PRISIFT: PRIOR-FITTED STRUCTURAL INFERENCE FROM TRAJECTORIES FOR SPARSE ODE DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse ODE discovery requires identifying a compact governing structure from an exponentially large space of candidate supports. Existing methods solve this structural selection problem independently for each new system, while exhaustive evaluation becomes increasingly costly as the candidate library grows. We introduce PriSIFT, a prior-fitted framework that learns transferable structural regularities from synthetic ODE systems and reuses them to guide discovery on unseen dynamics. Given observed trajectories, a Set-Time Transformer produces term-level structural scores in a single forward pass; these scores define a compact set of plausible support proposals, which are then fitted and verified using task-local extended-BIC rather than a learned final selector. This separation allows cross-system structural knowledge to be amortized while keeping the final scientific decision grounded in the observed data. On a 20-system external literature benchmark held out from method development, PriSIFT achieves the strongest dimension-averaged macro support F1 among evaluated methods (0.678). In matched compute experiments on the fresh holdout, it evaluates 154× fewer candidate supports and achieves a 69× measured wall-clock speedup over exhaustive EBIC. These results suggest that prior-data fitting can provide transferable structural guidance for efficient and reliable equation discovery.

1 INTRODUCTION

Sparse ordinary differential equation (ODE) discovery seeks to recover compact governing structures from observed trajectories. Given a library of p candidate terms, however, each state equation may correspond to one of up to 2^p possible supports. Classical approaches such as SINDy and its extensions (Brunton et al., 2016; Kaiser et al., 2018; Zheng et al., 2018; Fasel et al., 2021; Hirsh et al., 2021; Bertsimas & Gurnee, 2023) therefore solve a new sparse regression or subset-selection problem for every observed system. Lightweight heuristics are efficient but may rank the wrong structures, whereas exhaustive or exact procedures become increasingly expensive as the candidate library grows and may evaluate many redundant, collinear, or numerically unstable supports.

This repeated per-system search creates an opportunity for *amortization*. Machine learning increasingly learns reusable inference rules across distributions of related problems, as illustrated by simulation-based inference and prior-data fitted networks (PFNs) (Müller et al., 2022; Hollmann et al., 2023; Cranmer et al., 2020). In equation discovery, learned predictors such as ODEFormer (d'Ascoli et al., 2023) similarly reuse information across systems, but directly produce the final structural prediction. Classical and learned approaches therefore emphasize different sources of information: the former rely on task-local evidence but do not reuse cross-system structural regularities, whereas the latter amortize structural inference but place the final prediction inside the learned model. This raises a natural question: can cross-system structural knowledge guide *where to search* while leaving the final scientific decision to evidence from the current system?

Our central insight is that the transferable component is the *structural proposal*, not necessarily the final scientific decision. We introduce **PriSIFT** (**Pri**or-Fitted **S**tructural **I**nference **F**rom **T**rajectories), which learns reusable trajectory–structure regularities entirely from synthetic sparse ODE systems. Given trajectories from an unseen system, a Set-Time Transformer produces term-level structural scores in a single forward pass. In *PriSIFT-Search*, these scores define a compact set of plausible support hypotheses; each candidate is then fitted using only the current system, and a

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.

fixed EBIC criterion (Schwarz, 1978; Chen & Chen, 2008) selects the final support. *PriSIFT-Fast* optionally decodes the structural scores directly when minimal inference cost is preferred.

This proposal–selection decomposition uses learning to determine *what is worth considering* and task-local statistics to determine *what should be selected*. The learned proposal can therefore reduce combinatorial search while also suppressing unstable or structurally implausible supports before final selection. In matched fresh-holdout controls, PriSIFT-Search evaluates approximately $154\times$ fewer candidate supports than exhaustive EBIC and achieves a measured $69\times$ wall-clock speedup, with no detectable degradation in structural recovery. On the 20-system external literature benchmark, PriSIFT-Search attains a dimension-averaged macro support F1 of 0.678, compared with 0.666 for Full-EBIC and 0.649 for MIOSR-exact. The recovered equations also exhibit substantially lower coefficient and unseen-state vector-field error than exhaustive selection, showing that the learned proposal can improve the quality of the selected dynamics as well as the efficiency of discovery.

Our contributions are:

1. **Transferable structural proposal.** We show that prior-data fitting can learn reusable structural proposals from synthetic ODEs and transfer them to unseen systems.
2. **Proposal–selection decomposition.** PriSIFT separates cross-system structural inference from task-local model selection, combining a learned proposal with a fixed statistical selector.
3. **Structural regularization effect.** We show that restricting selection to the learned proposal region can improve equation quality as well as efficiency by excluding unstable and structurally implausible supports.
4. **External validation.** Across external benchmarks and varied data and noise regimes, PriSIFT achieves strong structural recovery at a fraction of exhaustive search cost while remaining competitive in coefficient, vector-field, and trajectory fidelity.

2 Related Work

Sparse equation discovery. This line of work recovers governing equations by regressing observed dynamics onto a candidate library and enforcing sparsity. SINDy (Brunton et al., 2016) introduced the template with sequentially thresholded least squares (Kaiser et al., 2018); SR3 (Zheng et al., 2018) replaced hard thresholding with a relaxed regularizer; E-SINDy (Fasel et al., 2021) stabilized selection across bootstrap ensembles; Bayesian-SINDy (Hirsh et al., 2021) placed sparsifying priors on coefficients; and MIOSR (Bertsimas & Gurnee, 2023) formulated support recovery as mixed-integer optimization, the strongest classical baseline in our evaluation. These methods share an assumption: every observed system is a new inference problem, so both the regression and the per-system model selection (BIC/EBIC (Schwarz, 1978; Chen & Chen, 2008), cross-validation) are solved from scratch. They have made single-system discovery practical at small libraries, but they leave untapped the structural knowledge that accumulates across systems, and their cost tracks the combinatorial size of the library. PriSIFT departs from this assumption by amortizing the search itself: structure learned across synthetic systems is reused as a proposal for every new system, while final selection remains a per-system statistical test.

Weak-form and noise-robust discovery. A complementary line reduces sensitivity to numerical differentiation through integral (weak-form) evidence (Messenger & Bortz, 2021; Rudy et al., 2017; Schaeffer & McCalla, 2017), which we adopt in our robustness analysis. These methods improve the *evidence* available to a per-system selector without amortizing anything across systems; in our framework the two compose naturally, since either evidence form can drive the task-local selection of a prior-fitted proposal.

Neural and symbolic equation discovery. A second family learns or searches expressions directly: ODEFormer (d'Ascoli et al., 2023) pretrained a Transformer to generate symbolic ODEs autoregressively; FIM-ODE (Mauel et al., 2026) trains a neural field and projects it onto a library; symbolic-regression methods search expression trees via policy gradients, physics-inspired decomposition, or evolutionary search (Petersen et al., 2021; Udrescu & Tegmark, 2020; Cranmer, 2023; Kusner et al., 2017). These approaches assume the learned output is the deliverable and evaluate

