3.07	–	0.167	0.585	4.07
BINN-P	0.0154	0.0925	0.790	2.65	–	n/a	n/a	n/a

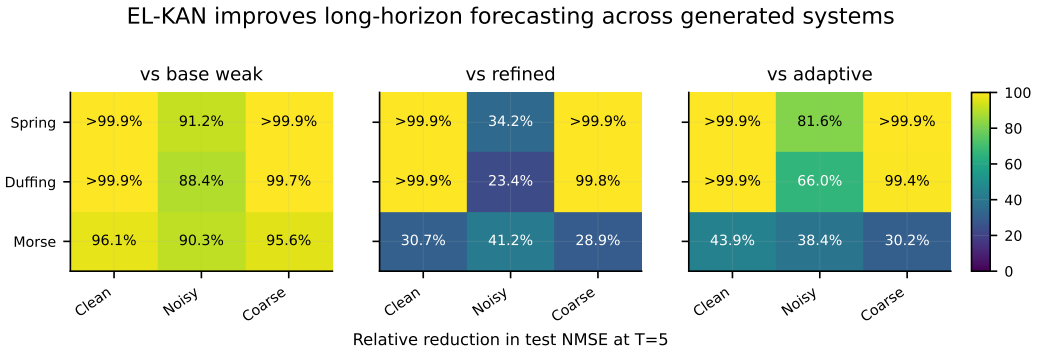

Figure 2: Long-horizon ($T = 5$) relative reduction of EL-KAN error against the weak-form control, the anchor control, and the adaptive variants across the nine generated settings. Clean and coarse reductions are often larger than noisy reductions because the final model nearly removes the remaining integration error in those settings.

The guard determines when the selected configuration is retained. In 33 of 45 runs it selects a configuration outside the anchor control; 12 retain the anchor, and confirmation accepts 34 and rejects 11. The selected family is a compact polynomial in 24 runs, the anchor in 12, and a hybrid dictionary in 9. No single dictionary wins everywhere, which motivates selection within the pool. The guard cuts Spring noisy $T = 5$ NMSE from 2.22×10^{-1} to 1.96×10^{-2}; per setting (Appendix Table 4) it is never worse than the weak-form control.

The public benchmarks (Appendix Table 2) probe different supervised targets and graph definitions. MD22/DHA has perfect ranking against the covalent-radius proxy (AUROC and top-k F1 both 1.000), but the threshold-selected mask retains 561 of 1,540 edges and has F1 0.179; this shows that score ranking and proxy-graph sparsification can disagree, without proving that the extra edges are spurious. NRI/Springs-5 predicts estimated accelerations from velocity-containing contexts and uses graph labels; it reaches edge F1 0.780 ± 0.003 and AUROC 0.899 ± 0.001 at 6.56 predicted edges against 5 true edges. rMD17 has the lowest force NMSE (0.0336 ± 0.0009) but retains 174 of 210 candidate edges, so its predictive and structural results should be read separately. These public experiments support applicability under their stated supervision, not position-only identification or a comparison with external force models.

4.3 MECHANISM RECOVERY

Pooled over the nine settings the final EL-KAN reaches a relative edge-law error of 0.0302 ± 0.075 against 0.145 ± 0.19 (weak-form control) and 0.0429 ± 0.087 (anchor control), graph F1 of 0.9911 ± 0.035, and the exact graph in 42 of 45 runs, including all 15 clean and all 15 coarse runs. On the hardest mechanism, Morse under noise, the law error falls from 0.405 and 0.274 to 0.200 – better, but still far above the clean Morse value of 0.0044, so we do not claim accurate Morse recovery under noise. Edge *ranking* was already perfect (AUROC 1.000 in all 45 runs for all three models): what improves is edge-count and threshold selection.

4.4 CONTROLS AND SAMPLING

With 1% position noise the finite-difference control selects the empty graph on all three laws (F1 $= 0$), ranks edges at chance (AUROC 0.31, 0.45, 0.50) and gives $T = 5$ NMSE 1.431, 1.227 and 1.375 against 0.222, 0.219 and 0.403 for score-ranked weak-form pruning, so weak-form fitting earns its keep in noise and on the laws a second difference cannot resolve. Keeping all ten edges scores F1 0.626 but AUROC 1.000 (trivial over-selection).

Discretization-consistent refit. The discretization-consistent refit is a separate candidate, so its effect is directly observable: it improves the learned law almost everywhere but rollout error does not follow, rising by 399% on clean and coarse Spring and Duffing and falling on Morse, where law error dominates (Appendix Table 6). Thus, lower discrete-equation error does not guarantee lower rollout error, which is why the correction is evaluated as a separate candidate.

Graph intervention and distribution shift. Deleting one edge chosen only by the training-time dense score moves the forecast toward the post-edit reference in all 45 runs, with intervention NMSE 0.00968 ± 0.014 against 0.0152, a 36% reduction, so learned edges behave like removable mechanisms. The coarse supplement doubles the horizons to $T = 10$ and preserves the ordering (Spring 3.0×10^{-7}, Morse 6.1×10^{-2}, Duffing 8.1×10^{-4}).

4.5 PREFIX STATE AND LONG HORIZONS

The final EL-KAN curve is lowest at every horizon in all nine settings (Appendix Figure 3). Because selection jointly changes the force field and the state estimator, initial-velocity and prefix-reconstruction aggregates do not by themselves isolate the two contributions.

