SysBio-SR: Benchmarking Symbolic Regression for Trajectory Fidelity and Mechanism Recovery in Systems Biology

Anonymous authors
Paper under double-blind review

Abstract

Symbolic regression seeks to discover interpretable governing equations for dynamical systems from data. In systems biology, however, accurately reproducing system trajectories does not necessarily imply recovery of the underlying reaction mechanisms. Existing benchmarks evaluate numerical accuracy, expression complexity, or equation-level similarity, but not agreement between recovered reactions and reference mechanisms in biological systems. We introduce SysBio-SR, a benchmark of 198 literature-based biological ODE systems with reaction annotations spanning mass action, Michaelis–Menten, Hill, and complex kinetics. Trajectory fidelity is evaluated across all systems under in-distribution reconstruction and out-of-distribution (OOD) extrapolation, while mechanism recovery is assessed on the subset of 161 systems. Of 11 state-of-the-art symbolic-discovery methods, only 3 support mechanism recovery, as most output unconstrained expressions that cannot map to biological mechanisms. Joint analysis across systems highlights how severely trajectory-based evaluation overestimates true mechanism recovery. Motivated by these findings, we develop KineSelect, a framework that searches candidate reaction topologies under biologically grounded kinetic laws to select optimal mechanisms that balance trajectory fidelity with structural parsimony. Compared with the strongest baseline, KineSelect improves aggregate OOD extrapolation R^2 by 19.2% and overall reaction F1 score by 14.3%.

1 Introduction

Biological systems evolve through interacting processes such as production, degradation, and regulation (Kholodenko, 2006; Fell & Cornish-Bowden, 1997; Aldridge et al., 2006), which are classically modeled using ordinary differential equations (ODEs). Traditionally, constructing these mathematical models relies on labor-intensive manual curation from domain literature and expert intuition. Symbolic regression (SR) offers a promising data-driven alternative to automate this workflow by inferring interpretable governing equations directly from time series measurements (Koza, 1994; Schmidt & Lipson, 2009; Brunton et al., 2016). Unlike black-box neural approximations, SR discovers explicit functional forms to expedite model development and uncover novel regulatory rules.

In systems biology, a mechanism comprises both the reaction network topology and its governing kinetic rate laws. However, existing symbolic regression benchmarks evaluate models primarily on numerical trajectory accuracy, expression complexity, or simple equation matching (La Cava et al., 2021; Matsubara et al., 2022; d'Ascoli et al., 2023; Bideh et al., 2026). Moreover, these suites focus heavily on physics-inspired scalar expressions (Udrescu & Tegmark, 2020; La Cava et al., 2021) or generic dynamical systems (d'Ascoli et al., 2023; Bideh et al., 2026; Liu et al., 2025), leaving mechanism recovery in coupled biological ODEs unexamined. This creates a critical evaluation gap: because distinct reaction pathways can produce virtually similar trajectory curves (Szederkényi et al., 2011), high trajectory fidelity does not imply biological correctness. Figure 1 illustrates this gap: on a representative system (Munz et al., 2009), multiple SR baselines achieve near-perfect trajectory fidelity $R^2 \approx 1.0$, yet their symbolic-term F1 scores remain low, ranging from 0 to 0.67.

To address this gap, we present **SysBio-SR**, a benchmark designed for mechanism-aware evaluation of symbolic regression in systems biology. SysBio-SR curates 198 biological ODE models

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.

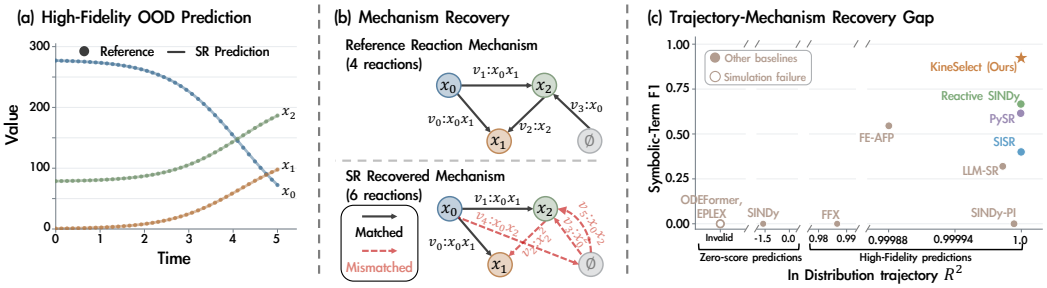


Figure 1: **Trajectory–mechanism gap.** (a) On a biological system Munz et al. (2009), an SR baseline closely matches the OOD trajectory. (b) Its recovered mechanism differs from ground truth. (c) Existing SR methods yield high trajectory fidelity but low, markedly different symbolic-term F_1.

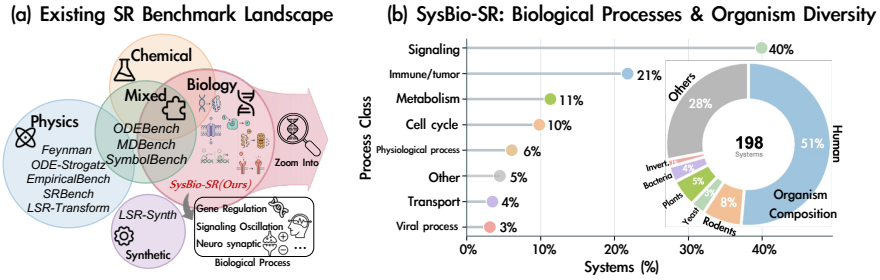


Figure 2: **Positioning and biological coverage of SysBio-SR.** (a) Landscape of SR benchmarks across scientific domains. (b) Biological process and organism composition across 198 systems.

covering signaling, metabolism, cell-cycle dynamics, and other biological processes across diverse organisms (Figure 2). To enable explicit mechanism verification, we provide reference annotations for all reactions and their kinetic rate laws, including mass action, Michaelis–Menten, and Hill dynamics (Klipp et al., 2016; Johnson, 2013; Weiss, 1997). SysBio-SR highlights two challenges in biological symbolic regression: (1) local kinetic-law selection (identifying the specific rate law for each reaction) (Xu, 2023), and (2) global mechanism identification (selecting ground-truth reaction network among trajectory-equivalent alternatives) (Wong et al., 2023; Szederkényi et al., 2011).

To tackle these challenges, we introduce **KineSelect**, a biological symbolic regression framework that searches candidate reaction topologies, identifies their kinetic laws, and selects the optimal model by balancing trajectory fidelity with structural parsimony. Our contributions are summarized as follows:

- We present SysBio-SR, the first reaction-annotated benchmark suite enabling mechanism recovery evaluation for biological ODE discovery, comprising 198 curated biological systems spanning mass action, Michaelis–Menten, Hill, and complex kinetics.
- We propose KineSelect, a kinetic selection framework that navigates expression spaces to biologically grounded kinetic rate laws, boosting aggregate OOD extrapolation R^2 by 19.2% and overall reaction F1 score by 14.3% over the strongest baseline.
- Through evaluation of 11 SR methods across 198 systems, we expose a fundamental limitation of symbolic regression in systems biology: high trajectory fidelity drastically overestimates real-world utility by failing to recover ground-truth mechanisms.

2 Related Work

2.1 Benchmarks for Symbolic Regression and ODE Discovery

Existing SR benchmarks like AI Feynman (Udrescu & Tegmark, 2020; Matsubara et al., 2022), SRBench family (La Cava et al., 2021; de Franca et al., 2024; Imai Aldeia et al., 2025), and LLM-

SRBench (Shojaee et al., 2025) focus primarily on scalar physics-inspired or synthetic equations, with limited coverage of ODE systems. Recent ODE benchmarks, including ODEBench (d'Ascoli et al., 2023), MDBench (Bideh et al., 2026), and SymbolBench (Liu et al., 2025), expand to multi-dimensional coupled systems (up to six variables), but draw from generic scientific domains and lack structural annotations for mechanism assessment. In Table 1, SysBio-SR fills this gap with 198 systems biology ODE models with reference reaction signatures for mechanism recovery.

Table 1: **Comparison of symbolic regression benchmarks.** # Equations counts equations across all systems. We count only SR equation-discovery instances. Per-benchmark composition is detailed in Appendix A.6. Structural annot. indicates whether explicit structural annotations are provided (e.g., physical interaction terms or reaction rate laws). ✓ = yes; × = no; ∼ = partial.

Benchmark	# Systems	# Equations	Max dim.	Domain	ODE	Coupled	Structural annot.
AI Feynman (Udrescu & Tegmark, 2020)	120	120	1	Physics	×	×	×
SRBench (La Cava et al., 2021)	123	130	2	Physics	∼	×	×
EmpiricalBench (Cranmer, 2023)	9	9	1	Physics	×	×	×
ODEBench (d'Ascoli et al., 2023)	63	117	4	Mixed	✓	✓	×
LLM-SRBench (Shojaee et al., 2025)	239	239	1	Phys.+Synth.	∼	×	×
SymbolBench (SR task) (Liu et al., 2025)	156	383	4	Mixed	✓	✓	×
MDBench (Bideh et al., 2026)	77	141	6	Mixed	✓	✓	×
SysBio-SR	**198**	**661**	**5**	**Sys. Biology**	✓	✓	✓

2.2 Symbolic Regression Methods for Dynamical Systems

Symbolic regression methods for dynamical systems span sparse regression, evolutionary search, and neural symbolic approaches. Sparse regression methods such as SINDy (Brunton et al., 2016), including extensions for rational dynamics (Kaheman et al., 2020) and reaction networks (Hoffmann et al., 2019), select active terms from predefined candidate libraries. Evolutionary methods like genetic programming (Koza, 1994) and PySR (Cranmer, 2023) explore broader expression spaces with complexity penalties, but frequently produce mathematically redundant or biologically uninterpretable expressions. Stoichiometrically informed symbolic regression (SISR) (Palma Banos et al., 2026) constrains search spaces by evolving candidate reaction mechanisms from time-series data; however, its scope is limited to case studies without standardized mechanism recovery evaluation and broad kinetic coverage. Recent neural symbolic approaches leveraging pre-trained transformers (Kamienny et al., 2022; d'Ascoli et al., 2023; Vastl et al., 2024; Chen et al., 2025) and LLMs (Holt et al., 2024; Shojaee et al., 2024; Li & Pan, 2026; Xia et al., 2025) enable rapid equation inference, but remain sensitive to pre-training distributions and underexplored for biological mechanism recovery.

Unlike unconstrained search or fixed-library methods, our KineSelect explicitly couples kinetic law selection with reaction topology search, scoring candidate mechanisms on structural parsimony and trajectory fidelity. SysBio-SR evaluates symbolic regression methods across diverse systems along two axes, trajectory fidelity and mechanism recovery.

3 The SysBio-SR Benchmark

SysBio-SR contains 198 literature-derived, coupled biological ODE systems (661 equations) spanning diverse kinetics, with standardized symbolic targets, reference trajectories, and reaction annotations. All 198 systems support trajectory evaluation, while 161 support mechanism recovery.

3.1 Problem Formulation

A system with d state variables follows a coupled vector field

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x};\boldsymbol{\theta}) = \mathbf{S}\mathbf{v}(\mathbf{x};\boldsymbol{\theta}), \qquad \mathbf{x}(t) \in \mathbb{R}^d, \tag{1}$$

where $\boldsymbol{\theta}$ is the parameter vector, $\mathbf{S} \in \mathbb{R}^{d\times r}$ records how r reaction channels change the state variables, and $\mathbf{v}$ contains the corresponding reaction-rate functions. These rate laws include mass action, Michaelis–Menten, and Hill kinetics commonly used in systems biology models (Johnson, 2013; Weiss, 1997). For system m, a symbolic regression method receives a fitting trajectory and returns a symbolic vector field for the coupled system:

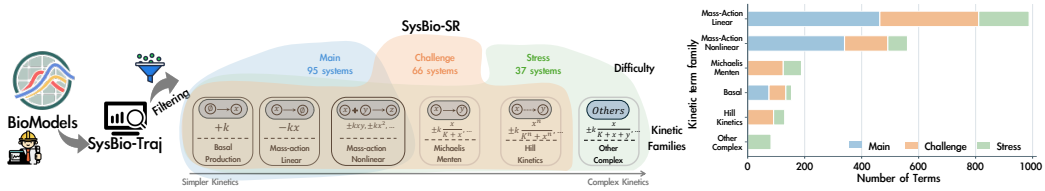
$$\mathcal{D}_m^{\text{fit}} = \{(t_i, \mathbf{x}_m(t_i))\}_{i=1}^N \xrightarrow{\text{SR method}} \widehat{\mathbf{f}}_m. \tag{2}$$

Here, N is the number of samples in the fitting trajectory.

3.2 Dataset Construction and Composition

Source models and curation. SysBio-Traj (Rao et al., 2026) collects literature-derived BioModels entries encoded in Systems Biology Markup Language (SBML) (Hucka et al., 2003; Malik-Sheriff et al., 2020; Lüders et al., 2022). From this collection, we construct SysBio-SR by retaining deterministic, autonomous ODE systems with at most five state variables, while excluding event-driven or piecewise dynamics, explicit time dependence, and systems that fail numerical validation. This process yields 198 systems with 661 equations. Figure 3(a) summarizes this curation pipeline.

Standardization and trajectories. For each system, we rename states as $x_0, \ldots, x_{d-1}$ and parameters as $c_0, \ldots, c_{p-1}$, and provide a standardized symbolic ODE representation. We generate a fitting trajectory with $N = 512$ samples from the default initial condition. For out-of-distribution (OOD) extrapolation, we generate a second 512-point trajectory from a perturbed initial condition while holding the vector field and parameters fixed.



(a) Overview of SysBio-SR Benchmark (b) Kinetics Families Distribution

Figure 3: **Construction and kinetic composition of SysBio-SR.** (a) Curation pipeline partitions systems into three kinetics-based subsets: Main (mass action only; $n = 95$), Challenge (with Michaelis–Menten or Hill kinetics; $n = 66$), and Stress (with complex non-canonical kinetics; $n = 37$). (b) Distribution of reaction terms across kinetic families for each subset.