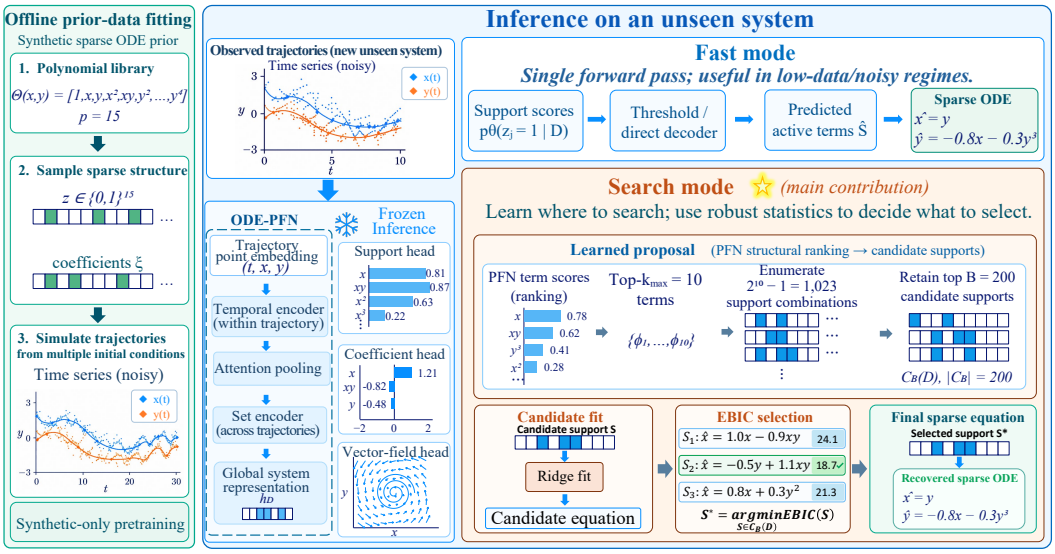

Figure 1: **PriSIFT framework.** Offline prior-data fitting learns transferable structural proposals from synthetic ODEs. At inference, a single forward pass supports either direct decoding in Fast mode or proposal-guided candidate fitting and EBIC selection in Search mode.

whole expressions rather than supports, providing no reusable, composable notion of candidate quality: a generator that misses one term fails outright rather than directing a search toward promising supports. Our setting differs in the unit of prediction (term-level support scores, not full expressions) and in inference (learned scores compose with classical selection instead of replacing it).

Prior-data fitted networks and scientific PFNs. PFNs (Müller et al., 2022) train a Transformer over synthetic Bayesian inference instances so that one forward pass approximates the posterior for a new instance; TabPFN (Hollmann et al., 2023) extended this to tabular classification, within the broader simulation-based and amortized inference literature (Cranmer et al., 2020). Recent scientific PFNs amortize *forward* problems—FIM-ODE (Mauel et al., 2026) pretrains across ODE systems to predict solutions, and PDE-PFN (Park et al., 2026) applies prior-data fitting to forward PDE solution prediction. PriSIFT addresses the complementary *inverse* structural question and changes the amortization target: the learned model approximates a distribution over *combinatorial structures* rather than parameters, and our transfer experiments show it is most robustly deployed as a *proposal mechanism* coupled to a classical selector rather than as a standalone predictor.

3 PRISIFT

3.1 OVERVIEW

PriSIFT (**Pri**or-Fitted **S**tructural **I**nference **F**rom **T**rajectories) separates sparse ODE discovery into two stages that exploit different sources of information: *cross-system structural proposal* and *task-local model selection*. Figure 1 summarizes the framework. Offline, PriSIFT is trained only on synthetically generated sparse ODE systems sampled from a structural prior. Because the generating equations are known, each synthetic task provides supervision for term inclusion, coefficients, and the induced vector field. A Set-Time Transformer learns to map a set of observed trajectories to term-level structural scores that capture recurring relationships between dynamics and governing structure.

At inference time, the pretrained network is frozen and applied to an unseen system in a single forward pass. PriSIFT supports two operating modes. *PriSIFT-Fast* directly decodes the structural scores into a sparse support, providing fully amortized inference. Our primary mode, *PriSIFT-Search*, instead treats these scores as a transferable proposal: it uses them to construct a compact set of plausible supports, fits each proposed equation using only the current system, and applies a fixed

information criterion to select the final model. Thus, learning is used to determine *what is worth considering*, while task-local statistical evidence determines *what should be selected*.

Formally, consider an autonomous system $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$ with $\mathbf{x}(t) \in \mathbb{R}^d$, observed through a collection of possibly noisy trajectories $\mathcal{D} = \{\mathcal{T}_m\}_{m=1}^M$. Given a candidate function library $\Theta(\mathbf{x}) = [\theta_1(\mathbf{x}), \ldots, \theta_p(\mathbf{x})]$, the r-th state equation is represented as $\dot{x}_r = \Theta(\mathbf{x})\,\boldsymbol{\xi}_r$, with sparse coefficient vector $\boldsymbol{\xi}_r \in \mathbb{R}^p$ and support $S_r = \{j : \xi_{rj} \neq 0\}$. Sparse ODE discovery therefore requires identifying a small support S_r from the combinatorial space of possible library subsets. PriSIFT amortizes the structural information used to construct this candidate space, rather than amortizing the final scientific decision.

3.2 Offline Prior-Data Fitting

PriSIFT is trained from a distribution of synthetic sparse ODE tasks rather than from a fixed collection of real dynamical systems. The prior specifies how governing structures, coefficients, initial conditions, trajectories, and observation noise are generated. For each synthetic task, we first sample a binary support $\mathbf{z} \sim p_{\text{struct}}(\mathbf{z})$, then sample coefficients conditional on that structure, $\boldsymbol{\xi} \sim p_{\text{coeff}}(\boldsymbol{\xi} \mid \mathbf{z})$, and construct the corresponding sparse vector field $\mathbf{f}_{\mathbf{z},\boldsymbol{\xi}}(\mathbf{x}) = \Theta(\mathbf{x})\boldsymbol{\Xi}$. Multiple initial conditions are sampled and numerically integrated to produce a set of trajectories $\mathcal{D} = \{(t_{mi}, \mathbf{x}_{mi})\}_{m,i}$, with optional observation perturbations drawn from the training prior. The resulting task distribution exposes the model to many different combinations of support patterns, coefficient scales, initial conditions, and trajectory geometries.

Unlike supervised learning on real benchmark systems, the complete generating mechanism is available for every synthetic task. PriSIFT therefore receives supervision at three complementary levels: *structure*, through the binary support $\mathbf{z}$; *parameters*, through the coefficients $\boldsymbol{\xi}$; and *dynamics*, through vector-field values $\mathbf{f}(\mathbf{x}_q)$ at sampled query states $\mathbf{x}_q$. These signals are combined through

$$\mathcal{L} = \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} + \lambda_{\text{coef}}\mathcal{L}_{\text{coef}} + \lambda_{\text{vf}}\mathcal{L}_{\text{vf}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{1}$$

where $\mathcal{L}_{\text{sup}}$ trains term inclusion, $\mathcal{L}_{\text{coef}}$ supervises coefficient prediction, $\mathcal{L}_{\text{vf}}$ supervises the induced vector field, and $\mathcal{L}_{\text{con}}$ encourages consistency between the predicted coefficients and vector field:

$$\mathcal{L}_{\text{con}} = \mathbb{E}_{\mathbf{x}_q}\left[\left\|\widehat{\mathbf{f}}(\mathbf{x}_q) - \Theta(\mathbf{x}_q)\widehat{\boldsymbol{\Xi}}\right\|_2^2\right]. \tag{2}$$

This multi-level supervision is important because structural discovery is not identifiable from term labels alone: two candidate structures may have similar inclusion patterns yet imply very different coefficient scales or off-trajectory dynamics. The auxiliary objectives encourage the shared representation to encode structure that is dynamically meaningful, while the support head remains the component used to construct proposals at test time. Exact prior distributions, loss weights, and optimization details are reported in Appendix A.1.

3.3 Set-Time Transformer

The input to PriSIFT is not a single regularly sampled sequence but a *set of trajectories*, potentially observed at different times and from different initial conditions. The architecture therefore factorizes representation learning along the two natural axes of the data: *within-trajectory temporal dynamics* and *across-trajectory system structure*.

For a point $(t_{mi}, \mathbf{x}_{mi})$ from trajectory m, PriSIFT first constructs a point embedding $\mathbf{e}_{mi} = \phi_{\text{point}}(t_{mi}, \mathbf{x}_{mi}, \mathbf{m}_{mi})$, where $\mathbf{m}_{mi}$ optionally denotes observation or padding masks. Time and state information are embedded separately and combined into a shared latent representation.