4.6 DICTIONARY ABLATION

Sweeping the polynomial count $P \in \{0, 2, 4\}$ and three slope sets S on Morse (Appendix Table 7) shows that neither slope-set size nor polynomial count is uniformly optimal: selected-candidate NMSE and law error remain lowest for different choices across clean, noisy, and coarse Morse. The observation-only extrapolation risk is more weakly correlated with held-out error (0.348–0.469) than validation error (0.614–0.962), so it is reported diagnostically and never used for selection.

4.7 WHERE VALIDATION-BASED SELECTION STILL FAILS

Confirmation rejects $11/45$ candidates (7 noisy, 2 clean, 2 coarse); all final selected forecasts are finite. The remaining horizon and stress regressions may reflect force-field extrapolation, state estimation, or integration error, motivating the frozen-field and true-state controls reported in the appendix.

5 CONCLUSION

EL-KAN recovers sparse conservative interaction mechanisms from position-only trajectories by combining conservative edge potentials, weak-form identification, prefix-based state estimation and protected selection over structured dictionaries. On the generated systems, the final selection improves long-horizon rollouts while retaining explicit and editable interaction mechanisms.

AI USE STATEMENT

We used generative AI tools for language polishing, formula typesetting, figure-caption editing, and preliminary literature-search assistance. We did not use them to develop the methodology, theoretical claims, experimental design or implementation, or to analyze or interpret results. We reviewed all AI-assisted material against the original arguments, and two authors independently verified the experiment descriptions; we take responsibility for the final manuscript and all associated artifacts.

ETHICS STATEMENT

This study evaluates EL-KAN on simulated, synthetic, and publicly available benchmark datasets. No new data are collected from human participants, and no experiments involving human subjects or deployed decision-making systems are conducted. The datasets are used for model training, validation, and evaluation within the experimental protocols described in this paper.

REPRODUCIBILITY STATEMENT

Code and Implementation. Anonymous source code, pre-processed dataset files, configuration files, and running scripts are provided in the supplementary material archive. All experiments are implemented in PyTorch, with training conducted on a single NVIDIA A100-SXM4-40GB GPU. Complete software dependency versions are listed in `requirements.txt` within the archive. Reproduction shell scripts are provided to launch experiments for all prediction horizons over each benchmark dataset.

Datasets. We use twelve datasets: nine generated systems and three public benchmarks. The generated systems contain $N = 5$ particles in two dimensions with unit masses and ten candidate edges, each present independently with probability 0.45 and weight in $[0.7, 1.3]$, under three interaction laws (Spring, Duffing and Morse); here Morse is non-polynomial and its exponents are never passed to the model. Each law is observed under clean, noisy (1% of the training position RMS) and coarse ($\Delta t = 0.04$) sampling, giving the nine settings. The public benchmarks are force-regression datasets: rMD17/Aspirin (21 atoms, 210 candidate edges), MD22/DHA (56 atoms, 1,540 candidate edges) and NRI/Springs-5 (5 balls, 10 edges), the last also in a trajectory-conditioned variant. The molecular targets are forces; NRI targets acceleration estimated from velocity and uses graph labels in its conditioned pipeline. Ground-truth graphs for the public benchmarks are a covalent-radius proxy rather than a measured bond table, and public NMSE is normalized by the squared training-target RMS. No custom filtering or manual data cleaning beyond standard pre-processing is applied.

Hyper-parameters and Experiment Protocols. The radial dictionary, weak-form window, regularization grid, candidate pool and estimator settings are summarized in Appendix C. Random seed is fixed to 0 for PyTorch, NumPy and CUDA to ensure deterministic behavior; all linear algebra is float64 and the gate optimization uses `torch.use_deterministic_algorithms`. Model selection uses independent tuning, guard and confirmation streams generated from seeds disjoint from the test and OOD streams, and every selection decision is written before test evaluation. Ablation protocols for the slope set, polynomial count, discretization-consistent refit, sampling regime and single-edge graph intervention are documented in Appendix C and Section 4.6. Evaluation metrics are position NMSE, acceleration NMSE, graph F1 and AUROC, relative L_2 edge-law error, and out-of-distribution and post-intervention NMSE computed on held-out test sets.

REFERENCES

Cédric Allier, Magdalena C. Schneider, Michael Innerberger, Larissa Heinrich, John A. Bogovic, and Stephan Saalfeld. Decomposing heterogeneous dynamical systems with graph neural networks. *arXiv preprint arXiv:2407.19160*, 2024. doi: 10.48550/arXiv.2407.19160.

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. *NeurIPS*, 2016.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS*, 113(15):3932–3937, 2016.

Rick Chartrand. Numerical differentiation of noisy, nonsmooth data. *ISRN Applied Mathematics*, 2011.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

Zhiheng Chen, Urban Fasel, and Anastasia Bizyaeva. Fourier Weak SINDy: Spectral test function selection for robust model identification. In *Proceedings of the 8th Annual Learning for Dynamics and Control Conference*, volume 331 of *Proceedings of Machine Learning Research*, pp. 557–573. PMLR, 2026.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. In *ICLR Workshop*, 2020.

Xiang Fu, Brandon M. Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17875–17893. PMLR, 2025.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *ICML*, 2021.

Zhongben Gong, Xiaoqun Wu, and Mingyang Zhou. IPSI: Enhancing structural inference with automatically learned structural priors. In *Advances in Neural Information Processing Systems*, volume 38, pp. 173394–173412, 2025.