Composition. SysBio-SR covers systems with one to five state variables with diverse reaction topologies and kinetic laws. We partition the 198 systems into three subsets based on maximum kinetic complexity: Main (95 systems, exclusively mass action kinetics), Challenge (66 systems containing Michaelis–Menten or Hill kinetics), and Stress (37 systems, containing complex non-canonical kinetics). Higher complexity subsets may also include reactions from lower kinetic tiers. Figure 3(b) details the kinetic-family composition across subsets, while Appendix Figure 7 summarizes distributions for system dimension, reaction count, and species per reaction.

3.3 Reaction Annotations

We provide reaction annotations for all 198 systems. Starting from the reference ODEs, we decompose each RHS into signed terms, separate rate coefficients from kinetic factors, and label the species-dependent kinetic terms. The resulting records specify consumption or production, participating species, and kinetic forms such as mass action, Michaelis–Menten, and Hill terms. We assess annotation consistency through reconstruction checks and manual review. The annotations provide reference signatures for mechanism-recovery evaluation on the 161 Main and Challenge systems. The 37 Stress systems retain annotations but include kinetic forms outside the matching protocol. Appendix A.5 details annotation construction, representation, and signature matching.

4 Mechanism-Aware Evaluation Protocol

Evaluation axes. We evaluate discovered ODE systems along two axes: trajectory fidelity and mechanism recovery. Trajectory fidelity is measured on all 198 systems via in-distribution (ID) reconstruction and out-of-distribution (OOD) extrapolation (d'Ascoli et al., 2023) from perturbed initial conditions. Mechanism recovery is evaluated on the Main and Challenge subsets by matching predicted terms against reference reaction signatures. To prevent data leakage, reaction annotations are strictly withheld during model fitting and selection, serving exclusively for post-hoc evaluation.

Trajectory metrics. For each system, we numerically integrate the recovered ODE from both ID and OOD evaluation initial conditions and compare with reference trajectory. We compute state-averaged coefficients of determination R^2_m across the evaluation time grid and aggregate them as

$$R^2\text{-score} = \frac{1}{M}\sum_{m=1}^{M} \max\left(R^2_m, 0\right), \tag{3}$$

where M is the system count in the target set. System trajectories that diverge, produce non-finite values, or encounter solver integration timeouts receive $R^2_m = 0$. We additionally report $\mathrm{Acc}@.90$, the fraction of systems with $R^2_m \geq 0.90$. Following ODEFormer (d'Ascoli et al., 2023), expression complexity is evaluated as the total symbolic tree node count summed across state equations. Model selection parsimony is defined in Section 5.3, with full metric specifications in Appendix A.4.

Reaction metrics. Equivalent ODE expressions can take different forms (d'Ascoli et al., 2023). We therefore match recovered reactions against reference mechanisms through canonical reaction signatures (Appendix A.5). A predicted signature is counted as a true positive if its reaction topology and kinetic law strictly match an annotated reaction, independent of numerical parameter values. Unmatched predicted and reference signatures constitute false positives and false negatives, respectively. We compute precision, recall, and F1 for each system and report their means across systems.

Observation noise. To evaluate robustness under noisy measurements, we corrupt clean fitting trajectories with independent state-wise Gaussian noise:

$$\widetilde{x}_i(t) = x_i(t) + \sigma_{\mathrm{rel}}\,\sigma_i \epsilon_{i,t}, \qquad \epsilon_{i,t} \sim \mathcal{N}(0,1), \qquad \sigma_{\mathrm{rel}} \in \{0.001, 0.01, 0.05, 0.1\}. \tag{4}$$

Here, σ_i denotes the temporal standard deviation of state i on the clean trajectory to adjust for varying state magnitudes. The relative noise magnitude $\sigma_{\mathrm{rel}} = 0$ represents noiseless baseline data. Regardless of the noise level, all discovered systems are strictly evaluated against clean ID reconstruction and OOD extrapolation reference trajectories. Appendix A.3 specifies the full protocol.

Evaluated baselines. We evaluate 11 symbolic discovery algorithms (12 total configurations) across four methodological families. (1) Symbolic search and genetic programming: EPLEX (2 configs) (La Cava et al., 2016), FE-AFP (Schmidt & Lipson, 2010; La Cava et al., 2021), FFX (McConaghy, 2011), and PySR (Cranmer, 2023). (2) Sparse system identification methods: SINDy (Brunton et al., 2016), SINDy-PI (Kaheman et al., 2020), and Reactive SINDy (Hoffmann et al., 2019). (3) Neural & language models: ODEFormer (d'Ascoli et al., 2023) and LLM-SR (Shojaee et al., 2024). (4) Reaction-structured methods: SISR (Palma Banos et al., 2026) and our KineSelect. All configurations undergo trajectory evaluation. Mechanism recovery is evaluated exclusively on Reactive SINDy, SISR, and KineSelect by mapping their reaction-structured outputs to reference reaction signatures. Implementation details and method specific configurations are provided in Appendix D.

KineSelect component studies. To evaluate architectural choices in KineSelect, we compare standard α-threshold (Appendix B.3) against Fidelity–Parsimony Model Selection (FPMS) on identical candidate archives. Additionally, component ablations isolate the individual contributions of derivative loss and specialized kinetic complexity across both trajectory fidelity and mechanism recovery.

5 KineSelect

KineSelect structures biological symbolic regression into three core components: Reaction Structure Search (RSS), Kinetic Form Identification (KFI), and Fidelity–Parsimony Model Selection (FPMS). In the search loop, RSS proposes regulator-guided reaction topologies, while KFI identifies their kinetic forms and evaluates the resulting mechanisms. The search maintains an archive of non-dominated candidate mechanisms, from which FPMS selects the final mechanism by balancing fidelity and parsimony.

5.1 Reaction Structure Search

Consistent with Eq. (1), we represent a candidate mechanistic model $\mathcal{M}$ with reaction set $\mathcal{R}(\mathcal{M})$ as $\dot{\mathbf{x}} = \sum_{r\in\mathcal{R}(\mathcal{M})} \mathbf{s}_r v_r(\mathbf{x};\boldsymbol{\theta}_r)$, where $\mathbf{s}_r \in \mathbb{Z}^d$ is the stoichiometric vector of reaction r, v_r is its kinetic rate law, and $\boldsymbol{\theta}_r$ contains the overall rate constant k_r and any kinetic shape parameters. Building on the reaction-level evolutionary representation of SISR (Palma Banos et al., 2026), RSS extends

structural search by incorporating regulator roles. For each proposed reaction topology, RSS assigns reactant, product, and regulator species. Regulators exert zero net stoichiometry ($\mathbf{s}_{r,i} = 0$) but directly modulate the rate law v_r (e.g., regulator x_2 in $x_0 \xrightarrow{x_2} x_1$ with rate $v = kx_0x_2$). Directly encoding discrete kinetic expressions into each reaction creates a combinatorial search explosion over topological and functional spaces. To resolve this, RSS restricts its search to reaction topologies and species roles, delegating rate-law selection to KFI during candidate evaluation. KineSelect maintains separate subpopulations partitioned by model size $c = |\mathcal{R}(\mathcal{M})| \in \{1, \ldots, c_{\max}\}$, pooling kinetic-resolved candidates into a shared archive $\mathcal{A}$ for selection by FPMS (Figure 4).

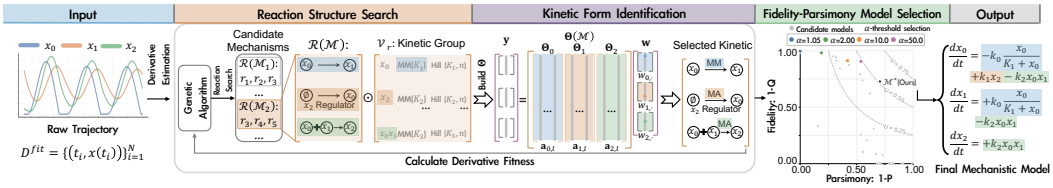


Figure 4: **Overview of the KineSelect framework.** Reaction Structure Search proposes reaction topologies with reactants, products, and optional regulators. Kinetic Form Identification jointly assigns kinetic expressions across reaction groups and evaluates candidates using derivative loss. Fidelity–Parsimony Model Selection selects the optimal mechanism from the candidate archive by balancing trajectory fidelity and structural parsimony.

5.2 Kinetic Form Identification

Conditioned on the reaction topologies proposed by RSS, KFI resolves the functional form of each rate law within the candidate evaluation loop. For every proposed reaction $r \in \mathcal{R}(\mathcal{M})$, KFI constructs a candidate library of unit-rate expressions $\mathcal{V}_r = \{\bar{v}_{r,\ell}\}$, where the bar indicates that each candidate excludes rate constants k_r. Each candidate factorizes into reactant and regulator terms:

$$\bar{v}_{r,\ell}(\mathbf{x}) = \phi_{r,\ell}(\mathbf{x})\, \psi_{r,\ell}(\mathbf{x}). \tag{5}$$

Here, $\phi_{r,\ell}$ models mass action or Michaelis–Menten reactant kinetics, while $\psi_{r,\ell}$ models optional regulatory influences (Explicit functional forms are given in Appendix B.2). Non-linear parameters (e.g., affinity constants or Hill coefficients) are discretized over predefined grid settings, yielding discrete elements in $\mathcal{V}_r$.

To evaluate all kinetic configurations across the system simultaneously, KFI constructs a reaction-block dictionary from these candidates. Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ denote the observed trajectory matrix and $\dot{\mathbf{X}} \in \mathbb{R}^{N\times d}$ its estimated time derivatives. To prevent state variables with large magnitudes from dominating model selection, we define a scale-normalization matrix $\Gamma = \mathrm{diag}(\gamma_0, \ldots, \gamma_{d-1})$ with $\gamma_j = \max_{1\le n\le N} |\dot{X}_{nj}|$ for $j = 0, \ldots, d-1$, setting the target vector to $\mathbf{y} = \mathrm{vec}(\dot{\mathbf{X}}\Gamma^{-1}) \in \mathbb{R}^{Nd}$.

For each candidate expression $\bar{v}_{r,\ell} \in \mathcal{V}_r$, define its stoichiometrically coupled dictionary column as:

$$\mathbf{a}_{r,\ell} = \frac{\tilde{\mathbf{a}}_{r,\ell}}{\|\tilde{\mathbf{a}}_{r,\ell}\|_2}, \qquad \tilde{\mathbf{a}}_{r,\ell} = \mathrm{vec}\left(\bar{v}_{r,\ell}(\mathbf{X})\, \mathbf{s}_r^\top \Gamma^{-1}\right) \in \mathbb{R}^{Nd}. \tag{6}$$

Grouping columns by reaction yields the reaction-block dictionary $\mathbf{\Theta}_r = [\mathbf{a}_{r,1} \mid \cdots \mid \mathbf{a}_{r,|\mathcal{V}_r|}]$, forming the system-wide dictionary $\mathbf{\Theta}(\mathcal{M}) = [\mathbf{\Theta}_r]_{r\in\mathcal{R}(\mathcal{M})}$. Unlike standard SINDy (Brunton et al., 2016), which fits state equations independently, Eq. (6) enforces stoichiometry vector $\mathbf{s}_r$ across all affected state derivatives while allowing zero-net regulators ($\mathbf{s}_{r,i} = 0$) to modulate rates without violating mass conservation.

KFI jointly evaluates competing rate laws by solving a non-negative regularized regression over the block dictionary:

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\ge 0} \|\mathbf{y} - \mathbf{\Theta}(\mathcal{M})\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2, \qquad \lambda \ge 0. \tag{7}$$

From the fitted weight vector $\widehat{\mathbf{w}}$, KFI selects the dominant candidate within each reaction group:

$$\ell_r^\star = \arg\max_{\ell\in\{1,\ldots,|\mathcal{V}_r|\}} \widehat{w}_{r,\ell}. \tag{8}$$

Finally, KFI refits the rate constants $k_r \ge 0$ for the selected expressions $v_r(\mathbf{x}) = k_r \bar{v}_{r,\ell_r^\star}(\mathbf{x})$ against derivative target $\mathbf{y}$, returning the resolved model and derivative loss to the evolutionary search loop.

5.3 Fidelity–Parsimony Model Selection

Post-search, FPMS identifies the final mechanism from the candidate archive $\mathcal{A}$ by evaluating candidates across four complementary criteria. Trajectory fidelity is captured by combining derivative loss $\mathcal{L}_i^{\mathrm{der}}$ and numerical trajectory simulation loss $\mathcal{L}_i^{\mathrm{sim}}$. To compress heavy-tailed loss distributions, both fidelity metrics undergo log-transformation prior to normalization. Appendix B.3 specifies the transformation and numerical policies.

Structural parsimony is evaluated using two distinct complexity measures. (1) Expanded tree complexity (C_i^{tree}): The total expression-tree node count obtained after expanding reaction rate laws into full state-level ODEs (Palma Banos et al., 2026). (2) Specialized kinetic complexity (C_i^{spec}): The count of non-mass-action kinetic terms (i.e., Michaelis–Menten reactant terms and Hill regulatory expressions), penalizing complex rate expressions independent of stoichiometric fan-out (how many state equations the reaction affects).

All four criteria are independently min–max normalized over $\mathcal{A}$ (denoted by $\widetilde{\cdot} \in [0,1]$). FPMS aggregates these into a composite lack-of-fidelity score Q_i and a model-complexity score P_i:

$$Q_i = \frac{\widetilde{\mathcal{L}}_i^{\mathrm{der}} + \widetilde{\mathcal{L}}_i^{\mathrm{sim}}}{2}, \qquad P_i = \frac{\widetilde{C}_i^{\mathrm{tree}} + \widetilde{C}_i^{\mathrm{spec}}}{2}. \tag{9}$$

Lower Q_i and P_i indicate higher fidelity and lower model complexity, respectively.