A temporal Transformer encoder then processes the points within each trajectory, $\{\mathbf{e}_{mi}\}_i \mapsto \{\mathbf{h}_{mi}^{\text{time}}\}_i$, allowing the model to capture local evolution, temporal dependencies, and trajectory geometry. Attention pooling compresses each trajectory into a single representation, $\mathbf{u}_m = \text{AttnPool}(\{\mathbf{h}_{mi}^{\text{time}}\}_i)$.

The resulting trajectory representations form an unordered set. A second Transformer therefore operates across trajectories, $\{\mathbf{u}_m\}_{m=1}^M \mapsto \{\mathbf{h}_m^{\text{set}}\}_{m=1}^M$, followed by pooling to obtain a global system representation $\mathbf{h}_{\mathcal{D}} = \text{Pool}(\{\mathbf{h}_m^{\text{set}}\}_{m=1}^M)$. This hierarchy makes the representation sensitive to temporal dynamics while remaining invariant to the order in which trajectories are presented.

Three prediction heads share $\mathbf{h}_{\mathcal{D}}$. The *support head* produces logits a_{rj} for each state–library-term pair and corresponding structural inclusion scores $p_{rj} = \sigma(a_{rj}) = p_\theta(z_{rj} = 1 \mid \mathcal{D})$. The *coefficient head* predicts the associated sparse coefficients, while the *vector-field head* predicts $\mathbf{f}(\mathbf{x}_q)$ at query states through query-conditioned decoding. The latter two heads provide richer dynamical supervision during prior fitting; the support scores are the central representation used for structural proposal at inference time.

We refer to p_{rj} as *structural inclusion scores*. They are probabilistic outputs learned from the synthetic task distribution, but we do not require them to coincide with an exact Bayesian posterior for the observed system. Their primary role is to provide a transferable ordering of candidate terms.

3.4 Inference on an Unseen System

Given trajectories $\mathcal{D}^\star$ from a previously unseen system, the pretrained PriSIFT network is frozen. A single forward pass produces the structural scores $\mathbf{P}^\star = f_\theta(\mathcal{D}^\star)$, after which the same representation can be used in either a fully amortized mode or a proposal-guided selection mode.

3.4.1 Fast Mode

PriSIFT-Fast directly converts the structural inclusion scores into a support estimate. For each state equation, $\widehat{z}_{rj} = \mathbb{I}[p_{rj} > \tau]$, where τ is a fixed decoding threshold, yielding $\widehat{S}_r^{\text{fast}} = \{j : \widehat{z}_{rj} = 1\}$. When coefficient estimates are required, the selected support can be paired with the corresponding coefficient-head predictions to form a sparse equation.

Fast mode performs no candidate fitting or combinatorial model selection after the network evaluation. It is therefore the fully amortized form of PriSIFT: the structural decision is made directly from the knowledge encoded during prior-data fitting. This mode is useful when inference cost is the primary constraint or when the learned structural prior itself provides sufficient evidence.

3.4.2 Search Mode

PriSIFT-Search uses the same forward pass differently. Rather than treating the structural scores as final decisions, it interprets them as a proposal over the support space and delegates final selection to the observed system.

For numerical stability and additive support scoring, the term-level probabilities are first converted to log-odds $r_{rj} = \log \frac{p_{rj}}{1-p_{rj}}$. The highest-ranked terms define a reduced term set $T_r = \mathrm{TopK}(\{r_{rj}\}_{j=1}^{p}, k_{\max})$. For a candidate support $S \subseteq T_r$, PriSIFT assigns the structural proposal score $R_\theta(S \mid \mathcal{D}) = \sum_{j \in S} r_{rj}$. The candidate pool is then defined by retaining the highest-scoring supports,

$$\mathcal{C}_{B,r}(\mathcal{D}) = \mathrm{TopB}_{S \subseteq T_r,\ S \neq \emptyset}\, R_\theta(S \mid \mathcal{D}), \qquad |\mathcal{C}_{B,r}| \leq B. \tag{3}$$

Thus, the network does not perform an iterative combinatorial search. A single neural forward pass instead induces an ordering over terms and supports, which concentrates subsequent evaluation on a small region of the hypothesis space.

Candidate fitting. Each proposed support is evaluated using only evidence from the current system. Let $\mathbf{y}_r$ denote the task-local regression target and $\mathbf{\Phi}_S$ the library design matrix restricted to support S. The coefficients are fitted by ridge regression,

$$\widehat{\boldsymbol{\xi}}_{r,S} = \arg\min_{\boldsymbol{\xi}} \|\mathbf{y}_r - \mathbf{\Phi}_S \boldsymbol{\xi}\|_2^2 + \alpha \|\boldsymbol{\xi}\|_2^2. \tag{4}$$

This step deliberately re-estimates coefficients from the unseen system instead of relying on the coefficient head learned from the prior.

Task-local selection. For each fitted candidate, we compute an extended Bayesian information criterion

$$\mathrm{EBIC}(S) = n \log\left(\frac{\mathrm{RSS}_S}{n}\right) + |S| \log n + 2\gamma |S| \log p, \tag{5}$$

where RSS_S is the candidate residual sum of squares, n is the number of task-local observations used for fitting, and the final term penalizes structural complexity. The discovered support is $\widehat{S}_r = \arg\min_{S \in \mathcal{C}_{B,r}(\mathcal{D})} \mathrm{EBIC}(S)$, and the final equation is $\widehat{f}_r(\mathbf{x}) = \Theta_{\widehat{S}_r}(\mathbf{x})\widehat{\boldsymbol{\xi}}_{r,\widehat{S}_r}$.

The distinction between proposal and selection is fundamental. PriSIFT uses cross-system experience only to determine which structures merit evaluation; coefficients and the final support are recomputed from task-local evidence. Consequently, the learned model need not encode a corpus-specific rule for deciding which fitted candidate is best.

The computational consequence follows directly from this decomposition. Exhaustive model selection requires fitting a number of supports that grows exponentially with the library size, whereas PriSIFT-Search limits expensive task-local fitting and model selection to at most B proposed candidates per equation. Candidate generation itself uses only the structural scores from a single frozen-network evaluation. PriSIFT therefore amortizes the expensive identification of promising regions of the hypothesis space while retaining an explicit statistical decision at the level of the individual system.

4 Experiments

We organize the evidence around five research questions: **RQ1**, does PriSIFT learn a stronger structural proposal than existing heuristics (Section 4.2); **RQ2**, can the learned proposal preserve exhaustive-level discovery while substantially reducing combinatorial cost (Section 4.3); **RQ3**, does improved structural recovery translate into better discovered equations end to end (Section 4.4); **RQ4**, why can restricting the hypothesis space improve final equation quality (Section 4.5); and **RQ5**, does the advantage persist under data scarcity, noise, and distribution shift (Section 4.6).

4.1 Experimental Setup

Corpora. Method development uses an 80-system corpus drawn from three public benchmark families—Strogatz (Strogatz, 2018), MDBench (Ziaei Bideh et al., 2025), and ODEBench (Gilpin, 2021). We evaluate transfer on two external corpora held out from method development. A 21-system fresh holdout, assembled from ODEbase/BioModels-derived models (Lüders et al., 2022) and canonical literature systems, is used for matched-budget proposal and same-selector controls. A separate 20-system literature corpus curated from predefined public sources serves as the primary confirmatory benchmark. Dataset construction, provenance, overlap audits, routing rules, and reproducibility details are reported in Appendix A.2.

Baselines. We compare PriSIFT against representative sparse-regression, ensemble, Bayesian, exact-subset, and exhaustive-selection approaches, including SINDy-STLSQ (Kaiser et al., 2018), SR3 (Zheng et al., 2018), E-SINDy (Fasel et al., 2021), Bayesian-SINDy (Hirsh et al., 2021), MIOSR (Bertsimas & Gurnee, 2023), and full-space BIC/EBIC selection. For proposal-specific experiments, classical signals such as ridge magnitude, STLSQ coefficients, ensemble frequency, Bayesian inclusion scores, and random ranking are converted into proposal heuristics under the same candidate budget and final selector. All methods receive matched trajectories, candidate libraries, noise realizations, and evaluation seeds; all tuning and method-selection decisions use development data only.