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

Zhichao Han, Olga Fink, and David S. Kammer. Collective relational inference for learning heterogeneous interactions. *Nature Communications*, 15:3191, 2024. doi: 10.1038/s41467-024-47098-7.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2688–2697. PMLR, 2018.

Quanjun Lang, Xiong Wang, Fei Lu, and Mauro Maggioni. Learning multi-type heterogeneous interacting particle systems. *arXiv preprint arXiv:2602.03954*, 2026a. doi: 10.48550/arXiv.2602.03954.

Quanjun Lang, Xiong Wang, Fei Lu, and Mauro Maggioni. Interacting particle systems on networks: Joint inference of the network and the interaction kernel. *Applied and Computational Harmonic Analysis*, 84:101867, 2026b. doi: 10.1016/j.acha.2026.101867.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–Arnold networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l_0 regularization. In *ICLR*, 2018.

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

Daniel A. Messenger, Joshua W. Burby, and David M. Bortz. Coarse-graining Hamiltonian systems using WSINDy. *Scientific Reports*, 14:14457, 2024. doi: 10.1038/s41598-024-64730-0.

Yuki Nagai and Masahiko Okumura. Kolmogorov–Arnold networks in molecular dynamics. *arXiv preprint arXiv:2407.17774*, 2024. doi: 10.48550/arXiv.2407.17774.

Atamert Rahma, Chinmay Datar, Ana Cukarska, and Felix Dietrich. Rapid training of Hamiltonian graph networks using random features. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hayden Schaeffer and Scott G. McCalla. Sparse model selection via integral terms. *Physical Review E*, 96(2):023302, 2017.

William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *Journal of Chemical Physics*, 76(1):637–649, 1982.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

Yulong Yang, Bowen Feng, Keqin Wang, Naomi Leonard, Adji Bousso Dieng, and Christine Allen-Blanchette. Behavior-inspired neural networks for relational inference. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4312–4320. PMLR, 2025.

Juyuan Zhang, Xiaoxiao Liang, Chenghua Gong, Rui Sun, Liming Pan, and Linyuan Lü. Decoupling forward and feedback flows: A dual-attention framework for relational inference. *Knowledge-Based Systems*, 348:116405, 2026. doi: 10.1016/j.knosys.2026.116405.

Algorithm 1 Standard EL-KAN workflow. All model and estimator choices are frozen before any mechanism or test-trajectory metric is computed.

Input: position trajectories y, candidate edges $\mathcal{E}$, radial dictionaries, ridge grid Λ, smoothness ν, complexity weight η
Output: frozen edge laws ϕ_{ij}, selected mask z, rollout $\hat{q}(t)$

$A, b \leftarrow$ weak design from equation 6 ▷ positions only; windows never cross trajectories
$c^{\mathrm{d}} \leftarrow \textsc{Fit}(A, b)$ by Cholesky ▷ dense fit on $\mathcal{E}$
$s_{ij} \leftarrow \mathrm{RMS}_t\ \phi_{ij}(r_{ij}(t))$ and $\pi \leftarrow \mathrm{argsort}(s, \downarrow)$
for $k = 0, \ldots, |\mathcal{E}|$ **do**
    form top-k mask $\mathcal{M}_k$ from π and refit c_k on the fit-validation split using equation 8
end for
build quantile-refined, uniform-refined, polynomial, and hybrid dictionaries from training distances only
refit every dictionary candidate and combine it with the quadratic, least-squares, direct-shooting, or adaptive prefix estimator
construct the 32-entry tuning pool and two frozen anchor controls
freeze the fallback anchor before confirmation
for each candidate m in the tuning pool **do**
    evaluate weak error, edge count, ID/stress means, and terminal 95th percentiles on streams independent of test data
    mark m eligible only if it satisfies the constraints relative to both anchors and all forecasts are finite
end for
select the eligible candidate minimizing mean relative rollout error across anchors, domains, and horizons
recheck only the winner on confirmation streams; on failure deploy the frozen fallback anchor
lock all choices before mechanism and test evaluation; roll out with Velocity-Verlet at internal step $\Delta t/4$

Appendix C lists the exact data, dictionary, regularization, and estimator settings.

A Workflow summary

B Structural properties

The properties below follow directly from equation 3; we list them to make clear which invariances are exact rather than learned.

Conservativity. With f_{ij} and $\phi_{ij} = U'_{ij}$ as in equation 2, the edge contribution satisfies $m_i \ddot{q}_i = -\nabla_{q_i} U_{ij}(r_{ij})$, with the pair potential U_{ij}. Summing over edges, the accelerations are the negative gradient of $U = \sum_{i<j} U_{ij}(r_{ij})$, and any Velocity-Verlet rollout of the learned field is a discretization of a conservative system.

Action and reaction. For each unordered pair the update adds f_{ij}/m_i to i and $-f_{ij}/m_j$ to j, so $f_{ji} = -f_{ij}$ and the internal forces cancel in $\sum_i m_i \ddot{q}_i$. Linear momentum is therefore conserved exactly at the level of the force field, before any time discretization.

Rigid-motion equivariance. Every quantity in equation 3 depends only on the relative displacement $q_j - q_i$. A common translation therefore leaves all relative displacements, distances, and unit directions untouched. For an orthogonal matrix R, relative displacements become $R(q_j - q_i)$, distances are unchanged, and each unit direction is multiplied by R, so the predicted accelerations $a_i = \ddot{q}_i$ of equation 2 transform as Ra_i. The model is exactly translation invariant and exactly equivariant to rotations and reflections.