FPMS selects the optimal model $\mathcal{M}^\star \in \mathcal{A}$ by maximizing a multiplicative fidelity–parsimony utility:

$$U_i = (1 - Q_i)(1 - P_i), \qquad \mathcal{M}^\star = \arg\max_{\mathcal{M}_i \in \mathcal{A}} U_i. \tag{10}$$

Unlike additive scoring, this multiplicative objective enforces strict balance: $U_i \to 0$ whenever either fidelity or parsimony is severely compromised ($Q_i \to 1$ or $P_i \to 1$), discouraging the selection of accurate yet overly complex models or simple yet inaccurate mechanisms.

6 Experimental Results

6.1 Benchmark-Wide Trajectory Recovery

KineSelect achieves the highest trajectory fidelity under both ID reconstruction and OOD extrapolation (Table 2). Relative to the strongest baseline SISR, KineSelect improves overall ID and OOD R^2 by 16.5% (0.770 vs. 0.661) and 19.2% (0.508 vs. 0.426), respectively, while increasing OOD $\mathrm{Acc}@.90$ by 14.0% (0.374 vs. 0.328). The concordant gains in average R^2 and high-fidelity coverage (Acc@.90) confirm that these improvements reflect robust, system-wide recovery rather than outlier-driven averages. OOD extrapolation serves as a crucial differentiator among methods with similar ID fitting performance. For instance, while FFX and SISR display similar ID R^2, SISR has substantially higher OOD R^2, revealing greater dynamical generalization (Table 2). KineSelect maintains the OOD advantage across all three subsets: relative to the strongest comparator in each regime, it improves OOD R^2 by 19.2% on Main (0.601 vs. 0.504), 20.1% on Challenge (0.437 vs. 0.364), and 8.2% on Stress (0.395 vs. 0.365). Stress is the most challenging subset, combining the lowest absolute OOD R^2 with the smallest relative gain. A systematic decline from ID to OOD performance is observed across all three subsets for all methods.

This higher trajectory fidelity is accompanied by increased expression complexity relative to SISR (62.13 vs. 37.66 nodes). This reflects KineSelect's capacity to express the non-mass-action kinetic rate laws (e.g., Hill equation) necessary for faithful dynamics.

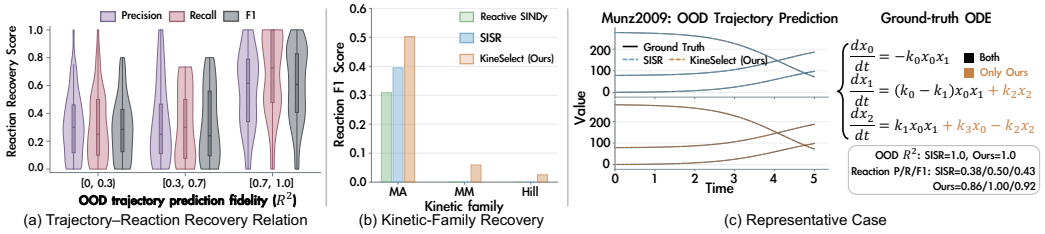
6.2 Trajectory Fidelity and Mechanism Recovery

High trajectory fidelity does not guarantee true mechanism recovery. In Figure 5(a), reaction precision, recall, and F_1 score distributions remain broadly spread even within the highest OOD R^2 regime, confirming that accurate curve-fitting can coexist with incorrect or incomplete reaction recovery. See Appendix C.1 for detailed F1 at OOD $R^2 \geq 0.90$. Furthermore, trajectory metrics mask severe performance gaps across kinetic families (Figure 5(b)). While trajectory fits remain high, reaction F_1 scores drop sharply for Michaelis–Menten and Hill kinetics compared to mass action

Table 2: **Trajectory fidelity on SysBio-SR.** Expression complexity, R^2, and Acc@.90 under ID reconstruction and OOD extrapolation for the full benchmark and its subsets. Acc@.90 denotes the fraction of systems with $R^2 \geq 0.90$. Bold = best; underlined = second-best.

Method	Overall					Main		Challenge		Stress	
	Comp.↓	R^2		Acc.		R^2		R^2		R^2	
		ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
EPLEX (default)	138.03	0.014	0.014	0.005	0.010	0.011	0.008	0.024	0.030	0.000	0.000
EPLEX (poly)	84.48	0.086	0.050	0.056	0.030	0.105	0.058	0.082	0.056	0.042	0.017
FE-AFP	51.82	0.123	0.094	0.106	0.071	0.161	0.121	0.103	0.081	0.061	0.049
FFX	45.92	0.658	0.367	0.525	0.222	0.755	0.402	0.592	0.364	0.529	0.280
PySR	**27.27**	0.342	0.233	0.303	0.177	0.488	0.332	0.234	0.174	0.161	0.086
SINDy	83.62	0.038	0.033	0.035	0.030	0.035	0.018	0.063	0.075	0.000	0.000
SINDy-PI	159.86	0.391	0.310	0.379	0.273	0.491	0.393	0.333	0.261	0.241	0.182
Reactive SINDy	46.95	0.166	0.142	0.101	0.091	0.246	0.220	0.094	0.068	0.089	0.074
ODEFormer	30.23	0.215	0.136	0.111	0.061	0.270	0.158	0.167	0.118	0.161	0.112
LLM-SR	85.26	0.313	0.222	0.288	0.182	0.403	0.299	0.224	0.129	0.242	0.191
SISR	37.66	0.661	0.426	0.561	0.328	0.737	0.504	0.594	0.347	0.587	0.365
KineSelect	62.13	**0.770**	**0.508**	**0.662**	**0.374**	**0.860**	**0.601**	**0.709**	**0.437**	**0.647**	**0.395**

due to increasing functional complexity. The Munz2009 system (Munz et al., 2009) illustrates this divergence (Figure 5(c)): while methods achieve a perfect OOD R^2 of 1.00, KineSelect attains a reaction F1 of 0.92, compared with 0.43 for SISR. See Appendix C.5 for additional case studies.



(a) Trajectory–Reaction Recovery Relation (b) Kinetic-Family Recovery (c) Representative Case

Figure 5: **Trajectory fidelity vs. mechanism recovery.** (a) KineSelect reaction precision, recall, and F1 distributions by OOD R^2 intervals. (b) Reaction F1 by kinetic family. (c) Predicted trajectories for the Munz2009 system under SISR (top) and KineSelect (bottom) against ground truth. Black and orange equation terms denote recovery by both methods and exclusively by KineSelect, respectively.

6.3 Mechanism Recovery Across Subsets and Kinetic Families

KineSelect ranks first in precision, recall, and F1 across the Overall, Main, and Challenge evaluations (Table 3). Relative to SISR, it improves Overall reaction F1 by 14.3% (0.447 vs. 0.391), with the gain persisting on Challenge (0.285 vs. 0.225). The large gain in recall compared to precision ($+0.066$, versus $+0.029$) indicates that KineSelect captures a larger proportion of true reactions per system while still reducing false positives. All three methods suffer performance drops on the Challenge subset, highlighting the inherent difficulty of mechanism recovery.

Mechanism recovery varies significantly across kinetic families (Figure 5(b)). On mass action dynamics, KineSelect raises reaction F_1 to 0.503 (vs. 0.395 for SISR). For non-mass-action kinetics, baseline zero scores (Reactive SINDy and SISR) reflect expressivity limits rather than search failure, as neither framework includes rational functional terms in its candidate basis (Hoffmann et al., 2019; Palma Banos et al., 2026). In contrast, only KineSelect achieves non-zero recovery ($F_1 = 0.059$ for Michaelis–Menten and 0.025 for Hill). While these low scores highlight the difficulty of identifying complex rate laws from trajectory data, they demonstrate that basis expressivity is a strict prerequisite for non-mass-action discovery. Kinetic-family evaluations provide a crucial diagnostic tool alongside trajectory metrics (see Appendix C.2 for detailed precision–recall breakdowns).

Table 3: **Reaction recovery on SysBio-SR.** Mean precision, recall, and F1 scores across systems for three methods that output reaction structures. Bold = best; underlined = second-best.

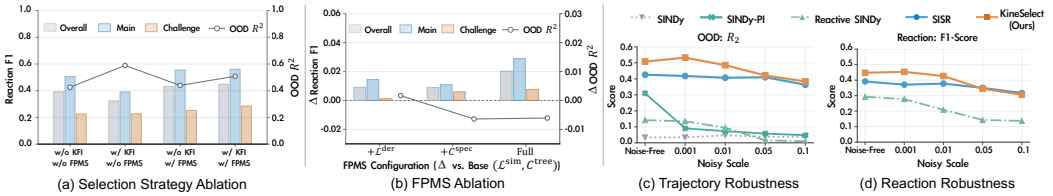
Method	Overall			Main			Challenge		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Reactive SINDy	0.320	0.370	0.292	0.383	0.477	0.363	0.228	0.215	0.191
SISR	<u>0.409</u>	<u>0.416</u>	<u>0.391</u>	<u>0.528</u>	<u>0.543</u>	<u>0.507</u>	<u>0.239</u>	<u>0.234</u>	<u>0.225</u>
KineSelect	**0.438**	**0.482**	**0.447**	**0.544**	**0.606**	**0.560**	**0.285**	**0.305**	**0.285**

6.4 Component Ablations and Noise Robustness

Figure 6(a) shows that the reaction-recovery benefit of KFI depends critically on the final model selection strategy. Under a standard α-threshold rule, enabling KFI improves OOD R^2 (0.598 vs. 0.426) but reduces reaction F1 (0.323 vs. 0.391). Applying FPMS elevates KFI reaction F1 from 0.323 to 0.447, yielding a 3.7% gain in F1 and a 15.5% increase in OOD R^2 for the main subset. Sweeping α further reveals divergent optima for trajectory fit versus mechanism recovery (Appendix C.3).

Ablating individual FPMS criteria demonstrates their additive value (Figure 6(b)). Starting from the base criteria using simulation loss and expression-tree complexity, introducing either derivative loss or specialized kinetic complexity improves overall reaction F1. Combining all four criteria achieves peak F1 across the Overall, Main, and Challenge benchmarks while preserving OOD trajectory fit. Crucially, reaction F1 provides fine-grained discrimination between criteria variants where the OOD trajectory scores remain indistinguishable. KineSelect adopts this complete four-criterion configuration (see Appendix C.3 for complete sensitivity and ablation results).

Finally, evaluating robustness across relative noise levels shows that KineSelect substantially outperforms SISR in both trajectory fit and reaction recovery at low-to-moderate noise ($\sigma_{\text{rel}} \leq 0.01$; Figure 6(c,d)). As noise increases ($\sigma_{\text{rel}} \geq 0.05$), the performance gap narrows, yet KineSelect retains superior OOD trajectory extrapolation ($R^2 = 0.421$ vs. 0.410 at $\sigma_{\text{rel}} = 0.05$; 0.383 vs. 0.363 at $\sigma_{\text{rel}} = 0.10$). Full noise-level distributions are provided in Appendix C.4.



(a) Selection Strategy Ablation (b) FPMS Ablation (c) Trajectory Robustness (d) Reaction Robustness

Figure 6: **KineSelect component ablations and noise robustness.** (a) KFI and FPMS ablations of KineSelect, with α-threshold selection used when FPMS is disabled. (b) Performance changes Δ when adding derivative loss and specialized kinetic complexity criteria individually or jointly to the base FPMS criteria (simulation loss and expression-tree complexity). (c,d) OOD extrapolation R^2 and reaction F1 under increasing relative noise levels.

7 Conclusion and Limitations

This work introduced SysBio-SR, a benchmark of 198 literature-based biological ODE systems designed to evaluate symbolic regression on both trajectory fidelity and mechanism recovery. Crucially, our benchmark reveals that the two objectives are decoupled: high trajectory accuracy does not guarantee true mechanistic discovery. KineSelect achieves superior OOD extrapolation R^2 and overall reaction F1 compared to strong baselines like SISR, while producing more complex expressions.

The SysBio-SR benchmark focuses on low-dimensional, deterministic ODE systems evaluated on simulated trajectories from curated reference models. Future work will investigate reliable recovery of reactions involving Michaelis–Menten or Hill kinetics, as well as extend the benchmark to higher-dimensional systems and more complex kinetic mechanisms. We recommend assessing biological symbolic regression through both trajectory fidelity and explicit mechanistic agreement.

AI USE STATEMENT

We used ChatGPT for literature searches, figure preparation using recorded experimental data, and assistance with code review. We used Gemini for manuscript language editing. We also used generative AI tools to obtain feedback on experimental design. We did not use these tools to formulate research questions or hypotheses or to design the methods or mathematical formulations. We manually reviewed the code and checked the mathematical formulations and method descriptions. We verified numerical values in the tables, figures, and text against the original experimental records. We have reviewed the entire manuscript and take responsibility for its final content, including material prepared with AI assistance.

ETHICS STATEMENT

This work uses publicly available, literature-derived systems-biology ODE models and trajectories simulated from these models. It involves no new experiments on human participants or animals and no collection or processing of personally identifiable data.

REPRODUCIBILITY STATEMENT

Appendix A documents the benchmark construction, trajectory generation, and evaluation protocols, including reaction matching, score aggregation, and failure handling. Appendix B specifies the KineSelect algorithm, including RSS, KFI, and FPMS. Appendix C reports supplementary analyses and experiments, including selection sensitivity and ablations, noise robustness, and representative cases. Appendix D describes the experimental pipeline, baseline implementations, method configurations, search budgets, and settings for the selection and noise experiments. We will submit the benchmark data and code for running and evaluating the experiments as anonymized supplementary material, together with software dependencies and environment specifications.

REFERENCES

Luis U Aguilera and Jesús Rodríguez-González. Studying hiv latency by modeling the interaction between hiv proteins and the innate immune response. *Journal of Theoretical Biology*, 360:67–77, 2014.

Bree B Aldridge, John M Burke, Douglas A Lauffenburger, and Peter K Sorger. Physicochemical modelling of cell signalling pathways. *Nature cell biology*, 8(11):1195–1203, 2006.