Evaluation protocol. Our primary structural metric is *dimension-averaged macro support F1*. For each system and evaluation seed, support F1 is computed separately for each state equation and averaged across state dimensions; repeated seeds are then aggregated within system, and systems—rather than individual trajectories or seeds—are the statistical unit for cross-system summaries and paired tests. We additionally evaluate exact support recovery, coefficient error, unseen-state vector-field error, and unseen-initial-condition trajectory error when assessing recovered equations end to end. Statistical comparisons use paired system-level tests with bootstrap confidence intervals and correction for multiple comparisons. Complete metric definitions, frozen hyperparameters, and statistical procedures are provided in Appendix A.

Table 1: **Matched-budget proposal comparison** on the fresh holdout ($B = 200$, shared min-EBIC selector). Higher is better.

Proposal	Ranking signal	Macro F1 ↑
PriSIFT	Learned structural score	**0.511**
Ridge	Coefficient magnitude	0.345
Bayesian-SINDy	Inclusion probability	0.280
Random	Random score	0.277
STLSQ	Coefficient magnitude	0.261
E-SINDy	Selection frequency	0.215

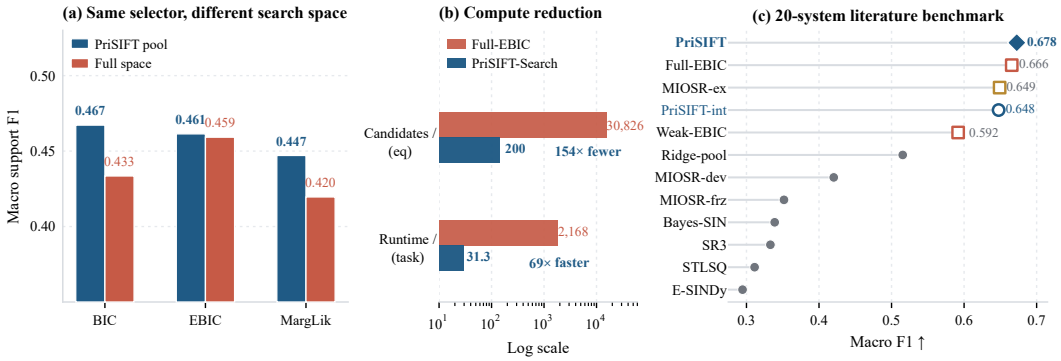

Figure 2: **PriSIFT-Search preserves exhaustive-level structural discovery at substantially lower cost.** (a) Same-selector control: identical BIC, EBIC, and marginal-likelihood selectors are applied to the PriSIFT proposal pool and the full audited support space. (b) Candidate-evaluation and measured runtime reduction relative to exhaustive Full-EBIC. (c) Structural discovery on the 20-system external literature benchmark, ordered by dimension-averaged macro support F1.

4.2 RQ1: Does PriSIFT Learn Better Structural Proposals?

Table 1 isolates the proposal mechanism by holding the trajectories, candidate budget, coefficient fitting, and final min-EBIC selector fixed. PriSIFT therefore differs from the baselines only in the structural signal used to decide which supports merit evaluation. Its advantage suggests that recurring trajectory–structure relationships learned across many synthetic systems provide information that cannot be recovered reliably from a single observed system by coefficient magnitude, inclusion frequency, or other task-local heuristics. Importantly, the learned scores are most useful as a *proposal*: they concentrate a limited candidate budget on plausible regions of the support space, while leaving the final scientific decision to task-local evidence. This controlled comparison directly supports the proposal–selection decomposition underlying PriSIFT.

4.3 RQ2: Can PriSIFT Match Exhaustive Discovery at a Fraction of the Cost?

The central question is whether proposal-based compression removes redundant structure or accidentally removes hypotheses needed for correct recovery. Figure 2a isolates this effect by holding coefficient fitting and the final selector fixed while changing only the candidate space. The restricted pool tracks the full space across three qualitatively different selectors, indicating that PriSIFT is not exploiting a particular information criterion. Instead, its structural ranking concentrates evaluation on the small region of the combinatorial space that contains the decisions relevant to downstream selection. The gain is therefore fundamentally a reduction in *candidate evaluation*, rather than an approximation to a weaker statistical decision rule.

Figure 2c asks whether this compression principle survives beyond the corpus used for method development. On the external literature benchmark, PriSIFT-Search remains at the top of the structural recovery ranking while evaluating only a bounded proposal set, alongside exhaustive and exact best-subset procedures that inspect orders of magnitude more hypotheses. On the primary structural

Table 2: **End-to-end ODE discovery** on the untouched corpus. Support F1 uses the same canonical dimension-averaged macro structural metric as Figure 2c; the remaining columns evaluate complementary properties of each method's native discovered equation. Best value per column in **bold**.

Method	Support F1 ↑	Exact ↑	Coeff. NCE ↓	VF-NRMSE ↓	Traj.-NRMSE ↓
PriSIFT-Search	**0.678**	**0.180**	2.11	**1.25**	**0.804**
Full-EBIC	0.666	0.170	5.88	6.10	0.818
MIOSR-exact	0.649	**0.180**	6.12	6.55	0.824
Bayesian-SINDy	0.339	0.000	**1.86**	3.07	0.868
SR3	0.333	0.000	44.99	60.38	1.385
STLSQ	0.311	0.000	20.03	43.83	1.373
E-SINDy	0.295	0.000	20.86	43.37	1.472

metric, PriSIFT-Search reaches 0.678 macro F1, compared with 0.666 for Full-EBIC and 0.649 for MIOSR-exact, while the comparison with exhaustive selection is statistically indistinguishable. These results show that cross-system structural knowledge can concentrate task-local selection onto a dramatically smaller hypothesis region with no detectable degradation in structural recovery.

Taken together, Figure 2 also characterizes the accuracy–compute trade-off. PriSIFT changes the scaling of inference by changing *which hypotheses are ever evaluated*: the expensive cross-system reasoning is paid once during prior-data fitting, whereas each new system requires one frozen-network evaluation followed by a bounded number of task-local candidate fits. Deployment cost therefore scales with the proposal budget rather than with the full combinatorial support space. This is the practical form of amortization in PriSIFT: reusable structural knowledge converts exhaustive enumeration into a small verification problem, while preserving an explicit statistical decision for the current system.

4.4 RQ3: Does Better Structural Recovery Yield a Better ODE?

Support recovery is an intermediate objective; the scientific deliverable is the recovered dynamical system itself. We therefore evaluate each method's *native final equation* on the untouched corpus using coefficient recovery, exact structural recovery, unseen-state vector-field fidelity, and trajectory prediction from unseen initial conditions. The complete protocol, including numerical integration and failure handling, is given in Appendix A.5.

Table 2 shows that the structural advantage carries through to the recovered dynamics. PriSIFT-Search attains the highest support F1 and substantially lower coefficient and off-trajectory vector-field error than exhaustive selection, while also achieving the lowest trajectory error. The larger gains in continuous dynamical fidelity than in discrete support F1 show that similarly scored supports can induce very different dynamics. A common-refit control (Appendix B.4) preserves this advantage, indicating that it primarily arises from the selected structure rather than the coefficient solver.

4.5 RQ4: Why Can Restriction Improve Discovery?

Figure 3 shows that the learned proposal acts as more than a computational shortcut. Exhaustive selection is exposed to many high-order, correlated supports that fit observed samples yet yield poorly conditioned coefficients and unstable off-trajectory dynamics. By pruning this region before EBIC selection, PriSIFT induces a smaller and structurally more coherent hypothesis class. This effect is strongest for degree-5 libraries, where exhaustive vector-field error increases sharply while PriSIFT remains substantially more stable. Together, these comparisons support a structural-regularization interpretation of the learned proposal.