Weak-form identity. For w with $w(a) = w(b) = w'(a) = w'(b) = 0$, integration by parts gives $\int_a^b w\ddot{q}\,dt = [w\dot{q} - w'q]_a^b + \int_a^b w''q\,dt = \int_a^b w''q\,dt$, which is equation 4. The weight $w(x) = (x(1-x))^3$ used in the experiments satisfies both boundary conditions.

C Experimental and implementation details

Dictionary. For generated systems the global basis uses $K = 24$ Gaussian centers spaced uniformly between $r_{\min} = 10^{-3}$ and $r_{\max} = \max(2,\ 1.2 \max_{t,i,j} r_{ij}(t))$ computed from training positions, with width $\sigma = 1.5\,(r_{\max} - r_{\min})/(K-1)$. Nested refinement appends $K_{\mathrm{loc}} = 48$ centers at $\mathrm{linspace}(0, 0.99, 48)$ of the pooled training edge-distance distribution, with widths 1.5 times the local center spacing, floored at $0.15\,\mathrm{span}/(K_{\mathrm{loc}} - 1)$, and the uniform variant uses 96 centers on

a uniform grid over the training range. The hybrid dictionary is parameterized by the polynomial count P and the slope set S, with $P+2|S|$ coefficients per edge (P polynomial and two independent exponentials per slope), with ς the RMS training radius and κ the median one; the default grid is $P \in \{2, 4\}$ and $S = \{\{.75, 1.5, 3\}\}$, and the ablation varies both.

Weak-form discretization. The window has 21 frames at $\Delta t = 0.02$ and 11 frames at $\Delta t = 0.04$, with stride 10 and 5 respectively; the weight is $w(x) = (x(1-x))^3$ on the normalized window coordinate, normalized to unit sum, with w'' formed by central differences of the zero-padded weight. Ridge values are searched over $\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$, the smoothness weight is 10^{-5}, the local multiplier is chosen from $\{1, 10\}$, the complexity weight in equation 11 is $\eta = 10^{-3}$, and refits use smoothness scales $\{1, 0.1, 0.01\}$. All linear solves are float64 with an explicit residual check.

Estimators. The quadratic estimate fits a degree-2 polynomial in normalized time to the prefix. The least-squares estimator solves each trajectory independently by damped Gauss-Newton with exact forward-mode Jacobians, 8 iterations and damping 10^{-6}. Shooting starts from the quadratic estimate, uses LBFGS with at most 20 iterations and a strong-Wolfe line search, a prior strength of 10^{-4} on the scaled correction, and an internal rollout step of $\Delta t/4$; the blend weight uses a support margin of 0.1 and caps the correction at 0.25 of the observation scale. Model parameters are frozen and their requires-grad flags restored afterwards, and reported rollouts use Velocity-Verlet with the same internal step.

Validation budgets. Each tuning, guard, and confirmation round uses a 32-trajectory ID stream and a 16-trajectory stress stream; the stress stream uses $1.4\times$ the initial-condition scale. All streams are generated from seeds independent of training, validation, test, and OOD. Confirmation is opened only for the final winner, and the anchor controls use the same stream budget without the additional candidates.

Data. Initial positions are drawn from $\mathcal{N}(0, 0.45^2)$ and initial velocities from $\mathcal{N}(0, 0.18^2)$, both centered to remove center-of-mass motion; reference trajectories are integrated with Velocity-Verlet at an internal step of 0.002. Splits contain $32/8/32/16$ trajectories for train/validation/test/OOD; training and validation weak-form fitting use the first 251 frames (126 in the coarse scenario), and test and OOD trajectories contain the 11-frame prefix (6 in coarse) plus 250 forecast frames (125 in coarse). Noise is i.i.d. Gaussian with standard deviation 1% of the training position RMS, and the OOD split multiplies the initial position and velocity scales by 1.8.

Public pipeline. The molecular benchmarks are force-regression snapshots; NRI is an acceleration-regression task whose targets are estimated from velocity. Candidate edges are all pairs, scanned over at most 2048 frames; $r_{\min} = 0.03$ and $r_{\max}$ is the $1.05\times$ maximum pairwise distance over train, validation and test snapshots, with the radial width either $1.5(r_{\max} - r_{\min})/(K-1)$ or a fixed 0.18 for rMD17; $K = 32$. Masks come from the sparsity-first rule with tolerance $\tau = 10^{-2}$, sweeping 24 thresholds and 24 budgets on rMD17 and 12/16 on MD22, followed by a single ridge refit at 10^{-4}. NRI uses the trajectory-conditioned model with 5 context frames, an encoder pretrained on graph labels for 30 epochs with weight 1.0 and an L_1 sparsity term of 0.003; its graph threshold is selected on validation edge labels. Public NMSE divides target MSE by the squared training-target RMS, using force targets for rMD17/MD22 and estimated acceleration targets for NRI.

External baseline adaptation. DARI-P (Zhang et al., 2026) retains dual attention between forward and feedback flows; IPSI-P (Gong et al., 2025) alternates structural pretraining with variational refinement; and BINN-P (Yang et al., 2025) predicts pairwise preferences over latent categories. The suffix `-P` denotes our fixed position-only adaptation, not an author-provided model variant. For each method we replace velocity or force inputs with a five-frame position prefix and retrain on the generated train/validation/test/OOD splits. Released defaults, batch sizes, training budgets, and checkpoint-selection rules are retained where compatible but are not identical to the EL-KAN protocol; in particular, DARI-P uses batch 1024 versus the released default 64. The comparison therefore evaluates a fixed-protocol adaptation rather than reproducing published scores. The baseline results use three seeds paired with the same five underlying graphs across the nine generated settings, while the main EL-KAN result uses five seeds. Macro averages are exploratory, with $\mathrm{df} = 4$ intervals and no multiplicity correction; repeated graphs mean the runs are not independent replicates. BINN-P does not output an adjacency, so its structural metrics are reported as N/A.