Amirmohammad Ziaei Bideh, Aleksandra Georgievska, and Jonathan Gryak. Mdbench: Benchmarking data-driven methods for model discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19746–19754, 2026.

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.

Jindou Chen, Jidong Tian, Liang Wu, Xiaokang Yang, Yaohui Jin, Yanyan Xu, et al. Kinformer: Generalizable dynamical symbolic regression for catalytic organic reaction kinetics. In *The Thirteenth International Conference on Learning Representations*, 2025.

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.

Fabricio O de Franca, Marco Virgolin, M Kommenda, MS Majumder, M Cranmer, G Espada, L Ingelse, A Fonseca, M Landajuela, B Petersen, et al. Srbench++: principled benchmarking of symbolic regression with domain-expert interpretation. *IEEE transactions on evolutionary computation*, 2024.

David Fell and Athel Cornish-Bowden. *Understanding the control of metabolism*, volume 2. Portland press London, 1997.

James E Ferrell, Tony Yu-Chen Tsai, and Qiong Yang. Modeling the cell cycle: why do certain circuits oscillate? *Cell*, 144(6):874–885, 2011.

Naama Geva-Zatorsky, Nitzan Rosenfeld, Shalev Itzkovitz, Ron Milo, Alex Sigal, Erez Dekel, Talia Yarnitzky, Yuvalal Liron, Paz Polak, Galit Lahav, et al. Oscillations and variability in the p53 system. *Molecular systems biology*, 2, 2006.

Albert Goldbeter. A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *Proceedings of the National Academy of Sciences*, 88(20):9107–9111, 1991.

Moritz Hoffmann, Christoph Fröhner, and Frank Noé. Reactive sindy: Discovering governing reactions from concentration data. *The Journal of chemical physics*, 150(2), 2019.

Samuel Holt, Zhaozhi Qian, Tennison Liu, James Weatherall, and Mihaela van der Schaar. Data-driven discovery of dynamical systems in pharmacology using large language models. *Advances in Neural Information Processing Systems*, 37:96325–96366, 2024.

Michael Hucka, Andrew Finney, Herbert M Sauro, Hamid Bolouri, John C Doyle, Hiroaki Kitano, Adam P Arkin, Benjamin J Bornstein, Dennis Bray, Athel Cornish-Bowden, et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

Guilherme Seidyo Imai Aldeia, Hengzhe Zhang, Geoffrey Bomarito, Miles Cranmer, Alcides Fonseca, Bogdan Burlacu, William G La Cava, and Fabrício Olivetti de França. Call for action: towards the next generation of symbolic regression benchmark. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 2529–2538, 2025.

Kenneth A Johnson. A century of enzyme kinetic analysis, 1913 to 2013. *FEBS letters*, 587(17):2753–2766, 2013.

Kadierdan Kaheman, J Nathan Kutz, and Steven L Brunton. Sindy-pi: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2242), 2020.

Pierre-Alexandre Kamienny, Stéphane d'Ascoli, Guillaume Lample, and François Charton. End-to-end symbolic regression with transformers. *Advances in Neural Information Processing Systems*, 35:10269–10281, 2022.

Boris N Kholodenko. Cell-signalling dynamics in time and space. *Nature reviews Molecular cell biology*, 7(3): 165–176, 2006.

Edda Klipp, Wolfram Liebermeister, Christoph Wierling, Axel Kowald, and Ralf Herwig. *Systems biology: a textbook*, volume 1. Wiley-VCH Weinheim, 2016.

John R Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4(2):87–112, 1994.

William La Cava, Lee Spector, and Kourosh Danai. Epsilon-lexicase selection for regression. In *Proceedings of the genetic and evolutionary computation conference 2016*, pp. 741–748, 2016.

William La Cava, Bogdan Burlacu, Marco Virgolin, Michael Kommenda, Patryk Orzechowski, Fabrício Olivetti de França, Ying Jin, and Jason H Moore. Contemporary symbolic regression methods and their relative performance. *Advances in neural information processing systems*, 2021(DB1):1, 2021.

Jean-Christophe Leloup, Didier Gonze, and Albert Goldbeter. Limit cycle models for circadian rhythms based on transcriptional regulation in drosophila and neurospora. *Journal of biological rhythms*, 14(6):433–448, 1999.

Yunlun Li and Sinno Jialin Pan. Robust equation structure learning with adaptive refinement. In *International Conference on Learning Representations*, volume 2026, pp. 70029–70053, 2026.

Zewen Liu, Juntong Ni, Xianfeng Tang, Max SY Lau, Wenpeng Yin, and Wei Jin. Can large language models adequately perform symbolic reasoning over time series? *arXiv preprint arXiv:2508.03963*, 2025.

Christoph Lüders, Thomas Sturm, and Ovidiu Radulescu. Odebase: a repository of ode systems for systems biology. *Bioinformatics Advances*, 2(1):vbac027, 2022.

Rahuman S Malik-Sheriff, Mihai Glont, Tung VN Nguyen, Krishna Tiwari, Matthew G Roberts, Ashley Xavier, Manh T Vu, Jinghao Men, Matthieu Maire, Sarubini Kananathan, et al. Biomodels—15 years of sharing computational models in life science. *Nucleic acids research*, 48(D1):D407–D415, 2020.

Yoshitomo Matsubara, Naoya Chiba, Ryo Igarashi, and Yoshitaka Ushiku. Rethinking symbolic regression datasets and benchmarks for scientific discovery. *arXiv preprint arXiv:2206.10540*, 2022.

Trent McConaghy. Ffx: Fast, scalable, deterministic symbolic regression technology. In *Genetic Programming Theory and Practice IX*, pp. 235–260. Springer, 2011.

Hongyu Miao, Joseph A Hollenbaugh, Martin S Zand, Jeanne Holden-Wiltse, Tim R Mosmann, Alan S Perelson, Hulin Wu, and David J Topham. Quantifying the early immune response and adaptive immune response kinetics in mice infected with influenza a virus. *Journal of virology*, 84(13):6687–6698, 2010.

Philip Munz, Ioan Hudea, Joe Imad, and Robert J Smith. When zombies attack!: mathematical modelling of an outbreak of zombie infection. *Infectious disease modelling research progress*, 4:133–150, 2009.

Manuel Palma Banos, Joel D Kress, Rigoberto Hernandez, and Galen T Craven. Stoichiometrically-informed symbolic regression for extracting chemical reaction mechanisms from data. *Digital Discovery*, 5(5):2325–2341, 2026.

Linda Petzold. Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM journal on scientific and statistical computing*, 4(1):136–148, 1983.

Heng Rao, Jason Zipeng Zhang, Yu Gu, Zhenghao Liu, Ge Yu, Jeffrey Su, Yang Cao, Fan Yang, and Minghan Chen. A regime-aware trajectory prediction framework for 1000+ systems biology models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=sI3UUkXJxs.

Alexander V Ratushny, Ramsey A Saleem, Katherine Sitko, Stephen A Ramsey, and John D Aitchison. Asymmetric positive feedback loops reliably control biological responses. *Molecular Systems Biology*, 8:577, 2012.

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

Michael Schmidt and Hod Lipson. Age-fitness pareto optimization. In *Genetic programming theory and practice VIII*, pp. 129–146. Springer, 2010.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*, 2024.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*, 2025.

Gábor Szederkényi, Julio R Banga, and Antonio A Alonso. Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC systems biology*, 5(1):177, 2011.

John J Tyson. Modeling the cell division cycle: cdc2 and cyclin interactions. *Proceedings of the National Academy of Sciences*, 88(16):7328–7332, 1991.

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

Martin Vastl, Jonáš Kulhánek, Jiří Kubalík, Erik Derner, and Robert Babuška. Symformer: End-to-end symbolic regression using transformer-based architecture. *IEEE Access*, 12:37840–37849, 2024.

James N Weiss. The hill equation revisited: uses and misuses. *The FASEB Journal*, 11(11):835–841, 1997.

Samuel WK Wong, Shihao Yang, and SC Kou. Estimating and assessing differential equation models with time-course data. *The Journal of Physical Chemistry. B*, 127(11):2362, 2023.

Shijie Xia, Yuhan Sun, and Pengfei Liu. Sr-scientist: Scientific equation discovery with agentic ai. *arXiv preprint arXiv:2510.11661*, 2025.

Jin Xu. Sbmlkinetics: a tool for annotation-independent classification of reaction kinetics for sbml models. *BMC bioinformatics*, 24(1):248, 2023.

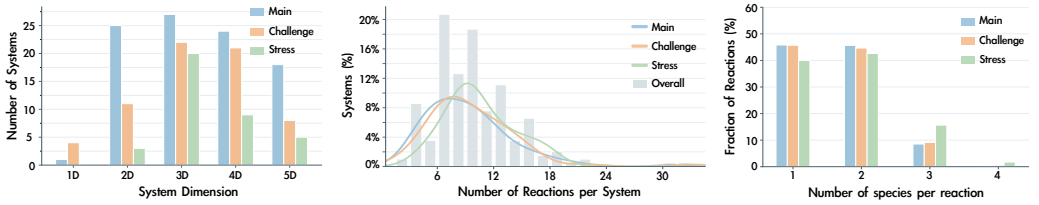
An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A SysBio-SR Benchmark and Evaluation Details

A.1 Benchmark Composition

Figure 7 summarizes system size and reaction structure across all 198 systems. State dimension and reaction count span overlapping ranges across the three subsets (Figure 7(a,b)). The subsets distinguish kinetic form while retaining variation in system size. In Figure 7(c), each species counts once across reactants, products, and the reaction-rate expression, including regulators. Percentages are normalized over all reactions within each subset.

The kinetic-family term counts appear in Figure 3(b). Basal denotes zero-order terms k, Mass-Action Linear denotes first-order terms kx_i, and Mass-Action Nonlinear denotes higher-order monomials $k\prod_i x_i^{a_i}$ of total degree at least two. The remaining families are Michaelis–Menten, Hill, and other nonlinear terms.



(a) System Dimension Distribution (b) System Reaction Distribution (c) Species-per-Reaction Distribution

Figure 7: **Structural composition of SysBio-SR.** (a) Number of systems by state dimension and subset. (b) Distribution of reactions per system; gray bars show the overall distribution, and colored curves show the distributions within each subset. (c) Distribution of the number of species involved per reaction, by subset.

A.2 Benchmark Organization and Evaluation Scope

To pin down what enters each evaluation, we organize the release around the three kinetics-based subsets of Section 3.2 and a per-system file schema. The three subsets contain 95 Main, 66 Challenge, and 37 Stress systems, totaling 198 systems. All 198 systems have reaction annotations and enter trajectory evaluation. The reaction-evaluation subset comprises the 161 Main and Challenge systems. The 37 Stress systems are excluded from reaction scoring.

The per-system release format is defined as follows:

```
benchmark/
|-- benchmark.yaml
|-- main/                       (95 systems)
|   |-- BIOMD0000000006/        (example: Tyson1991)
|   |   |-- Tyson1991.xml
|   |   |-- equation.json
|   |   |-- model.py
|   |   |-- reaction_annotation.json
|   |   `-- traj/
|   |       |-- traj_default.csv
|   |       |-- traj_perturbed.csv
|   |       `-- trajectory_metadata.json
|   `-- ...
|-- challenge/                  (66 systems)
`-- stress/                     (37 systems)
```

Each system includes its source SBML file (Hucka et al., 2003; Xu, 2023). `equation.json` provides the standardized ODEs, and `reaction_annotation.json` stores their reaction-level reference annotations. Appendix A.5 describes annotation construction, kinetic components, and signature matching. `model.py` provides an executable Python implementation of the reference

ODEs. The `traj/` directory holds the default and perturbed trajectories used for ID reconstruction and OOD extrapolation. The expanded branch shows BIOMD0000000006 (Tyson, 1991), a two-state system released with trajectories sampled at 512 time points.

Stress annotations include complex rational rate laws and other kinetic forms outside the current reaction-matching protocol. These annotations remain available even though Stress systems do not enter reaction-metric aggregates.

A.3 ID Reconstruction, OOD Extrapolation, and Observation Noise

To separate what varies between fitting and evaluation, we fix the fitting trajectory, the evaluation trajectory, and the corruption applied in the noise study.

Trajectory generation. Each system m uses a system-specific time horizon $[0, T_m]$ with 512 uniformly spaced observations. The default and perturbed trajectories share the same time grid and differ only in their initial conditions. System-specific horizons are provided with the released benchmark, together with a standalone reference simulator using LSODA (Petzold, 1983).

ID reconstruction. The ID condition evaluates the recovered system against the clean default trajectory of the same system. In the noise-free setting this is the trajectory used for fitting. In the noise study, the fitting trajectory is corrupted first, and evaluation remains on the clean default trajectory.

OOD extrapolation. The OOD condition shifts the initial condition while retaining the same vector field, parameter setting, and evaluation time span. The recovered system is evaluated against the trajectory generated from the perturbed initial condition. The two conditions therefore differ only in the evaluation initial condition.

Observation noise. To model imperfect measurement, we corrupt only the fitting trajectory with state-scaled Gaussian noise at relative levels $\sigma_{\mathrm{rel}} \in \{0.001, 0.01, 0.05, 0.1\}$ (Equation 4), where the per-state scaling uses the temporal standard deviation of that state. We leave the initial observation unchanged so that noise does not alter the fitting initial condition. After perturbation, we set negative concentrations to zero. We use noise seed 2026 and score all recovered systems against the clean default and clean perturbed trajectories. We label the clean setting **Noise-free** and the corrupted settings by their **relative noise level** σ_{rel}.

We use one noise realization per level to assess sensitivity to observation noise.

A.4 Metrics, System-Level Averaging, and Failures

To define the reported aggregates, we first compute each metric at the system level.

System-level averaging. We compute each metric separately for each system and report the unweighted mean across the fixed target set of the corresponding evaluation (198 systems for trajectory metrics, 161 for reaction metrics).