4.6 RQ5: Robustness under Data Scarcity, Noise, and Distribution Shift

Figure 4 shows that PriSIFT-Search remains effective as trajectories become scarce and observation noise increases, without re-tuning on the untouched corpus. The integral-evidence variant further shows that task-local evidence can be changed without retraining the learned proposal. Larger libraries remain an important scaling challenge because higher-order terms introduce greater redun-

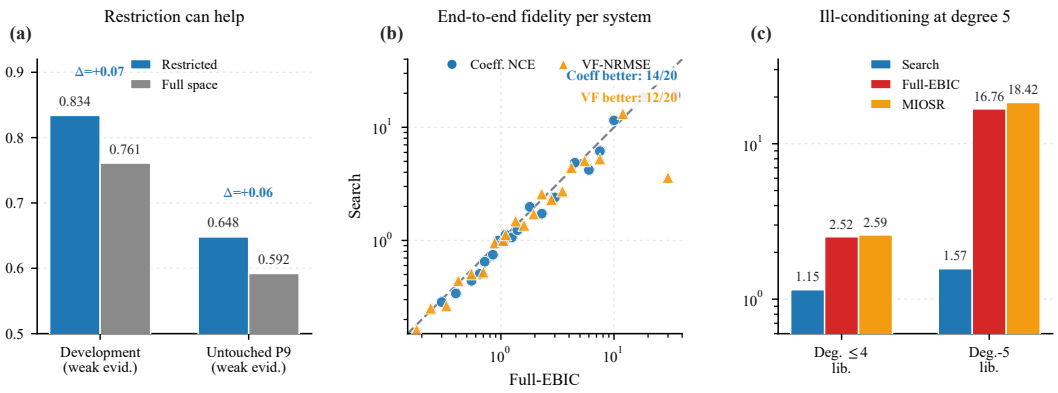

Figure 3: **Evidence for learned structural regularization.** (a) With the selector and evidence fixed, restricting selection to the PriSIFT proposal region improves recovery on both development and untouched corpora. (b) Per-system comparisons against Full-EBIC show that the benefit extends beyond support recovery to coefficient and vector-field fidelity. (c) The regularization effect is most pronounced for degree-5 libraries, where exhaustive methods suffer sharply increased vector-field error while PriSIFT-Search remains substantially more stable.

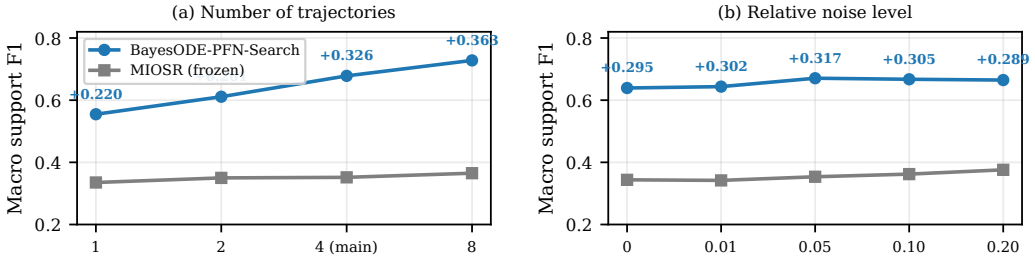

Figure 4: **Robustness across observation regimes.** PriSIFT-Search is evaluated as the number of observed trajectories varies (left) and as relative observation noise increases (right)

dancy and conditioning difficulty. The degree-5 results nevertheless show that restricting the candidate space can substantially stabilize the recovered dynamics when exhaustive selection becomes ill-conditioned.

5 Conclusion

We studied whether the combinatorial search underlying sparse equation discovery can be amortized like prediction. We introduced PriSIFT, a prior-data fitted Set-Time Transformer whose term-level structural scores serve as a learned proposal, coupled with a fixed task-local EBIC selector that keeps the final scientific decision grounded in the observed system. On the 21-system fresh holdout, the learned proposal substantially outperforms classical heuristics under identical candidate budgets; on the untouched 20-system literature corpus, PriSIFT-Search attains the highest dimension-averaged macro support F1 among the evaluated methods. It also requires far fewer candidate evaluations than exhaustive selection while producing equations with substantially lower coefficient and unseen-state vector-field error. Together, these results show that prior-data fitting can amortize *where to search* without replacing task-local scientific model selection, and that restricting the search space can improve not only efficiency but also the quality of the recovered dynamics.

Future work will extend PriSIFT to higher-dimensional systems, richer candidate libraries, and degree-agnostic proposal models, while improving task-local selection in larger and more ill-conditioned search spaces. The same proposal–selection principle may also apply to other scientific discovery problems with combinatorial hypothesis spaces.

AI USE STATEMENT

We used generative AI tools to improve the readability and language of the manuscript. We did not use generative AI tools to develop the research methodology, formulate mathematical claims, design experiments, interpret experimental results, or generate synthetic datasets. All AI-assisted edits were reviewed and verified by the authors. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

REFERENCES

Dimitris Bertsimas and Wes Gurnee. Learning sparse nonlinear dynamics via mixed-integer optimization. *arXiv preprint arXiv:2206.00176*, 2023.

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.

Jiahua Chen and Zehua Chen. Extended and modified likelihoods for model selection. *Statistica Sinica*, 18(2):275–294, 2008.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

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. *arXiv preprint arXiv:2111.10992*, 2021.

William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. *arXiv preprint arXiv:2110.05266*, 2021.

Seth M. Hirsh, David A. Barajas-Solano, and J. Nathan Kutz. Sparsifying priors for Bayesian uncertainty quantification in model discovery. *Royal Society Open Science*, 8(10):211899, 2021.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.

Eurika Kaiser, J. Nathan Kutz, and Steven L. Brunton. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society A*, 474(2219): 20180335, 2018.

Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International Conference on Machine Learning*, 2017.

Christoph Lüders, Thomas Sturm, and Ovidiu Radulescu. ODEbase: A repository of ODE systems for systems biology. *arXiv preprint arXiv:2201.08980*, 2022.

Maximilian Mauel, Johannes R. Hübers, David Berghaus, Patrick Seifner, and Ramses J. Sanchez. Foundation inference models for ordinary differential equations. *arXiv preprint arXiv:2602.08733*, 2026.

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

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022.

Jaehyeon Park, Mingu Kang, Dongseok Lee, Woojin Cho, Kookjin Lee, Anthony Gruber, Youngjoon Hong, and Noseong Park. PDE-PFN: Prior-data fitted neural pde solver. In *Advances in Neural Information Processing Systems*, 2026.

Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Claudio P. Santiago, et al. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations*, 2021.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

Hayden Schaeffer and Scott G. McCalla. Learning differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A*, 473(2197):20160846, 2017.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. CRC Press, 2nd edition, 2018.

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

Peng Zheng, Travis Askham, Steven L. Brunton, J. Nathan Kutz, and Aleksandr Y. Aravkin. A unified framework for sparse relaxed regularized regression: SR3. *IEEE Control Systems Letters*, 2(4):748–753, 2018.

Amirmohammad Ziaei Bideh, Aleksandra Georgievska, and Jonathan Gryak. MDBench: Benchmarking data-driven methods for model discovery. *arXiv preprint arXiv:2509.20529*, 2025.

A Detailed Experimental Protocol

This appendix provides the implementation, corpus, evaluation, and statistical details omitted from the main text. All evaluation-facing choices described below were frozen before the corresponding external evaluation.

A.1 Training Prior and Loss Specification

Synthetic structural prior. PriSIFT is trained entirely on synthetically generated sparse polynomial ODE systems. Synthetic tasks are sampled on the fly, so training does not reuse a fixed finite dataset. In the primary prior, systems are two-dimensional with polynomial degree at most four, corresponding to a 15-term candidate library. Each library term is independently included with probability 0.2, yielding approximately three active terms per equation on average. Active coefficients are sampled from the frozen coefficient prior with scale 2.5.