Metrics. Position NMSE divides the mean squared future position error by the squared training-position RMS. The law metric is the relative L_2 error of the learned radial law against the reference

law on each true edge's training 5–95% radius interval, restricted to true edges. Graph metrics compare the selected mask with the ground-truth edge set, and AUROC uses all candidate edges; the top-k variants take the k highest scoring edges with k the true edge count. For generated position-only experiments, prohibited inputs are enforced by the archived manifests: no true velocities, forces, graph labels, or simulator coefficients enter fitting or selection. The public NRI protocol is a separate supervised exception: it supplies velocity-containing context and graph labels as specified above.

D Additional scope and limitations

Four boundaries affect how these results should be read.

Capacity versus placement. The nested model has 72 basis functions per edge against 24 for the base model, so the refinement comparison mixes extra capacity with quantile-based placement. The 96-center uniform variant is a partial control and is kept in the candidate pool, but an equal-capacity global control is still the clean experiment.

Guard versus risk. Validation reduces test degradation without removing it. The final EL-KAN loses to the anchor control in 4 of 45 runs and to the equal-budget control in 1 (by $1.02\times$, inside noise). Of the 11 confirmation rejections, 7 are noisy, 2 clean, and 2 coarse. The failures trace to prefixes whose true-edge distances leave the observed range, state estimation, or integration error; separating these causes requires frozen-field and true-state controls on a locked split.

Public graph labels. rMD17 and MD22 have no measured bond table in this suite, so their graph metrics are agreement with a covalent-radius proxy; only NRI carries per-sample labels. Their NMSE uses force targets for rMD17/MD22 and estimated acceleration targets for NRI, and no external baseline was run, so the public tables support statements about our own pipeline rather than about the state of the art.

Scope. Our conclusions cover two-dimensional, five-particle, single-law, time-invariant systems with known unit masses, observed through positions, plus three fixed public force/acceleration-regression tasks. Long-range extrapolation of the learned law, multi-body or non-pairwise interactions, time-varying graphs, unknown masses, and three-dimensional generated systems are not covered. The four-model comparison of Table 1 compares EL-KAN with three fixed-protocol position-only adaptations of relational-inference methods on the generated settings; it is not a claim about their original published benchmarks or about learned-dynamics and interatomic-potential models on the public tasks.

E Additional figures and tables

Table 4 gives the per-setting horizon detail behind Table 1; Table 5 adds the per-setting mechanism, OOD and intervention metrics; Table 6 isolates the discretization-consistent refit; Table 7 gives the full dictionary sweep; and Table 2 reports the public benchmarks summarised in Section 4.2. Figure 3 gives the multi-horizon rollout profiles behind Section 4.4.

Table 2: Public benchmark results (mean $\pm$ standard deviation over three seeds). NMSE is normalized target MSE: force for rMD17/MD22 and estimated acceleration for NRI. Top-k metrics take the k highest scoring edges with k the true edge count. Molecular graph metrics use the covalent-radius proxy, and NRI uses per-sample labels.

Dataset	NMSE	F1	Recall	AUROC	Top-k F1	Edges
rMD17 / Aspirin	$0.0336_{\pm 0.0009}$	$0.218_{\pm 0.029}$	1.000	$0.973_{\pm 0.014}$	$0.651_{\pm 0.073}$	$174_{\pm 24}$
MD22 / DHA	$0.0693_{\pm 0.0012}$	$0.179_{\pm 0.000}$	1.000	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$561_{\pm 0}$
NRI / Springs-5	$0.1090_{\pm 0.0056}$	$0.780_{\pm 0.003}$	$0.900_{\pm 0.005}$	$0.899_{\pm 0.001}$	$0.778_{\pm 0.001}$	$6.6_{\pm 0.01}$