Trajectory metrics. For system m, let $x_{m,j}(t_n)$ and $\widehat{x}_{m,j}(t_n)$ denote the reference and recovered values of state j at the N_m evaluation times. We compute the state-wise scores and their equal-weight system average as

$$R^2_{m,j} = 1 - \frac{\sum_{n=1}^{N_m} \left[x_{m,j}(t_n) - \widehat{x}_{m,j}(t_n)\right]^2}{\sum_{n=1}^{N_m} \left[x_{m,j}(t_n) - \bar{x}_{m,j}\right]^2}, \qquad R^2_m = \frac{1}{d_m} \sum_{j=1}^{d_m} R^2_{m,j}, \tag{11}$$

where d_m is the number of states and $\bar{x}_{m,j} = N_m^{-1} \sum_{n=1}^{N_m} x_{m,j}(t_n)$. The state-wise fraction applies to nonconstant reference states.

For a constant reference state, an exactly matching prediction receives $R^2 = 1$; otherwise, we mark the state-wise R^2 as undefined. We assess trajectory fidelity with the coverage-aware R^2-score of Equation 3: negative per-system R^2 values and failed or non-finite simulations contribute zero, and

Acc@.90 is the fraction of target systems with $R^2 \geq 0.90$. We report expression complexity as a cross-method evaluation metric. Section 5.3 defines the parsimony criteria used for FPMS model selection. Recovered equations that cannot be parsed are reported separately from the complexity metric.

Reaction metrics. We operationalize mechanism recovery through exact reaction-signature matching (Section A.5). The three reaction-eligible methods listed in Section 4 contribute reaction metrics. For system m, let TP_m count matched signatures, FP_m unmatched predicted signatures, and FN_m unmatched reference signatures. Per-system precision, recall, and F1 follow as

$$P_m = \frac{\mathrm{TP}_m}{\mathrm{TP}_m + \mathrm{FP}_m}, \qquad R_m = \frac{\mathrm{TP}_m}{\mathrm{TP}_m + \mathrm{FN}_m}, \qquad F_{1,m} = \frac{2P_m R_m}{P_m + R_m}, \tag{12}$$

and we report their unweighted means over the $M = 161$ reaction-scored systems:

$$\bar{P} = \frac{1}{M}\sum_{m=1}^{M} P_m, \qquad \bar{R} = \frac{1}{M}\sum_{m=1}^{M} R_m, \qquad \bar{F}_1 = \frac{1}{M}\sum_{m=1}^{M} F_{1,m}. \tag{13}$$

An eligible method with an empty predicted set scores $P_m = R_m = F_{1,m} = 0$ (Section A.5).

Kinetic-family metrics. Each predicted and reference record receives one diagnostic family label. A record containing a Hill component is labeled Hill. Otherwise a record containing a Michaelis–Menten component is labeled Michaelis–Menten. All remaining records are labeled mass action. True positives, false positives, and false negatives are pooled per label across the 161 systems, and per-label precision, recall, and F1 follow from the pooled counts. Family metrics therefore use label-level pooling rather than the system-level averaging of the reaction metrics above. This unit of counting differs from the reference-equation terms of Figure 3(b).

A.5 Reaction Annotations and Matching

To compare recovered and reference reactions, we canonicalize both sides into term-projected, unit-stoichiometric reaction signatures and match them exactly. Exact matching means equality over the signature fields defined below. This subsection defines the construction and matching rules summarized in Section 4.

Annotation construction. We construct reference annotations from the standardized ODE expressions. For each state equation, we decompose the RHS into nonzero additive terms and separate each term's sign, coefficient, and kinetic factor. A negative term defines a consuming record for the target species, whereas a positive term defines a producing record. We identify the species-dependent factors and assign their kinetic-law classes and associated parameters. Each equation-level record retains its originating equation index. The annotations also provide a reaction-level representation that consolidates equation-level contributions with shared kinetics, as illustrated by the nonlinear example below.

Annotation representation. Each reaction is stored as a structured record containing its stoichiometry, rate coefficient, kinetic factor, and species-dependent kinetic components. For example, $x_0 \to \varnothing$ with $v = 0.25\, x_0 x_2/(0.02 + x_0)$ is represented by the following abridged record:

```
{
  "reactants": [1, 0, 0, 0, 0],
  "products":  [0, 0, 0, 0, 0],
  "coefficient": 0.25,
  "kinetic_factor": "x_0*x_2/(x_0 + 0.02)",
  "substrates": [
    {"species_idx": 0, "law": "mm", "order": 1, "K": 0.02}
  ],
  "modifiers": [
    {"species_idx": 2, "law": "linear", "order": 1}
  ]
}
```

Here, `substrates` denote reactant-dependent kinetic components, whereas `modifiers` denote regulatory dependencies.

For the 161 reaction-scored systems, the component vocabulary includes mass action (x^a), monomial dependence (x^q), Michaelis–Menten saturation ($x/(K+x)$), and Hill-type activation and inhibition. In the examples, monomial dependence is written as $\mathrm{Mono}(q)$, corresponding to the label `linear`.

Annotation checks. We reconstruct the RHS expressions from the annotated contributions and compare them with the reference ODEs using symbolic and numerical checks. We also verify that the annotated kinetic components reproduce the corresponding kinetic factors. All annotations are subsequently manually reviewed for consistency with the reference ODEs and kinetic forms.

A projection example. Consider the single reaction $2A + B \to C$ with mass-action rate $v = k\,x_A^2 x_B$. The reaction expands into the coupled ODE terms

$$2A + B \xrightarrow{v} C, \qquad v = k\,x_A^2 x_B, \qquad \begin{bmatrix} \dot{x}_A \\ \dot{x}_B \\ \dot{x}_C \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix} k\,x_A^2 x_B.$$

All three terms share the same rate v and originate from this single reaction. Term projection produces one signature per affected state equation. The arrows below encode unit consumption or production of the target species; the remaining columns record the corresponding kinetic components. $\mathrm{MA}(q)$ denotes a mass-action reactant-dependent term of order q. $\mathrm{Mono}(q)$ denotes a monomial kinetic dependency x^q. Kinetic roles are assigned relative to each projected signature.

Equation term	Unit-projected stoichiometry	Reactant-dependent term	Regulatory term(s)
$-2kx_A^2x_B$ in $\dot{x}_A$	$A \to \varnothing$	A: MA(2)	B: Mono(1)
$-kx_A^2x_B$ in $\dot{x}_B$	$B \to \varnothing$	B: MA(1)	A: Mono(2)
$+kx_A^2x_B$ in $\dot{x}_C$	$\varnothing \to C$	none	A: Mono(2), B: Mono(1)

The example shows three properties of the protocol. First, the stoichiometric coefficient 2 on A yields exactly one consuming signature for $\dot{x}_A$. Unit projection normalizes stoichiometric multiplicities to one, while the order 2 remains part of the mass-action component. Second, the kinetic roles depend on the target equation. In each consuming signature, dependence on the consumed species is reactant-dependent and the remaining dependence is regulatory. The producing signature represents both dependencies as regulatory terms, although A and B are reactants of the original reaction. The $\mathrm{Mono}(2)$ entries preserve the kinetic order after reassignment. Third, a recovered term $-k'x_A^2x_B$ in $\dot{x}_A$ matches the first signature even when its coefficient differs from k. A recovered term $-k'x_A$ does not match.

A nonlinear kinetic example. The released annotation of BIOMD0000000004 (Goldbeter, 1991), a five-state model over C, M, X, MI, and XI, contains 11 equation-level records, which are consolidated into seven reaction-level records. The table shows the reaction-level representation. $\mathrm{MM}(K)$ denotes a Michaelis–Menten reactant-dependent term with affinity K, and $\mathrm{Hill}^{+}(K, n)$ a Hill-activation regulatory term.

Reaction	Reactant-dependent term(s)	Regulatory term(s)	Kinetic form
$\mathrm{C} \to \varnothing$	C: MA(1)	none	$v = 0.01\, x_{\mathrm{C}}$
$\varnothing \to \mathrm{C}$	none	none	$v = 0.025$
$\mathrm{C} \to \varnothing$	C: MM(0.02)	X: Mono(1)	$v = 0.25\, \frac{x_{\mathrm{C}}}{0.02+x_{\mathrm{C}}}\, x_{\mathrm{X}}$
$\mathrm{X} \to \mathrm{XI}$	X: MM(0.005)	none	$v = 0.5\, \frac{x_{\mathrm{X}}}{0.005+x_{\mathrm{X}}}$
$\mathrm{XI} \to \mathrm{X}$	XI: MM(0.005)	M: Mono(1)	$v = 1.0\, \frac{x_{\mathrm{XI}}}{0.005+x_{\mathrm{XI}}}\, x_{\mathrm{M}}$
$\mathrm{M} \to \mathrm{MI}$	M: MM(0.005)	none	$v = 1.5\, \frac{x_{\mathrm{M}}}{0.005+x_{\mathrm{M}}}$
$\mathrm{MI} \to \mathrm{M}$	MI: MM(0.005)	C: Hill^{+}(0.5, 1)	$v = 3.0\, \frac{x_{\mathrm{MI}}}{0.005+x_{\mathrm{MI}}}\, \frac{x_{\mathrm{C}}}{0.5+x_{\mathrm{C}}}$

The 11 equation-level records and 7 reaction-level records yield the same signature multiset after unit projection. The final reaction-level record consolidates two equation-level records with the same rate coefficient and kinetic factor. The consuming record $\mathrm{MI} \to \varnothing$ retains the Michaelis–Menten dependence on MI as a reactant-dependent term and the Hill-activation dependence on C as a regulatory term. The producing record $\varnothing \to \mathrm{M}$ represents both dependencies as regulatory terms and represents the Michaelis–Menten factor as Hill activation at $n = 1$. At $n = 1$, Michaelis–Menten saturation and Hill activation share the form $x/(K + x)$.

Reference reactions. For each of the 161 reaction-scored systems, the annotation records the reference mechanism as a list of reactions. Each reaction carries reactants and products as per-species integer counts, a rate coefficient, and a list of kinetic components. Each kinetic-term entry records its species index, kinetic law, order, and, where applicable, an affinity constant K and a Hill exponent n. The component law vocabulary is: mass action (x^a), monomial dependence (x^q), Michaelis–Menten saturation ($x/(K + x)$), Hill activation ($x^n/(K^n + x^n)$), Hill inhibition ($K^n/(K^n + x^n)$), and Hill saturation (a Hill factor in a reactant-dependent role).

Signature construction. We canonicalize each reaction into one equation-level signature for every state with a nonzero net stoichiometric coefficient. Let $\mathbf{s}_r = \mathbf{s}_r^{+} - \mathbf{s}_r^{-}$ denote the net stoichiometric vector of reaction r. For each state x_j with $s_{r,j} \neq 0$, we project $s_{r,j}$ to its sign, retaining whether the reaction consumes or produces x_j while discarding the magnitude of the stoichiometric coefficient. Kinetic dependence on a consumed target species is represented by a reactant-dependent term. All remaining kinetic dependencies are represented as regulatory terms. For a production signature, all kinetic dependencies are regulatory. We canonicalize regulatory forms while preserving kinetic orders: Michaelis–Menten saturation becomes Hill activation with $n = 1$, and Hill saturation reassigned from a reactant-dependent role becomes Hill activation.

Matching. Each signature contains unit reactant and product vectors and two sorted kinetic-term lists, one reactant-dependent and one regulatory. Each entry records its species index, kinetic law, and law-dependent details. Two signatures match exactly when all four parts agree. We discard predicted terms with fitted coefficient magnitude below 10^{-12} before signature construction. We use reference annotations as recorded. We pool signatures across source reactions and compute multiset agreement as $\mathrm{TP} = \sum_{\sigma} \min(\mathrm{true}_{\sigma}, \mathrm{pred}_{\sigma})$. Unmatched predicted and reference signatures count as false positives and false negatives. We compute and average per-system precision, recall, and F1 as defined in Section A.4.

We compare every component's kinetic-law class but exclude numerical rate coefficients, affinity constants K, and Hill exponents n. The score measures agreement between projected signatures. If an eligible method returns no model, or we cannot map its output to the reaction representation, we assign it an empty predicted multiset. Because all 161 reference multisets are nonempty, such cases receive zero precision, recall, and F1.

A.6 Composition of Benchmarks Compared in Related Work

Table 1 compares SysBio-SR with existing symbolic regression benchmarks. This subsection documents how the counts reported for each external benchmark are composed. Only instances posed as symbolic-regression tasks with ground-truth equation targets are counted.

AI Feynman. The AI Feynman database (Udrescu & Tegmark, 2020) contains 120 equations: 100 regression mysteries drawn from the Feynman Lectures on Physics plus a bonus set of 20 additional equations.

SRBench. The SRBench ground-truth suite (La Cava et al., 2021) comprises 130 datasets with known model forms, drawn from the Feynman Symbolic Regression Database and the ODE-Strogatz repository. Its Feynman portion contains 116 scalar equations. The underlying database contains 120 equations, but one is absent from the suite and three are excluded for containing arcsin or arccos terms. Its Strogatz portion poses one dataset per state variable of seven two-state ODE systems, giving 14 single-state datasets in total. Counting full systems rather than per-state datasets yields 123 systems (116 scalar and seven two-state) with 130 right-hand-side equations, as reported in Table 1. The remaining 122 black-box regression problems of SRBench have no ground-truth equation target and are excluded from the comparison.

EmpiricalBench. EmpiricalBench (Cranmer, 2023) contains nine historical empirical relations from the physical sciences, including Hubble's law, Kepler's third law, and the ideal gas law. Each instance poses one scalar target equation, so the reported system and equation counts both equal nine.

ODEBench. ODEBench (d'Ascoli et al., 2023) contains 63 ODE systems spanning state dimensions 1–4. The reported 117 right-hand-side equations are the implied sum over the dimension distribution of the suite.

LLM-SRBench. The LLM-SRBench paper (Shojaee et al., 2025) reports 239 equation-discovery tasks, each with one scalar target. The set comprises 111 LSR-Transform tasks that reformulate established physical models and 128 LSR-Synth tasks spanning chemistry, biology, physics, and materials science, so the reported system and equation counts both equal 239.