For each sampled system, four trajectories of 32 observations are generated from independently sampled initial conditions over a time horizon drawn from the prior, with additive Gaussian observation noise $\sigma = 0.02$. We additionally sample 128 query states for vector-field supervision. These choices define the prior-data distribution seen by PriSIFT; no real benchmark system is used as a training example.

Training uses 8,000 optimization steps with batch size 16. Checkpoint selection is based on support F1 on a held-out synthetic proxy set rather than the final training loss. We found during development that continuing to optimize the synthetic objective after structural validation had peaked could reduce transfer to real systems. This observation motivates the use of PriSIFT scores as transferable proposals that are subsequently verified by task-local evidence, rather than as self-sufficient final decisions.

Training objective. The multi-objective loss is

$$\mathcal{L} = 10\,\mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{vf}} + \mathcal{L}_{\text{coef}} + 0.5\,\mathcal{L}_{\text{con}}. \tag{6}$$

The support objective is binary cross-entropy over the state–library-term grid, with positive-class weight 2.0. The coefficient and vector-field objectives are Gaussian negative log-likelihoods with learned predictive variances. The consistency objective couples the two auxiliary heads by penalizing disagreement between direct vector-field prediction and the sparse library reconstruction:

$$\mathcal{L}_{\text{con}} = \frac{1}{N_q} \sum_{q=1}^{N_q} \left\| \widehat{\mathbf{f}}(\mathbf{x}_q) - \Theta(\mathbf{x}_q)\widehat{\boldsymbol{\Xi}} \right\|_2^2. \tag{7}$$

Loss weights, prior parameters, and query counts were fixed during method development and were not changed after the external evaluation corpora were frozen.

A.2 Evaluation Corpora and Provenance

Development corpus. Method development uses 80 real systems drawn from three public benchmark families: Strogatz (Strogatz, 2018), MDBench (Ziaei Bideh et al., 2025), and ODEBench (Gilpin, 2021). These systems were used for method-design decisions and therefore are never reported as untouched test data.

Fresh holdout. The first external corpus contains 21 systems: 14 biomedical systems derived from ODEbase/BioModels (Lüders et al., 2022) and seven canonical literature systems. This corpus supports the matched-budget proposal comparison, same-selector controls, proposal diagnostics, and runtime measurements. Evaluation seeds 421000–421009 were frozen independently of the development seeds.

Untouched literature corpus. The primary confirmatory corpus contains 20 systems curated mechanically from thirteen predefined public source families: Jordan & Smith; Perko; Guckenheimer & Holmes; Kuznetsov; Murray; Nayfeh & Mook; Nagumo et al.; Wiggins; Scholarpedia; Sprott; Keener & Sneyd; Strogatz; and ODEbase. Candidate systems were enumerated in source order and screened only by predeclared criteria: compatibility with the polynomial autonomous setting and numerical stability under the observation protocol. The final manifest was frozen and content-addressed before any discovery inference.

Fifteen systems use candidate libraries of degree at most four and five use degree-five libraries. The declared candidate-library degree is corpus metadata used to construct the method-side library; it does not encode the active support. Evaluation seeds 90001–90005 were frozen with the corpus manifest.

Provenance and overlap audit. A provenance audit of the earlier fresh holdout identified one exact duplicate of a development-corpus system and eight support-pattern overlaps under library symmetries. We therefore report an explicit overlap-sensitivity analysis in Appendix B.7. The untouched literature corpus was constructed mechanically after this audit and remained frozen throughout confirmatory evaluation.

A.3 Frozen Configuration and Degree Routing

All components used during untouched-corpus inference were frozen in advance. The evaluation family contains a degree-four checkpoint `m20_deg4_seed42` and a degree-five specialist checkpoint `m37_p7_3_d5_seed42`. PriSIFT-Search uses the frozen candidate budget, maximum proposal cardinality, ridge regularization, EBIC convention, and preprocessing pipeline.

Routing depends only on the *declared maximum degree of the candidate library*. Systems whose declared library degree is at most four use the degree-four prior-fitted checkpoint; degree-five libraries use the degree-five specialist. No routing decision inspects the active ground-truth support, coefficients, or evaluation outcome.

The degree-five model is used as a specialist matched to the larger candidate library of degree-five systems, while lower-degree systems use the primary degree-four checkpoint.

A.4 Baselines and Comparison Protocol

We evaluate representative methods spanning sparse regression, relaxed optimization, ensemble selection, Bayesian sparsity, exact subset selection, and exhaustive statistical model selection.

The primary comparison set includes SINDy-STLSQ (Kaiser et al., 2018), SR3 (Zheng et al., 2018), E-SINDy (Fasel et al., 2021), Bayesian-SINDy (Hirsh et al., 2021), and MIOSR (Bertsimas & Gurnee, 2023). We additionally evaluate full-space BIC and EBIC selection as exhaustive statistical references and an exact best-subset variant based on the MIOSR formulation.

For matched-budget proposal experiments, existing estimators are converted into *proposal heuristics* rather than being compared through their native final output. Ridge and STLSQ rank terms by fitted coefficient magnitude, E-SINDy by ensemble selection frequency, Bayesian-SINDy by posterior inclusion score, and the random baseline by an independent random ordering. The candidate budget, coefficient fitting procedure, and final selector are then held fixed across all proposal mechanisms.

All methods receive matched trajectories, candidate libraries, noise realizations, and evaluation seeds. Hyperparameter tuning and variant selection are restricted to development data. No baseline threshold or hyperparameter is selected from the untouched corpus.

A.5 Metrics and Aggregation

We distinguish structural recovery from the quality of the final recovered dynamics.

Support recovery. The primary structural metric throughout the paper is *dimension-averaged macro support F1*. For each system and evaluation seed, support F1 is computed separately for each state equation and averaged across state dimensions. Repeated seeds are then aggregated within system before cross-system summaries and paired comparisons. This single canonical protocol is used in Figure 2c, the observation-regime analysis, and the per-system structural results; no alternative support-F1 aggregation is used. Precision, recall, exact support recovery, AUROC, AUPRC, Recall@k, true-support-in-pool rate, and OraclePoolF1 are used as diagnostics where appropriate.

End-to-end equation quality. End-to-end evaluation scores each method's *native final discovered equation* without post-hoc support correction. We report normalized coefficient error (NCE), unseen-state vector-field NRMSE (VF-NRMSE), and trajectory NRMSE from unseen initial conditions.

For VF-NRMSE, each discovered equation is evaluated on 100 unseen states sampled from the ground-truth trajectory tube. For trajectory fidelity, the discovered ODE is integrated from 16 unseen initial conditions per system, with seeds frozen before evaluation. Errors are reported at predeclared short and full horizons. Numerical integration failures receive fixed capped penalties rather than being discarded.

All linear solves used in the final evaluation employ float64 arithmetic. A float32 implementation was found during audit to induce large coefficient errors in exhaustive baselines; the precision correction changes coefficient evaluation but does not alter the selected structural support.

A.6 Statistical Procedures

All confirmatory inference treats the *system* as the statistical unit, with repeated seeds nested within system. Paired comparisons use the two-sided Wilcoxon signed-rank test on per-system paired differences. Confidence intervals are obtained using a hierarchical bootstrap that resamples systems with replacement and then resamples seeds within each selected system; reported intervals use 10,000 bootstrap draws.

When several methods are compared against the same reference within one metric family, p-values are corrected using the Holm–Bonferroni procedure. We distinguish statistical indistinguishability from formal non-inferiority: when the confidence interval includes the predefined non-inferiority margin, we make no non-inferiority claim.