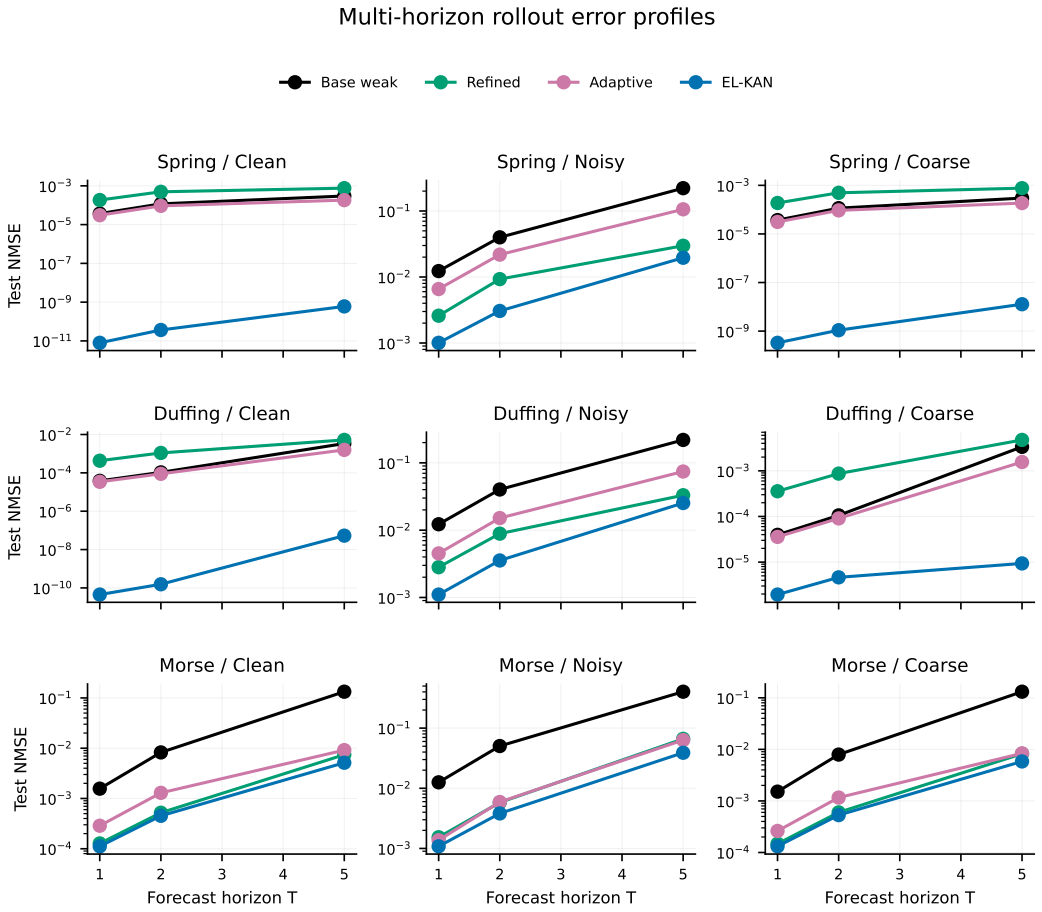

Figure 3: Multi-horizon rollout error ($T = 1, 2, 5$, log scale), one panel per generated setting; legend names map to the weak-form/base control ("Base weak"), the anchor control ("Refined"), the adaptive candidate ("Adaptive") and EL-KAN.

Table 3: The twelve datasets, their role in the study, and the difficulty of their observation regime.

Dataset	Type	What it probes	Difficulty
rMD17 / Aspirin	public	real molecular force field; dense candidate graph	med.–high
MD22 / DHA	public	many nodes, long-range and multi-scale interactions	high
NRI / Springs-5	public	per-sample sparse interaction graph	high
Spring clean	generated	linear law, simple structure; end-to-end sanity check	low
Spring noisy	generated	noise against graph and law estimation	medium
Spring coarse	generated	low sampling rate, less informative dynamics	medium
Duffing clean	generated	nonlinear stiffening, still-regular structure	medium
Duffing noisy	generated	nonlinearity plus noise; long rollouts degrade	med.–high
Duffing coarse	generated	latent velocity hardest to recover	high
Morse clean	generated	non-polynomial law with strong local curvature	med.–high
Morse noisy	generated	nonlinear law plus noise; hardest generated case	high
Morse coarse	generated	nonlinear law plus coarse sampling	high

Table 4: Per-setting test position NMSE at $T = 1, 2, 5$ (mean over five seeds). Weak is the weak-form control, Anchor is the frozen anchor control, and EL-KAN is the final selected model.

		Weak			Anchor			EL-KAN		
Law	Scen.	T=1	T=2	T=5	T=1	T=2	T=5	T=1	T=2	T=5
Spring	clean	3.7e-5	1.2e-4	3.0e-4	1.9e-4	4.9e-4	7.6e-4	**8.0e-12**	**3.7e-11**	**6.0e-10**
Spring	noisy	1.2e-2	4.0e-2	2.2e-1	2.6e-3	9.3e-3	3.0e-2	**1.0e-3**	**3.1e-3**	**2.0e-2**
Spring	coarse	3.7e-5	1.15e-4	3.0e-4	1.9e-4	4.9e-4	7.7e-4	**3.3e-10**	**1.1e-9**	**1.3e-8**
Duffing	clean	3.8e-5	1.1e-4	3.4e-3	4.3e-4	1.1e-3	5.2e-3	**4.5e-11**	**1.55e-10**	**5.3e-8**
Duffing	noisy	1.2e-2	4.0e-2	2.2e-1	2.8e-3	8.9e-3	3.3e-2	**1.1e-3**	**3.5e-3**	**2.5e-2**
Duffing	coarse	4.0e-5	1.1e-4	3.4e-3	3.6e-4	8.7e-4	4.8e-3	**1.9e-6**	**4.6e-6**	**9.4e-6**
Morse	clean	1.6e-3	8.2e-3	1.3e-1	1.3e-4	5.2e-4	7.4e-3	**1.1e-4**	**4.5e-4**	**5.1e-3**
Morse	noisy	1.3e-2	5.0e-2	4.0e-1	1.5e-3	5.8e-3	6.6e-2	**1.1e-3**	**3.8e-3**	**3.9e-2**
Morse	coarse	1.5e-3	7.9e-3	1.3e-1	1.5e-4	6.0e-4	8.2e-3	**1.3e-4**	**5.3e-4**	**5.8e-3**

Table 5: Per-setting mechanism, OOD ($1.8\times$ initial conditions) and single-edge intervention metrics at $T = 5$, mean over five seeds. Law L_2 is reported for the weak control, anchor control, and final EL-KAN; F1, exact-graph count, OOD and intervention NMSE are for EL-KAN.