SymbolBench. SymbolBench (Liu et al., 2025) evaluates symbolic regression over coupled ODE systems. Its regression task extends ODEBench's 63 systems with additional systems, yielding 156 instances that span state dimensions 1–4. The reported 383 right-hand-side equations are the implied sum over this dimension distribution. Its Boolean-network and causal-discovery tasks are not equation-recovery tasks and are excluded from Table 1.

MDBench. MDBench (Bideh et al., 2026) combines 63 ODE systems (dimensions 1–4, contributing 117 right-hand-side equations) with 14 PDE systems (up to six state variables, contributing 24 right-hand-side equations). All 77 systems carry ground-truth dynamics, giving 141 equations in total.

B KINESELECT: ALGORITHMIC DETAILS

RSS proposes reaction structures, and KFI selects one kinetic realization per reaction during candidate evaluation. FPMS selects the final model when the search produces a finite simulation-based archive. The selected model is an executable reaction mechanism with fitted rate constants. Ground-truth reaction annotations and reaction-matching scores remain evaluation-only.

B.1 REACTION STRUCTURE SEARCH

KineSelect represents each proposed reaction by its reactant stoichiometry, product stoichiometry, and an optional regulator role. For reaction r, we write

$$g_r = \left(\mathbf{s}_r^-, \mathbf{s}_r^+, m_r\right), \qquad \mathbf{s}_r^-, \mathbf{s}_r^+ \in \mathbb{Z}_{\geq 0}^d, \tag{14}$$

where $\mathbf{s}_r^-$ and $\mathbf{s}_r^+$ encode reactants and products, and $m_r \in \{\varnothing, 0, \ldots, d-1\}$ identifies an optional regulator. We restrict each reaction gene to at most one proposed regulator to avoid combinatorial enumeration of regulator subsets during structural and kinetic search. When $m_r \neq \varnothing$, we require $[\mathbf{s}_r^-]_{m_r} = [\mathbf{s}_r^+]_{m_r} = 0$, so the regulator can affect the rate without entering the reactant or product stoichiometry. The net stoichiometric vector is $\mathbf{s}_r = \mathbf{s}_r^+ - \mathbf{s}_r^-$. For example, $x_0 \to x_1$ with rate $v(\mathbf{x}) = kx_0x_2$ assigns x_2 as a regulator. Its gene uses $\mathbf{s}_r^- = (1,0,0)^\top$, $\mathbf{s}_r^+ = (0,1,0)^\top$, and $m_r = 2$.

RSS follows SISR's reaction-set evolutionary framework (Palma Banos et al., 2026) and extends each reaction gene with the optional regulator role defined above. It maintains separate subpopulations for models with $|\mathcal{R}(\mathcal{M})| = c$, where $c \in \{1, \ldots, c_{\max}\}$. Each individual in subpopulation c contains exactly c reaction genes. Crossover pools the genes of two parents, removes duplicates, and randomly draws c genes to form an offspring. Mutation replaces one randomly selected gene with a gene from the pre-enumerated admissible reaction library. For example, with $c = 3$, parent mechanisms

$$\mathcal{R}(\mathcal{M}_1) = \{g_1, g_2, g_3\}, \qquad \mathcal{R}(\mathcal{M}_2) = \{g_2, g_4, g_5\}$$

form the deduplicated pool $\{g_1, g_2, g_3, g_4, g_5\}$, from which one possible offspring is $\{g_1, g_4, g_5\}$. A mutation may then replace g_4 with another admissible gene g_6. KFI assigns one kinetic realization to each reaction in every proposal and returns the derivative loss used to rank that proposal within its subpopulation. After pre-screening candidates by derivative loss, each reaction-count subpopulation retains the locally non-dominated candidates under three archive objectives: simulation loss, expanded ODE expression-tree complexity, and active-reaction count. The search merges these local sets, deduplicates candidates by reaction structure, and retains the globally non-dominated candidates under these archive objectives as the fixed archive $\mathcal{A}$. FPMS then re-scores $\mathcal{A}$ using the four fidelity and parsimony criteria defined in Section B.3 without modifying its candidates. If the search cannot form a finite simulation-based archive, it uses a predefined derivative-based fallback to return a model.

B.2 KINETIC LIBRARY AND KINETIC FORM IDENTIFICATION

KFI separates kinetic-candidate selection from overall rate-constant estimation. For each proposed reaction, it constructs unit-rate kinetic candidates by pairing a reactant-dependent kinetic term with a regulatory term:

$$\bar{v}_{r,(a,b)}(\mathbf{x}) = \phi_{r,a}\left(\mathbf{x}; \mathbf{s}_r^-\right) \psi_{r,b}\left(\mathbf{x}; m_r\right). \tag{15}$$

Here (a, b) indexes an admissible term pair, and a single index ℓ enumerates the resulting set $\mathcal{V}_r = \{\bar{v}_{r,\ell}\}$. For a fixed proposed mechanism $\mathcal{M}$, directly encoding one kinetic candidate in every reaction gene would admit $\prod_{r\in\mathcal{R}(\mathcal{M})} |\mathcal{V}_r|$ joint kinetic assignments. KFI avoids their explicit evolutionary enumeration by representing the alternatives with $\sum_{r\in\mathcal{R}(\mathcal{M})} |\mathcal{V}_r|$ dictionary columns, fitting them jointly, and selecting one candidate per reaction group. For each species x_q, the canonical configuration derives a finite candidate set $\mathcal{K}_q$ from the 0.05 and 0.50 quantiles of its positive fitting-trajectory values. When no positive observations are available, we use the fallback grid recorded in the full configuration. Hill candidates use $n \in \{1, 2, 4\}$, and each admissible component setting defines a distinct kinetic choice. KFI then refits only the overall rate constant:

$$v_r(\mathbf{x}) = k_r \bar{v}_{r,\ell_r^\star}(\mathbf{x}), \qquad \dot{\mathbf{x}} = \sum_{r\in\mathcal{R}(\mathcal{M})} \mathbf{s}_r k_r \bar{v}_{r,\ell_r^\star}(\mathbf{x}). \tag{16}$$

Table 4: **Kinetic terms used to construct KFI candidates.** KFI pairs a reactant-dependent kinetic term with a regulatory term. Here $a_{r,q} = [\mathbf{s}_r^-]_q$, $m = m_r$, and $\mathcal{K}_q$ denotes the finite species-specific candidate set for x_q.

Term type	Law	Form	Availability
Reactant-dependent	Mass action	$\phi_r^{\mathrm{MA}}(\mathbf{x}) = \prod_{q=0}^{d-1} x_q^{a_{r,q}}$	Available for every reaction; it equals 1 for a source reaction.
Reactant-dependent	Michaelis–Menten	$\phi_{r,j,K}^{\mathrm{MM}}(\mathbf{x}) = x_j^{a_{r,j}-1} \frac{x_j}{K + x_j} \prod_{q \neq j} x_q^{a_{r,q}}$	Available when $a_{r,j} > 0$ and $K \in \mathcal{K}_j$. Unavailable for source reactions.
Regulatory	None	$\psi^{\mathrm{none}}(\mathbf{x}) = 1$	Available for every reaction.
Regulatory	Linear	$\psi^{\mathrm{lin}}(\mathbf{x}; m) = x_m$	Requires m.
Regulatory	Hill activation	$\psi_{K,n}^{\mathrm{act}}(\mathbf{x}; m) = \frac{x_m^n}{K^n + x_m^n}$	Requires m, $K \in \mathcal{K}_m$, and $n \in \{1, 2, 4\}$.
Regulatory	Hill inhibition	$\psi_{K,n}^{\mathrm{inh}}(\mathbf{x}; m) = \frac{K^n}{K^n + x_m^n}$	Requires m, $K \in \mathcal{K}_m$, and $n \in \{1, 2, 4\}$.

Table 4 lists the admissible reactant-dependent and regulatory terms used to construct these candidates, together with their availability conditions.

KFI builds the scale-normalized, stoichiometrically coupled dictionary in Eq. (6) and solves the joint regularized fit in Eq. (7). This system-level fit lets kinetic alternatives compete in the context of the complete proposed mechanism. A regulator can change the temporal amplitude through $\bar{v}_{r,\ell}(\mathbf{X})$, while its zero stoichiometric entry prevents a direct derivative contribution to the regulator state.

Before normalization, we replace near-zero derivative scales with a safe value and set near-zero dictionary columns to zero. Before the joint fit, we prune nearly collinear candidates within each reaction group, retaining simpler representatives. The full configuration records the cosine thresholds and numerical tolerances.

KFI computes each retained candidate's kinetic contribution score from its fitted weight $\widehat{w}_{r,\ell}$ obtained in Eq. (7). For nondegenerate unit-normalized columns, non-negative fitted weights, and $\|\mathbf{y}\|_2 > 0$,

$$\eta_{r,\ell} = \frac{\|\mathbf{a}_{r,\ell}\widehat{w}_{r,\ell}\|_2}{\|\mathbf{y}\|_2} = \frac{\widehat{w}_{r,\ell}}{\|\mathbf{y}\|_2}, \qquad \arg\max_{\ell} \eta_{r,\ell} = \arg\max_{\ell} \widehat{w}_{r,\ell}. \tag{17}$$

Thus, $\eta_{r,\ell}$ is a target-normalized fitted amplitude that yields the same maximizing-index set as $\widehat{w}_{r,\ell}$. KFI selects the highest-scoring kinetic candidate in each reaction group and refits the overall rate constants $\widehat{k}_r$ using the selected kinetic forms. Algorithm 1 summarizes the procedure.

Algorithm 1 Kinetic Form Identification (KFI)

Input: Proposed reaction model $\mathcal{M}$, fitting trajectory $\mathbf{X}$, derivative estimate $\dot{\mathbf{X}}$, kinetic-library configuration, and regularization coefficient λ

Output: Kinetic-resolved model $\widehat{\mathcal{M}}$ and derivative loss $\mathcal{L}^{\mathrm{der}}$

Construct Γ and $\mathbf{y} = \mathrm{vec}(\dot{\mathbf{X}}\Gamma^{-1})$
for all $r \in \mathcal{R}(\mathcal{M})$ **do**
    Enumerate the admissible unit-rate candidates $\mathcal{V}_r$
    Construct the normalized columns $\{\mathbf{a}_{r,\ell}\}_\ell$
end for
Prune nearly collinear columns within each group and update $\{\mathcal{V}_r\}_r$
Assemble $\mathbf{\Theta}(\mathcal{M})$ and solve the joint fit in Eq. (7)
for all $r \in \mathcal{R}(\mathcal{M})$ **do**
    Compute $\eta_{r,\ell}$ for all $\ell \in \{1, \ldots, |\mathcal{V}_r|\}$
    Select $\ell_r^\star$ as the first maximizer in Eq. (17) under the fixed kinetic-library order
end for
Jointly refit the selected unnormalized unit-rate columns to obtain $\{\widehat{k}_r\}_{r\in\mathcal{R}(\mathcal{M})}$
Form $\widehat{\mathcal{M}}$ and compute $\mathcal{L}^{\mathrm{der}}$
return $\widehat{\mathcal{M}}$ and $\mathcal{L}^{\mathrm{der}}$

The fitted weights $\widehat{\mathbf{w}}$ guide kinetic-candidate selection, while the refitted rate constants $\{\widehat{k}_r\}_r$ define the executable ODE.

B.3 Fidelity–Parsimony Model Selection

α-threshold selection. We use α-threshold selection as a deterministic baseline for archive-level model selection. It operationalizes the loss–complexity principle used by SISR (Palma Banos et al., 2026). This study defines the threshold rule for the controlled selector comparison.

For a fixed candidate archive $\mathcal{A}$, let $\mathcal{L}^{\mathrm{sim}}_{\min} = \min_{\mathcal{M}_i\in\mathcal{A}} \mathcal{L}^{\mathrm{sim}}_i$. The rule retains

$$\mathcal{A}_\alpha = \left\{\mathcal{M}_i \in \mathcal{A} : \mathcal{L}^{\mathrm{sim}}_i \le \alpha \mathcal{L}^{\mathrm{sim}}_{\min}\right\}. \tag{18}$$

Here, $\mathcal{L}^{\mathrm{sim}}_i$ compares the simulated fitting trajectory with the observed trajectory. We set $\alpha = 1.05$, which allows up to a 5% increase above the minimum simulation loss. We chose this default after inspecting SISR's reaction F1 and OOD R^2 on this benchmark. The implementation adds $\epsilon = 10^{-12}$ to the right-hand side for numerical comparison. Within this set, the selector chooses the model with the lowest expanded ODE expression-tree complexity C^{tree}_i, which FPMS also uses as its tree-parsimony criterion.

FPMS. FPMS scores complete mechanisms in the shared archive $\mathcal{A}$. For each candidate $\mathcal{M}_i \in \mathcal{A}$, it uses derivative loss $\mathcal{L}^{\mathrm{der}}_i$ and simulation loss $\mathcal{L}^{\mathrm{sim}}_i$. Simulation loss integrates each fitted candidate from the first observed state over the fitting time grid and compares the solution with the same observed trajectory $\mathbf{X}$. Failed or non-finite simulations receive infinite loss.

A reaction is active when $|\widehat{k}_{i,r}| > 10^{-14}$. Expanded ODE expression-tree complexity C^{tree}_i follows the SISR-inspired node count after expanding each active reaction into every affected state equation (Palma Banos et al., 2026). It includes sign, stoichiometric, rate-constant, kinetic-expression, and within-equation addition nodes. Specialized kinetic-component count C^{spec}_i separately counts Michaelis–Menten reactant-dependent terms and Hill regulatory terms.

For example, an active mass-action reaction $x_0 \to x_1$ with rate $v(\mathbf{x}) = kx_0$ expands to

$$\dot{x}_0 = -kx_0, \qquad \dot{x}_1 = +kx_0.$$

Each term contributes four nodes: one sign node, one unit-stoichiometry node, one rate-constant node, and one kinetic node. Because each equation contains only one active term, the count includes no addition node, giving $C^{\mathrm{tree}}_i = 8$ and $C^{\mathrm{spec}}_i = 0$. An active reaction combining a Michaelis–Menten reactant-dependent term with a Hill regulatory term contributes 2 to C^{spec}_i.