For the end-to-end comparison in Table 2, multiplicity families are defined separately for each metric. The coefficient and vector-field improvements of PriSIFT-Search over the two exhaustive ref-

erences survive Holm correction; trajectory differences are reported as point estimates when not significant. Any inferential comparison involving support F1 uses the same canonical dimension-averaged metric defined in Section A.5; no alternative support-F1 aggregation is used.

A.7 Reproducibility and Frozen Artifacts

The evaluation-facing artifacts are content-addressed and frozen. The degree-four checkpoint is `m20_deg4_seed42` (SHA256 `7fa2b954...`), and the degree-five specialist is `m37_p7_3_d5_seed42` (SHA256 `058997e5...`). The untouched-corpus manifest has SHA256 `3f42ebc2...`; the end-to-end evaluation configuration has SHA256 `6c1d7776...`.

The structural evaluation uses seeds 90001–90005; unseen-initial-condition evaluation uses frozen seeds 92001–92016. Proposal-level and same-selector results on the fresh holdout derive from a separate frozen statistics bundle.

No evaluation-facing checkpoint, proposal budget, preprocessing rule, selector, baseline configuration, or corpus manifest was retrained or reselected after the corresponding external corpus was frozen. Code, configuration files, manifests, and evaluation scripts will be released.

A.8 Runtime Measurement Protocol

Wall-clock results in Figure 2b are medians over the fresh holdout measured on a single NVIDIA A100 using PyTorch 2.6 and CUDA 12.4. Per-task timing includes preprocessing and candidate evaluation. PriSIFT-Search requires 31.3 ms per task under the reported implementation, compared with 2,168 ms for exhaustive EBIC over 30,826 audited supports; the mixed-integer baseline requires approximately 5 s per task.

The theoretical non-empty support space for the primary 15-term library is $2^{15} - 1 = 32{,}767$. The audited exhaustive control restricts support cardinality according to the frozen protocol and evaluates 30,826 candidates. The original PriSIFT-Search configuration evaluates 200 proposed supports per equation; the later frozen evaluation family uses a larger proposal budget where specified. Candidate-count reduction is a property of the algorithmic protocol, whereas wall-clock reduction depends on the implementation and hardware.

B Supplementary Controls and Robustness Analyses

B.1 Same-Selector Controls

Figure 2a holds the fitting procedure and final selector fixed while varying only the candidate space. Table 3 reports the underlying results on the fresh holdout. The PriSIFT-restricted pool and exhaustive support space are statistically indistinguishable under all three selectors; no formal non-inferiority claim is made because the corresponding confidence intervals cross the predefined margin.

Table 3: **Same selector, different candidate space.** Macro support F1 on the fresh holdout. PriSIFT evaluates 200 candidates; the exhaustive control evaluates 30,826.

Selector	PriSIFT pool	Full space	Δ	95% CI	Wilcoxon p
BIC	0.4672	0.4335	+0.034	$[-0.008, +0.075]$	0.126
EBIC	0.4614	0.4592	+0.002	$[-0.040, +0.045]$	0.879
Marg. likelihood	0.4471	0.4196	+0.028	$[-0.008, +0.064]$	0.149

B.2 Observation-Regime Matrix

Table 4 gives the complete numerical results underlying Figure 4. The integral-evidence column uses the same PriSIFT proposal with integral-form task-local evidence; it is a robustness variant of the inference pipeline rather than a separately trained model.

Table 4: **Support recovery across observation regimes.** Canonical dimension-averaged macro support F1 on the untouched corpus. The final column is PriSIFT-Search minus the frozen MIOSR baseline.

Regime	PriSIFT-Search	Integral evidence	MIOSR	Δ
Main	0.678	0.648	0.352	+0.326
1 trajectory	0.555	0.588	0.335	+0.220
2 trajectories	0.611	0.616	0.350	+0.261
8 trajectories	0.728	0.598	0.365	+0.363
$\sigma_{\mathrm{rel}} = 0$	0.639	0.665	0.344	+0.295
$\sigma_{\mathrm{rel}} = 0.01$	0.644	0.668	0.342	+0.302
$\sigma_{\mathrm{rel}} = 0.05$	0.671	0.645	0.354	+0.317
$\sigma_{\mathrm{rel}} = 0.10$	0.667	0.631	0.362	+0.305
$\sigma_{\mathrm{rel}} = 0.20$	0.665	0.587	0.376	+0.289

B.3 Exact-Equation Recovery

Exact equation recovery is considerably stricter than support F1. For a coefficient tolerance τ, an equation is counted as recovered only when all active coefficients fall within τ of their ground-truth values and inactive coefficients remain zero. Table 5 reports recovery rates across several tolerances.

Table 5: **Exact-equation recovery rate** on the untouched corpus.

Tolerance	PriSIFT	MIOSR-ex.	Full-EBIC	MIOSR	MIOSR-tuned	SR3	STLSQ	Bayes-S.
$\tau = 0.1$	0.06	0.06	0.06	0.00	0.02	0.00	0.00	0.00
$\tau = 0.2$	0.14	0.14	0.13	0.01	0.04	0.00	0.00	0.00
$\tau = 0.5$	0.17	0.17	0.16	0.02	0.05	0.00	0.00	0.00

B.4 Common-Refit Control

To separate support quality from coefficient-estimation differences, we refit every method's predicted support with the same ridge estimator. Table 6 shows that the qualitative end-to-end ranking is preserved. In particular, the coefficient advantage of PriSIFT-Search over the exhaustive references remains after the common refit, indicating that the gain primarily arises from the selected support rather than from a different coefficient solver.

Table 6: **Common-refit control.** All predicted supports are refitted with the same ridge estimator.

Method	Refit NCE $\downarrow$	Refit Traj.-NRMSE $\downarrow$
PriSIFT-Search	2.11	0.804
Full-EBIC	5.88	0.818
MIOSR-exact	6.12	0.824
MIOSR	11.07	1.297
Bayesian-SINDy	14.85	0.868
SR3	20.38	1.385
STLSQ	20.03	1.373
E-SINDy	20.86	1.472

B.5 Degree-Five Scaling

Degree-five libraries provide a harder setting with a larger and more redundant hypothesis space. On the degree-five development benchmark, the specialized PriSIFT proposal achieves a selection F1 of 0.719 and a ranking AUROC of 0.718. Higher-order terms also increase collinearity and conditioning difficulty, making exhaustive model selection substantially more fragile.

As shown in Figure 3c, this difficulty is especially visible in off-trajectory vector-field error: exhaustive procedures deteriorate sharply on degree-five systems, whereas PriSIFT-Search remains

substantially more stable. Together, these results show that specialized prior fitting can provide strong structural proposals for larger libraries while restricted search improves the stability of downstream equation selection.

B.6 Learned Selection Fails under Corpus Shift

During development, we explored several learned alternatives to the fixed task-local selector. These included logistic and multilayer-perceptron candidate rerankers, a Group-DRO variant, and a domain-adaptive weighted information-criterion selector. Candidate features included fit quality, conditioning statistics, term composition, and related task-local quantities.

Each learned selector improved performance on the corpus used to fit or design it, yet each degraded relative to the simple fixed selector when evaluated on a fresh external corpus. The degradation grew as the candidate-feature distribution shifted between development and external systems.

This negative result directly motivated the final PriSIFT decomposition. We retain learning where cross-system structure transfers—the proposal— but use a fixed information criterion for the final decision. The empirical failure of learned reranking is therefore not an auxiliary ablation but part of the evidence supporting the architectural separation in Section 3.4.2.

B.7 Overlap Sensitivity on the Fresh Holdout

The fresh-holdout provenance audit identified one exact development-corpus duplicate and eight systems sharing an active-term pattern with a development system under library symmetries. To verify that the proposal comparison is not driven by this overlap, we recomputed the matched-budget analysis after removing all overlapping systems.

The qualitative conclusion is unchanged: PriSIFT remains ahead of the strongest classical proposal heuristic under the same candidate budget, coefficient fit, and final selector. The overlap audit therefore does not explain the proposal advantage reported in RQ1.