		Law L_2						
Law	Scen.	Weak	Anchor	EL-KAN	F1	Exact	OOD	Interv.
Spring	clean	6.9e-4	2.8e-4	**6.5e-6**	1.000	5/5	4.2e-9	3.8e-10
Spring	noisy	3.7e-2	3.7e-2	**2.4e-2**	0.967	4/5	0.850	2.3e-2
Spring	coarse	9.0e-4	8.0e-4	**4.7e-5**	1.000	5/5	1.1e-7	1.1e-8
Duffing	clean	1.1e-3	3.9e-4	**3.0e-5**	1.000	5/5	1.2e-7	1.0e-9
Duffing	noisy	5.5e-2	5.5e-2	**3.8e-2**	0.953	3/5	0.736	2.0e-2
Duffing	coarse	1.2e-3	5.3e-4	**2.8e-4**	1.000	5/5	0.361	9.8e-6
Morse	clean	0.404	9.1e-3	**4.4e-3**	1.000	5/5	2.6e-4	4.3e-3
Morse	noisy	0.405	0.274	**0.200**	1.000	5/5	4.6e-2	3.5e-2
Morse	coarse	0.402	9.7e-3	**4.7e-3**	1.000	5/5	3.0e-4	5.0e-3

Table 6: Discretization-consistent refit as a separate candidate: $T = 5$ test NMSE and relative law L_2 before ($\rightarrow$) and after the corrected discrete equation, for the global (G) and refined (R) potentials, mean over five seeds.

		$T = 5$ NMSE		Law L_2	
Law	Scen.	G	R	G	R
Spring	clean	4.15e-4→2.1e-3	7.6e-4→3.6e-3	6.9e-4→2.8e-4	1.9e-4→1.1e-4
Spring	noisy	3.0e-2→3.0e-2	3.2e-2→3.2e-2	3.7e-2→3.7e-2	0.141→0.141
Spring	coarse	4.2e-4→2.1e-3	7.7e-4→3.6e-3	9.0e-4→3.7e-4	6.1e-4→2.2e-4
Duffing	clean	2.1e-3→3.6e-3	3.1e-3→4.2e-3	1.1e-3→5.5e-4	2.4e-4→1.5e-4
Duffing	noisy	3.3e-2→3.3e-2	3.6e-2→3.6e-2	5.5e-2→5.7e-2	0.206→0.206
Duffing	coarse	2.0e-3→3.5e-3	2.7e-3→4.2e-3	1.2e-3→5.9e-4	4.0e-4→1.9e-4
Morse	clean	0.143→0.115	7.3e-3→7.5e-3	0.404→0.273	9.1e-3→5.6e-3
Morse	noisy	0.109→0.0879	6.3e-2→6.2e-2	0.405→0.280	0.265→0.265
Morse	coarse	0.148→0.128	8.1e-3→8.1e-3	0.402→0.272	9.75e-3→4.9e-3

Table 7: Morse dictionary ablation: $T = 5$ test NMSE and relative law L_2 of the selected model per slope set, averaged over the polynomial counts $P = 0, 2, 4$ (the three counts agree to six significant digits except for the widest slope set), with the number of runs, out of fifteen, in which a dictionary candidate was selected.

Scenario	Slope set S	selected	Test NMSE	Law L_2
clean	$\{0.5, 1, 2, 4\}$	1/15	0.00780	0.00954
clean	$\{0.75, 1.5, 3\}$	9/15	**0.00514**	0.00442
clean	$\{1.5\}$	9/15	**0.00514**	**0.00441**
noisy	$\{0.5, 1, 2, 4\}$	9/15	**0.03531**	0.15254
noisy	$\{0.75, 1.5, 3\}$	7/15	0.03808	0.18056
noisy	$\{1.5\}$	7/15	0.03806	**0.17273**
coarse	$\{0.5, 1, 2, 4\}$	2/15	0.00807	0.01068
coarse	$\{0.75, 1.5, 3\}$	9/15	**0.00584**	0.00469
coarse	$\{1.5\}$	9/15	**0.00584**	**0.00467**