The criteria overlap because Michaelis–Menten and Hill components also contribute to C^{tree}_i. The tree count covers the expanded state equations, whereas C^{spec}_i counts each active specialized kinetic component once, irrespective of the number of affected state equations.

FPMS maps negative or non-finite losses to $+\infty$ before applying the log transform:

$$\mathcal{L}_i^{\mathrm{der}} \leftarrow \log\left(\mathcal{L}_i^{\mathrm{der}} + \varepsilon\right), \qquad \mathcal{L}_i^{\mathrm{sim}} \leftarrow \log\left(\mathcal{L}_i^{\mathrm{sim}} + \varepsilon\right), \qquad \varepsilon = 10^{-12}. \tag{19}$$

FPMS independently min–max normalizes the transformed losses and the two complexity criteria over $\mathcal{A}$. If the finite values of a criterion have zero range, they map to zero and FPMS marks the criterion as degenerate. Non-finite entries map to one whenever the criterion has a finite value. If no finite value exists, FPMS also marks the criterion as degenerate and maps all entries to zero.

Tildes on $\mathcal{L}_i^{\mathrm{der}}$ and $\mathcal{L}_i^{\mathrm{sim}}$ denote the normalized values of the log-transformed losses. FPMS forms a lack-of-fidelity score Q_i and a model-complexity score P_i:

$$Q_i = \frac{\widetilde{\mathcal{L}}_i^{\mathrm{der}} + \widetilde{\mathcal{L}}_i^{\mathrm{sim}}}{2}, \qquad P_i = \frac{\widetilde{C}_i^{\mathrm{tree}} + \widetilde{C}_i^{\mathrm{spec}}}{2}, \tag{20}$$

where $Q_i, P_i \in [0, 1]$. Lower Q_i and P_i indicate higher fidelity and lower model complexity, respectively.

FPMS selects the final model by maximizing the multiplicative utility:

$$U_i = (1 - Q_i)(1 - P_i), \qquad \mathcal{M}^\star = \operatorname*{arg\,max}_{\mathcal{M}_i \in \mathcal{A}} U_i. \tag{21}$$

FPMS resolves ties in U_i by minimizing Q_i and then selecting the lexicographically smallest canonical candidate identifier.

C SUPPLEMENTARY EXPERIMENTAL RESULTS

This appendix extends the joint trajectory and mechanism-recovery analysis in Section 6. We examine mechanism recovery at high OOD trajectory fidelity, kinetic-family recovery, final-model selection, and OOD performance distributions under observation noise. We then analyze representative recovery and failure cases.

C.1 MECHANISM RECOVERY AMONG HIGH-FIDELITY PREDICTIONS

OOD $R^2 \geq 0.90$ does not guarantee high reaction F1 (Table 5). Among the 65 KineSelect predictions with OOD $R^2 \geq 0.90$, 18 systems (27.7%) have reaction F1 below 0.50. For SISR and Reactive SINDy predictions with OOD $R^2 \geq 0.90$, the corresponding counts are 17 of 56 systems (30.4%) and 7 of 16 systems (43.8%), respectively.

KineSelect reaches OOD $R^2 \geq 0.90$ on more reaction-eligible systems than SISR: 65 of 161 (40.4%) versus 56 of 161 (34.8%). Among predictions with OOD $R^2 \geq 0.90$, mean reaction F1 is similar for KineSelect and SISR, at 0.658 and 0.653, respectively; their median F1 values are 0.714 and 0.727.

Table 5: **Mechanism recovery for predictions with OOD $R^2 \geq 0.90$.** Coverage gives the number and percentage of the 161 reaction-eligible systems with OOD $R^2 \geq 0.90$. For each method, we report mean and median reaction F1, together with the count and percentage of systems with F1 < 0.50, among systems with OOD $R^2 \geq 0.90$.

Method	**Coverage**	**Mean F1**	**Median F1**	**F1** < 0.50
Reactive SINDy	16 (9.9%)	0.552	0.519	7 (43.8%)
SISR	56 (34.8%)	0.653	0.727	17 (30.4%)
KineSelect	65 (40.4%)	0.658	0.714	18 (27.7%)

C.2 KINETIC-FAMILY RECOVERY AND REFERENCE SUPPORT

Mass-action records dominate the kinetic-family evaluation (Table 6). Mass action appears in 158 systems with 1,309 reference records, compared with 32 systems and 47 records for Michaelis–Menten kinetics. Hill kinetics account for 163 records across 61 systems.

KineSelect has the highest mass-action precision, recall, and F1, but its nonlinear-family scores expose different recovery difficulties. For Michaelis–Menten kinetics, precision is 0.034 while recall is 0.213, indicating many unmatched predictions in addition to missed reference signatures. Hill recovery has low precision and recall, at 0.021 and 0.031. In the evaluated configurations, Reactive SINDy and SISR lack these nonlinear components (Appendix D.2). KineSelect represents both component types, yet their F1 scores remain 0.059 and 0.025, compared with 0.503 for mass action. Improving nonlinear kinetic recovery requires addressing the low precision of Michaelis–Menten predictions and the low precision and recall of Hill predictions.

C.3 SENSITIVITY AND ABLATIONS OF FINAL-MODEL SELECTION

Sensitivity to α. The preferred threshold depends on both the candidate archive and the evaluation metric (Figure 8). We compare α-threshold selection and FPMS on a fixed candidate archive from each search method. Appendix B.3 defines the threshold rule and the default $\alpha = 1.05$. For SISR, F1 rises to about 0.411 at $\alpha \approx 2$–5, then generally declines as the selected expressions become simpler. OOD R^2 peaks at 0.434 with $\alpha = 1.5$. FPMS exceeds both tested maxima, reaching F1 of 0.431 and OOD R^2 of 0.440.

KineSelect candidates gain more reaction F1 from relaxing α, with the highest tested F1 occurring at $\alpha = 20$. From the default $\alpha = 1.05$ to $\alpha = 20$, F1 rises from 0.323 to 0.417, while OOD R^2 decreases from 0.598 to 0.527. Mean expression complexity falls from 125.16 to 61.10, and mean specialized kinetic-component count falls from 5.64 to 2.45. Larger thresholds continue to simplify the models but do not improve on the F1 at $\alpha = 20$. Thus, relaxing the fitting-loss threshold favors

Table 6: **Kinetic-family recovery and reference support.** The *Systems* column counts systems containing reference records of the family; the *Records* column counts projected unit-stoichiometry reference records with multiplicity. System supports can overlap. Precision, recall, and F1 use TP, FP, and FN counts pooled by kinetic-family label across the 161 reaction-eligible systems (Appendix A.4). Bold and underlined values denote the best and second-best results within each family, including ties.

Family	Systems	Records	Method	Precision↑	Recall↑	F1↑
Mass action	158	1309	Reactive SINDy	0.260	0.380	0.309
			SISR	<u>0.379</u>	<u>0.412</u>	<u>0.395</u>
			KineSelect	**0.552**	**0.461**	**0.503**
Michaelis–Menten	32	47	Reactive SINDy	<u>0.000</u>	<u>0.000</u>	<u>0.000</u>
			SISR	<u>0.000</u>	<u>0.000</u>	<u>0.000</u>
			KineSelect	**0.034**	**0.213**	**0.059**
Hill	61	163	Reactive SINDy	<u>0.000</u>	<u>0.000</u>	<u>0.000</u>
			SISR	<u>0.000</u>	<u>0.000</u>	<u>0.000</u>
			KineSelect	**0.021**	**0.031**	**0.025**

simpler models, but the thresholds that maximize reaction F1 differ across candidate archives and from those that maximize OOD R^2.

For KineSelect candidates, FPMS and $\alpha = 20$ yield mean expression complexities of 62.13 and 61.10, respectively, but F1 differs: 0.447 versus 0.417. FPMS yields lower OOD R^2, at 0.508 versus 0.527. FPMS avoids choosing a simulation-loss tolerance while achieving higher reaction F1 than every tested α setting on both candidate sets.

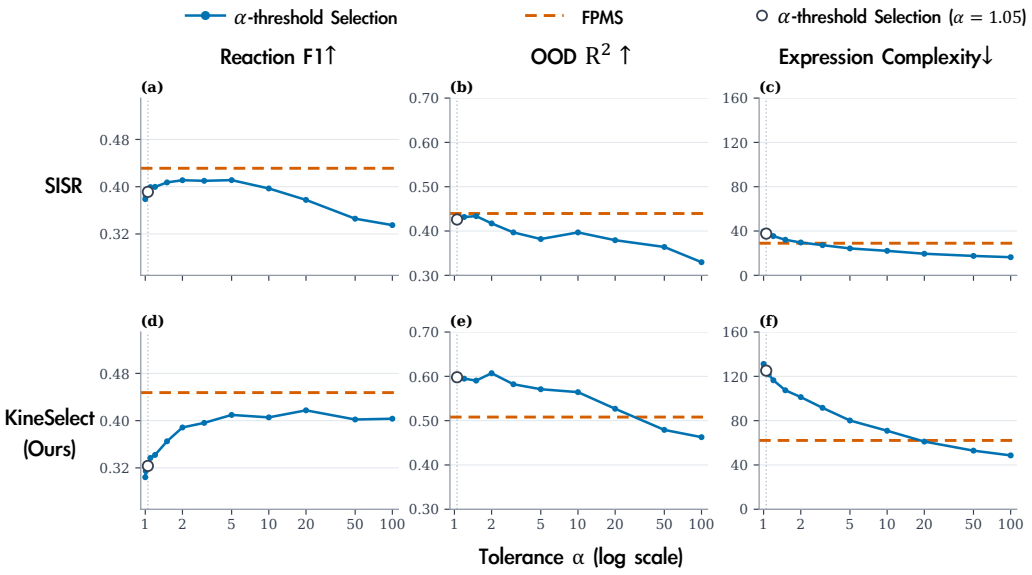


Figure 8: **Sensitivity of final-model selection to α.** The top and bottom rows use candidates from SISR and the KineSelect search, respectively. Columns report reaction F1 (a,d), OOD R^2 (b,e), and benchmark expression complexity (c,f). Solid curves show 13 α settings from 1 to 100 on a log axis, dashed lines show FPMS, and outlined circles mark the default $\alpha = 1.05$. Trajectory and complexity means use 198 systems, and reaction F1 uses 161.

FPMS component ablations. Full FPMS achieves the highest reaction F1 on both Main and Challenge (Table 7). Adding derivative loss or specialized kinetic-component count to Base raises Overall F1 from 0.427 to about 0.436; using both reaches 0.447. The individual additions have different effects across subsets: derivative loss yields higher F1 on Main, whereas specialized kinetic-component count yields higher F1 on Challenge.

The similar Overall OOD R^2 values mask opposite changes across Main and Challenge. Relative to Base, Full FPMS lowers Main OOD R^2 from 0.624 to 0.601 and raises Challenge OOD R^2 from 0.404 to 0.437. Thus, combining the criteria improves reaction F1 on both subsets, while its trajectory advantage is specific to Challenge.

Table 7: **FPMS component ablations in KineSelect.** Base combines fitting-trajectory simulation loss and expanded ODE expression-tree complexity. The next two rows add derivative loss $\mathcal{L}^{\mathrm{der}}$ or specialized kinetic-component count C^{spec} to Base; Full FPMS includes both. Overall OOD R^2 uses all 198 systems, including Stress, whereas Overall reaction F1 uses the 161 reaction-eligible systems. Main and Challenge use 95 and 66 systems for both metrics. Bold and underlined values denote the best and second-best results in each column.

Selection	**OOD R^2↑**			**Reaction F1↑**		
	Overall	Main	Challenge	Overall	Main	Challenge
Base	<u>0.5141</u>	**0.6241**	0.4040	0.4272	0.5315	0.2772
$+\mathcal{L}^{\mathrm{der}}$	**0.5158**	0.6162	<u>0.4194</u>	<u>0.4363</u>	<u>0.5459</u>	0.2784
$+C^{\mathrm{spec}}$	0.5077	<u>0.6176</u>	0.4060	0.4361	0.5424	<u>0.2832</u>
Full FPMS	0.5081	0.6011	**0.4373**	**0.4475**	**0.5604**	**0.2848**

C.4 OOD Performance Distributions under Observation Noise

KineSelect achieves the highest OOD R^2 and Acc@.90 across all five tested noise conditions (Table 8). Figure 9 shows how the OOD score distributions change with noise. KineSelect reaches OOD $R^2 \geq 0.90$ on 74 systems without noise and 43 at 10% noise, compared with 65 and 40 for SISR. The difference in the number of systems with OOD $R^2 \geq 0.90$ therefore narrows at high noise.

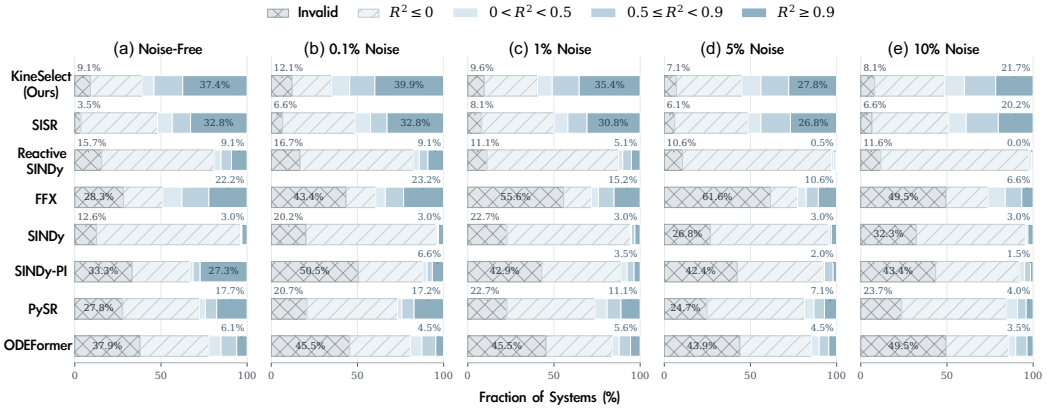


Figure 9: **OOD performance distributions under observation noise.** (a–e) Results for noise-free observations and relative noise levels of 0.1%, 1%, 5%, and 10%, respectively. Each bar represents all 198 systems. Cross-hatched segments denote invalid or undefined OOD scores, and diagonally hatched segments denote valid scores with raw OOD $R^2 \leq 0$. Valid predictions occupy four bins based on raw OOD R^2: $R^2 \leq 0$, $0 < R^2 < 0.5$, $0.5 \leq R^2 < 0.90$, and $R^2 \geq 0.90$. Labels give percentages of invalid/undefined scores and predictions with OOD $R^2 \geq 0.90$ to one decimal place.