B.8 Proposal Diagnostics

Ranking and pool quality. On the fresh holdout under the full 10-seed analysis, the PriSIFT term ranking attains AUROC 0.694, AUPRC 0.493, and Recall@10 0.838. The 200-candidate proposal pool contains the true support in 61.7% of evaluated tasks, and its OraclePoolF1 is 0.878. The gap between this oracle pool quality and the final selected-model quality indicates substantial remaining headroom in task-local selection.

Error decomposition. We partition discovery failures into *proposal failures*, where the true support is absent from the candidate pool, and *selection failures*, where the true support is available but not selected. Under the most common holdout failure profile, 38.3% of errors are attributable to proposal failure and 47.4% to selection failure, with the remainder involving mixed cases. Thus, neither stage alone explains the remaining error.

Observation-regime detail. The numerical values underlying Figure 4 are provided in Table 4. Per-regime results for the broader baseline family are retained in the frozen evaluation bundle.

C Per-System Results

Aggregate metrics can hide substantial system-to-system variation. We therefore report both structural and end-to-end results for every system in the untouched literature corpus.

Table 7 reports structural macro F1 for each system, including the five degree-five systems used to evaluate performance under larger candidate libraries.

Averaging the per-system structural means in Table 7 gives 0.6782 for PriSIFT-Search, 0.6657 for Full-EBIC, and 0.6489 for MIOSR-exact, matching the 0.678, 0.666, and 0.649 values in Figure 2c up to rounding. This provides a direct consistency check on the canonical Support-F1 aggregation.

Table 7: **Per-system structural macro F1** on the untouched corpus. Values are means over the frozen evaluation seeds; best value in each row is shown in bold.

System (deg.)	PriSIFT	Full-EBIC	MIOSR-ex.	MIOSR	STLSQ	E-SINDy
gh_quadratic_hamiltonian (4)	**1.000**	**1.000**	0.980	0.334	0.294	0.286
js_cubic_damping_quad_stiff (4)	**0.893**	**0.893**	**0.893**	0.470	0.267	0.229
js_helmholtz (4)	**0.865**	**0.865**	**0.865**	0.314	0.247	0.238
js_helmholtz_duffing (4)	0.749	**0.768**	**0.768**	0.389	0.277	0.276
js_lienard_pos_damping (4)	**0.736**	0.716	0.644	0.259	0.255	0.243
js_mixed_damping_stiffness (4)	0.715	**0.724**	0.697	0.338	0.283	0.280
js_no_linear_stiffness (4)	**0.917**	0.892	0.842	0.322	0.258	0.231
js_quadratic_damping (4)	0.927	0.917	**0.937**	0.405	0.224	0.236
murray_strong_allee_pp (4)	0.427	**0.464**	0.459	0.076	0.347	0.347
nm_cubic_damp_hard_stiff (4)	0.750	**0.754**	0.587	0.378	0.297	0.283
nm_nagumo_1962 (4)	**0.496**	0.447	0.457	0.089	0.299	0.347
nm_vdp_duffing (4)	0.353	0.353	**0.377**	0.355	0.370	0.276
pe_lienard_cubic_restoring (4)	0.701	**0.719**	**0.719**	0.398	0.309	0.229
pe_quadratic_system (4)	**0.797**	0.742	0.742	0.517	0.251	0.237
pe_quadratic_system_b (4)	**0.950**	**0.950**	**0.950**	0.557	0.259	0.235
gh_hopf_quintic (5)	0.196	0.231	0.229	0.337	0.590	**0.600**
js_quintic_double_well (5)	**0.593**	0.496	0.485	0.299	0.210	0.205
kz_bautin_nf (5)	0.219	0.252	0.305	0.589	**0.739**	0.727
nm_quartic_vdp (5)	**0.714**	0.702	0.618	0.286	0.236	0.170
pe_quintic_lienard_mixed (5)	**0.566**	0.429	0.424	0.321	0.215	0.214

Table 8 reports the corresponding per-system vector-field errors for the principal end-to-end methods. Together with Table 7, these results illustrate why structural F1 alone is not a complete measure of equation quality: systems with similar structural scores can have markedly different off-trajectory vector-field behavior.

Table 8: **Per-system vector-field NRMSE** on the untouched corpus. Lower is better.

System	PriSIFT-Search	Full-EBIC	MIOSR-exact	Bayesian-SINDy
gh_hopf_quintic	0.1	0.2	0.2	0.2
gh_quadratic_hamiltonian	0.2	0.2	0.2	0.2
js_cubic_damping_quad_stiff	1.1	1.1	1.1	0.7
js_helmholtz	0.5	0.4	0.4	0.6
js_helmholtz_duffing	1.1	0.8	0.8	0.7
js_lienard_pos_damping	1.8	3.2	3.2	1.7
js_mixed_damping_stiffness	1.0	1.3	1.3	0.5
js_no_linear_stiffness	0.6	0.6	1.0	0.7
js_quadratic_damping	0.2	0.4	0.4	0.3
js_quintic_double_well	2.4	5.9	6.0	1.5
kz_bautin_nf	1.0	1.0	1.1	0.4
murray_strong_allee_pp	1.7	4.3	4.3	0.8
nm_cubic_damp_hard_stiff	1.0	5.4	5.4	0.7
nm_nagumo_1962	2.3	7.1	7.2	2.2
nm_quartic_vdp	0.6	0.6	0.6	0.6
nm_vdp_duffing	0.4	0.4	0.4	0.4
pe_lienard_cubic_restoring	0.4	0.4	0.4	0.4
pe_quadratic_system	4.7	12.4	12.4	47.2
pe_quadratic_system_b	0.2	0.2	0.2	0.2
pe_quintic_lienard_mixed	3.7	76.0	84.2	1.4

D Extended Discussion and Limitations

Discussion. Taken together, the experiments suggest a broader design principle for data-driven scientific discovery: the component that transfers most reliably across systems need not be the component that makes the final decision. PriSIFT uses prior-data fitting to amortize *structural proposal*,

where cross-system regularities can be reused, while leaving final model selection to task-local statistical evidence.

This separation helps explain several otherwise distinct observations. The structural proposal transfers across corpora more reliably than development-fitted learned selectors; restricting the hypothesis space can retain exhaustive-level structural recovery while avoiding most candidate evaluations; and, in some regimes, the same restriction also improves coefficient and vector-field fidelity by excluding implausible or poorly conditioned supports before statistical selection. The integral-evidence variant further illustrates the modularity of the decomposition: task-local evidence can be changed without retraining the structural proposal.

We therefore view PriSIFT not as a replacement for sparse model selection, but as a transferable mechanism for deciding which hypotheses are worth exposing to it. In this view, learning and statistics play complementary roles: cross-system experience narrows the hypothesis space, while evidence from the current system determines the final scientific model.

Scope and limitations. The present study focuses on two-dimensional autonomous polynomial ODEs under a declared candidate library. The frozen deployment routes between degree-specific prior-fitted checkpoints rather than learning the candidate library itself. Extending PriSIFT to higher-dimensional systems, non-polynomial libraries, and degree-agnostic proposal models therefore remains open.

Larger libraries introduce stronger redundancy and collinearity. The degree-five experiments demonstrate improved dynamical stability under this more difficult regime and motivate extension to still larger and richer candidate libraries.

The external literature corpus is method-blind and assembled from predefined public sources, but it contains only twenty systems. PriSIFT-Search also retains nontrivial task-local computation because each proposed support must still be fitted and evaluated.

These limitations motivate larger and more diverse external corpora, degree-agnostic and higher-capacity proposal models, richer candidate libraries, and more robust task-local selectors. More broadly, these results establish the proposal–selection decomposition as a foundation for amortized scientific model discovery, with higher-dimensional systems, richer candidate libraries, and degree-agnostic proposal models as natural next steps.