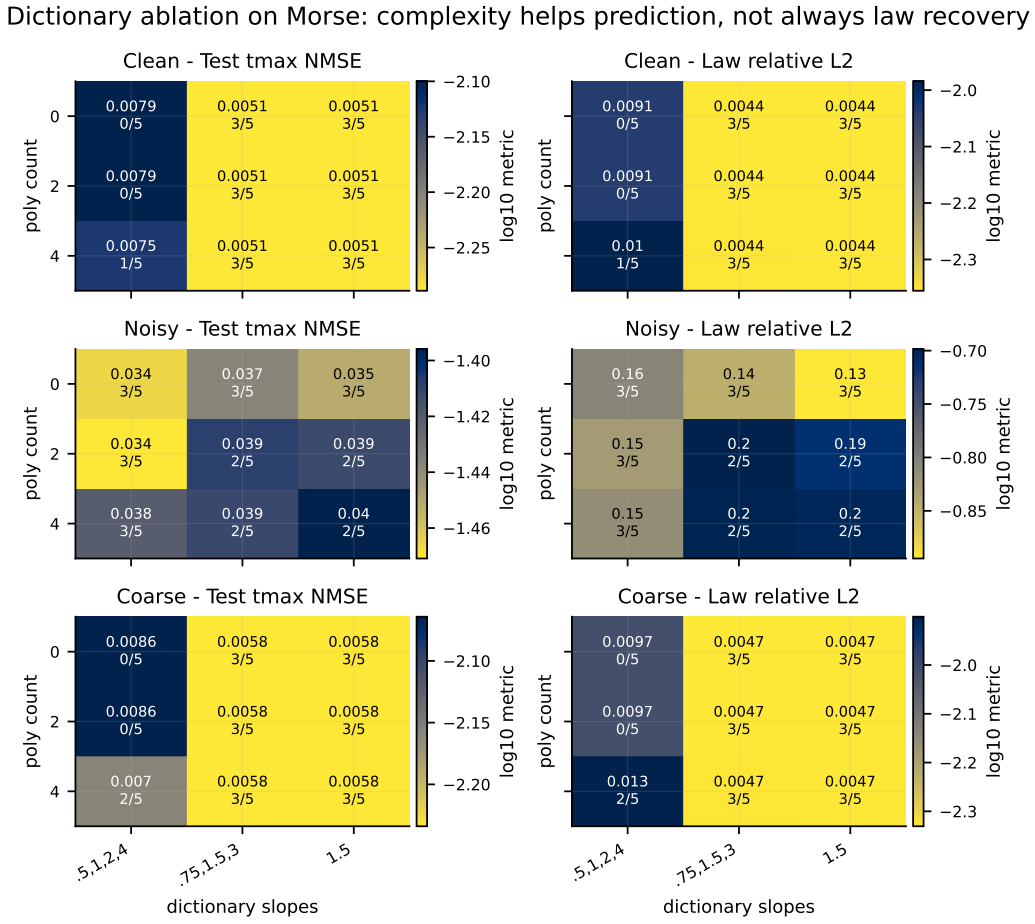

Figure 4: Morse dictionary ablation. Rows are slope sets and columns are polynomial counts; cells report $T = 5$ test NMSE (left) and relative edge-law error (right). The best cell for prediction is not the best cell for the law.

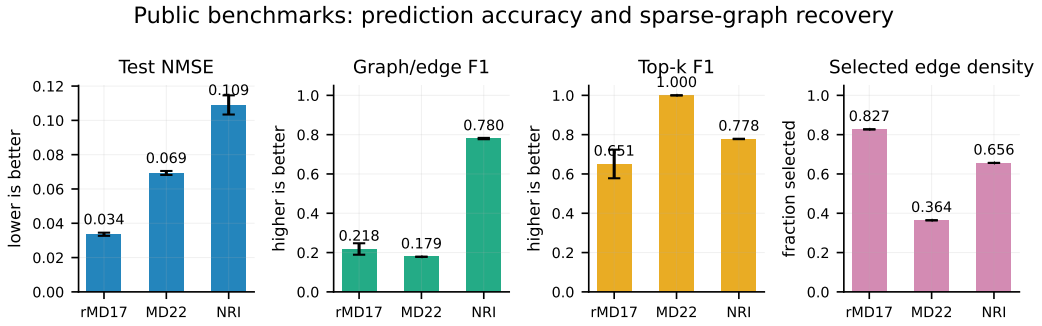

Figure 5: Public benchmark summary: test NMSE, threshold-selected F1, top-k F1 at the true edge count, and the density of the selected graph.

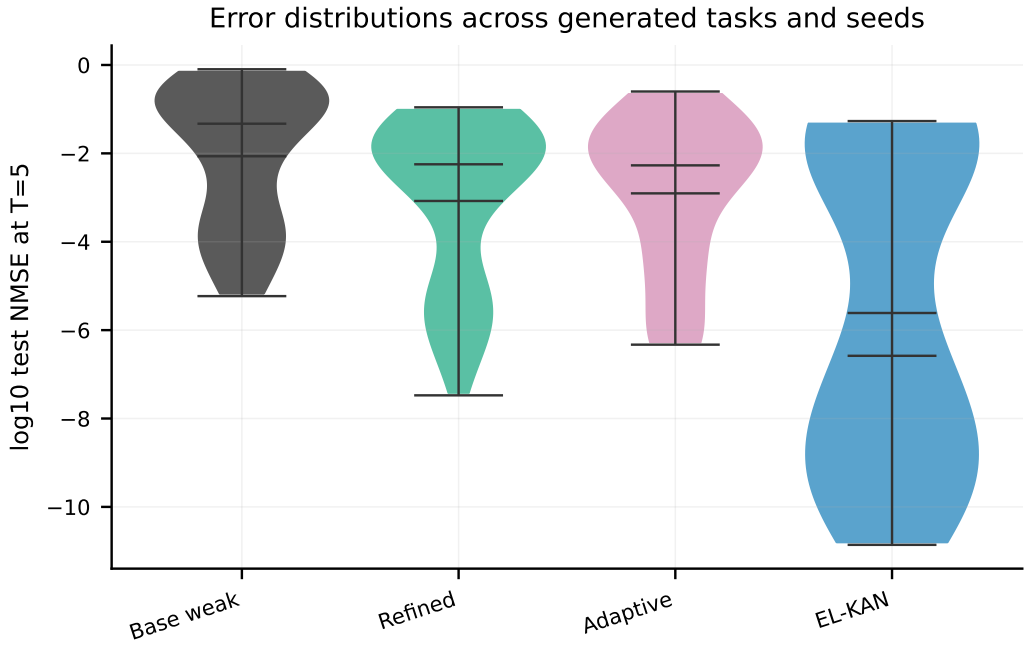

Figure 6: Distribution of $\log_{10}$ test NMSE at $T = 5$ over all generated tasks and seeds for the compared families.