The trajectory advantage does not extend to Overall reaction F1 at the higher noise levels (Table 9). At 5% noise, KineSelect has higher Challenge F1 than SISR (0.277 versus 0.232), but lower Main F1 (0.392 versus 0.432), yielding lower Overall F1. At 10% noise, SISR has higher F1 on both subsets, whereas KineSelect retains higher OOD R^2. The richer kinetic representation may admit alternative fits to noisy observations that preserve trajectory accuracy while matching the reference reactions less well.

The baseline distributions distinguish missing or invalid OOD scores from valid but inaccurate dynamics. FFX, SINDy-PI, and ODEFormer have large fractions of systems without valid OOD scores, reaching 61.6% for FFX at 5% noise. For Reactive SINDy, valid but non-positive OOD R^2 dominates at high noise: inaccurate trajectories remain a problem even when a score is available. SINDy reaches OOD $R^2 \geq 0.90$ on only 6 of 198 systems (3.0%) in each of the five conditions.

Table 8: **Trajectory fidelity under observation noise.** Each noise level reports OOD R^2 and Acc@.90 (Acc.) over all 198 systems; higher is better for both. Bold and underlined values denote the best and second-best results for each metric and noise level, including ties.

Method	Noise-free		0.1%		1%		5%		10%	
	R^2	Acc.	R^2	Acc.	R^2	Acc.	R^2	Acc.	R^2	Acc.
KineSelect	**0.508**	**0.374**	**0.532**	**0.399**	**0.485**	**0.354**	**0.421**	**0.278**	**0.383**	**0.217**
SISR	0.426	0.328	0.418	0.328	0.407	0.308	0.410	0.268	0.363	0.202
Reactive SINDy	0.142	0.091	0.135	0.091	0.094	0.051	0.016	0.005	0.010	0.000
FFX	0.367	0.222	0.318	0.232	0.226	0.152	0.175	0.106	0.162	0.066
SINDy	0.033	0.030	0.034	0.030	0.047	0.030	0.038	0.030	0.037	0.030
SINDy-PI	0.310	0.273	0.091	0.066	0.073	0.035	0.058	0.020	0.048	0.015
PySR	0.233	0.177	0.227	0.172	0.186	0.111	0.126	0.071	0.094	0.040
ODEFormer	0.136	0.061	0.121	0.045	0.109	0.056	0.096	0.045	0.094	0.035

Table 9: **Reaction F1 under observation noise.** Overall, Main, and Challenge report mean reaction F1 over 161, 95, and 66 systems, respectively. Bold and underlined values denote the best and second-best results within each noise level and subset.

Noise	Method	Overall↑	Main↑	Challenge↑
Noise-free	KineSelect	**0.447**	**0.560**	**0.285**
	SISR	0.391	0.507	0.225
	Reactive SINDy	0.292	0.363	0.191
0.1%	KineSelect	**0.453**	**0.566**	**0.291**
	SISR	0.371	0.467	0.233
	Reactive SINDy	0.278	0.335	0.195
1%	KineSelect	**0.426**	**0.511**	**0.304**
	SISR	0.378	0.464	0.253
	Reactive SINDy	0.209	0.255	0.142
5%	KineSelect	0.344	0.392	**0.277**
	SISR	**0.350**	**0.432**	0.232
	Reactive SINDy	0.143	0.169	0.106
10%	KineSelect	0.306	0.359	0.229
	SISR	**0.317**	**0.377**	**0.231**
	Reactive SINDy	0.137	0.165	0.096

C.5 Representative Recovery and Failure Cases

We use six representative systems to illustrate how trajectory fidelity and mechanism recovery align in some cases and separate in others. Figure 10 shows OOD trajectories after noise-free fitting, and Table 10 compares the reference and recovered equations alongside the full-system metrics.

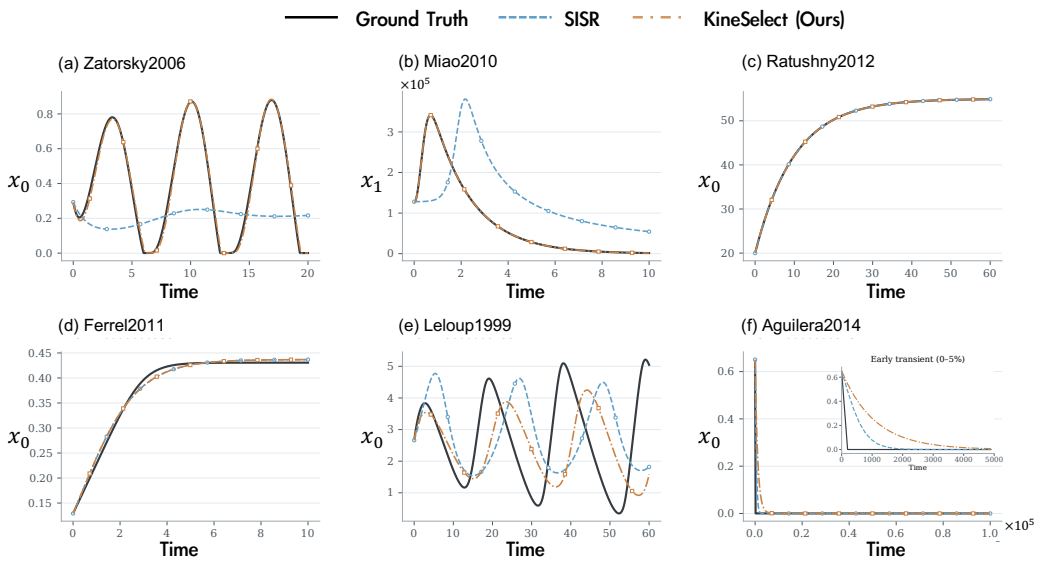


Figure 10: **Representative OOD trajectories after noise-free fitting.** (a-f) OOD trajectories of six representative systems, comparing Ground Truth, SISR, and KineSelect. Table 10 reports the corresponding equations and complete system-level OOD and reaction metrics. Each plot displays one state; the inset in (f) shows the first 5% of the time interval.

KineSelect recovers more reference signatures and produces more accurate OOD trajectories than SISR in Zatorsky2006 and Miao2010 (Figure 10(a,b)). In Zatorsky2006 (Geva-Zatorsky et al., 2006), KineSelect matches the Michaelis–Menten consumption signature in $\dot{x}_0$ together with four mass-action signatures. It closely reproduces the reference oscillations, reaching OOD $R^2 = 0.995$ and reaction F1 of 0.714. SISR matches none of the reference signatures and produces a nearly constant trajectory, with OOD $R^2 = -0.236$ and F1 of 0. In Miao2010 (Miao et al., 2010), both methods recover the x_0x_2 interaction signatures, but KineSelect also matches the linear consumption and production signatures that SISR misses. KineSelect closely follows the reference transient (OOD $R^2 = 1.000$, F1 $= 0.909$), whereas SISR extrapolates poorly (OOD $R^2 = -1.507$, F1 $= 0.333$). Even with this near-perfect OOD trajectory, KineSelect misses one reference linear-production signature.

For Ratushny2012 (Ratushny et al., 2012), both methods recover the two reference signatures for constant production and linear consumption, attaining OOD $R^2 = 1.000$ and reaction F1 of 1.000 (Figure 10(c)). Ferrel2011 (Ferrell et al., 2011) gives the contrasting case: both methods again reach OOD $R^2 \approx 0.993$, but their reaction F1 is only 0.400 (Figure 10(d)). Both methods match the constant production signature while missing the reference Hill-regulated consumption. Near-perfect OOD trajectories can therefore correspond to either complete or partial mechanism recovery.

Leloup1999 (Leloup et al., 1999) shows the reverse mismatch: KineSelect matches more reference signatures, yet both methods extrapolate poorly (Figure 10(e)). KineSelect matches inhibitory-Hill and Michaelis–Menten signatures and reaches reaction F1 of 0.588, compared with 0.375 for SISR. Both models deviate from the reference oscillations, with OOD $R^2 = -0.407$ for KineSelect and -1.502 for SISR. The matched Hill term uses $n = 2$ instead of the reference $n = 4$. Numerical Hill exponents are not part of the reaction-signature matching criterion (Appendix A.5).

SISR performs better on both trajectory fidelity and mechanism recovery in Aguilera2014 (Aguilera & Rodríguez-González, 2014) (Figure 10(f)). SISR matches three reference mass-action signatures,

whereas the selected KineSelect model retains only one reference match and includes unmatched Michaelis–Menten and Hill terms. SISR attains OOD $R^2 = 0.495$ and reaction F1 of 0.545, compared with -0.152 and 0.182 for KineSelect, respectively.

KineSelect recovers Michaelis–Menten and Hill signatures beyond SISR's mass-action representation in these examples. Reliable recovery of nonlinear mechanisms together with accurate OOD dynamics remains an open challenge. Across these cases, trajectory fidelity and mechanism recovery can improve together or diverge in either direction, motivating their joint evaluation.

Table 10: **Reference and recovered equations for the six representative cases.** The three columns show Ground Truth, SISR, and KineSelect. Ground-truth equations are black. Predicted reaction-signature occurrences are shown in black (matched) or muted red (unmatched). Parameter indices are local to each model and retain shared parameters across its states. OOD R^2 denotes the unclipped system-level score. Negative values are retained for the individual case studies. Precision, recall, and F1 are computed from reaction-signature matches. TP/FP/FN are occurrence counts. Case labels match Figure 10.

(a) BIOMD0000000157 (Zatorsky2006)

Ground Truth	SISR	KineSelect
$\begin{cases} \dot{x}_0 = k_1 - \dfrac{k_2\, x_0\, x_1}{K_{\mathrm{M},2} + x_0} \\ \dot{x}_1 = k_3\, x_2 - k_4\, x_1 \\ \dot{x}_2 = k_5\, x_0 - k_6\, x_2 \end{cases}$	$\begin{cases} \dot{x}_0 = -x_0\, x_1\ (k_1 + k_3) \\ \qquad + k_2\, {x_1}^2 + k_4\, x_0 \\ \dot{x}_1 = x_0\, x_1\ (k_1 - k_3) \\ \qquad - 2\, k_2\, {x_1}^2 \\ \dot{x}_2 = x_0\, x_1\ (k_1 + k_3) \\ \qquad + k_2\, {x_1}^2 - 2\, k_5\, {x_2}^2 \end{cases}$	$\begin{cases} \dot{x}_0 = k_2\, \dfrac{K_{\mathrm{H},2}}{K_{\mathrm{H},2} + x_1} + k_4\, x_2 \\ \qquad - \dfrac{k_6\, x_0\, x_1}{K_{\mathrm{M},6} + x_0} \\ \dot{x}_1 = k_3\, x_2 - k_5\, x_1 \\ \dot{x}_2 = k_1\, x_0 - k_3\, x_2 - k_4\, x_2 \end{cases}$

Method	OOD R^2	Precision	Recall	F1	TP/FP/FN
SISR	-0.236	0.000	0.000	0.000	0/8/6
KineSelect	0.995	0.625	0.833	0.714	5/3/1

(b) BIOMD0000000546 (Miao2010)

Ground Truth	SISR	KineSelect
$\begin{cases} \dot{x}_0 = k_1\, x_0 - k_2\, x_0\, x_2 \\ \dot{x}_1 = -k_3\, x_1 + k_4\, x_0\, x_2 \\ \dot{x}_2 = k_5\, x_1 - k_6\, x_2 \end{cases}$	$\begin{cases} \dot{x}_0 = -k_3\, x_0\, x_2 \\ \dot{x}_1 = k_3\, x_0\, x_2 - k_4\, x_1\, x_2 \\ \dot{x}_2 = 2\, k_1\, x_0\, x_2 + k_2\, x_1\, x_2 \\ \qquad - k_5\, {x_2}^2 \end{cases}$	$\begin{cases} \dot{x}_0 = -k_4\, x_0\, x_2 \\ \dot{x}_1 = -k_3\, x_1 + k_4\, x_0\, x_2 \\ \dot{x}_2 = k_1\, x_1 - k_2\, x_2 \end{cases}$

Method	OOD R^2	Precision	Recall	F1	TP/FP/FN
SISR	-1.507	0.333	0.333	0.333	2/4/4
KineSelect	1.000	1.000	0.833	0.909	5/0/1

(c) BIOMD0000000417 (Ratushny2012)

Ground Truth	SISR	KineSelect
$\dot{x}_0 = k_1 - k_2\, x_0$	$\dot{x}_0 = k_1 - k_2\, x_0$	$\dot{x}_0 = k_1 - k_2\, x_0$

Method	OOD R^2	Precision	Recall	F1	TP/FP/FN
SISR	1.000	1.000	1.000	1.000	2/0/0
KineSelect	1.000	1.000	1.000	1.000	2/0/0

Table 10: Reference and recovered equations for the six representative cases (continued).

(d) BIOMD0000000936 (Ferrel2011)

Ground Truth	SISR	KineSelect
$\dot{x}_0 = k_1 - \frac{k_2\, {x_0}^9}{{K_{\mathrm{H},2}}^8 + {x_0}^8}$	$\dot{x}_0 = -k_1\, {x_0}^2 + k_2\, x_0 + k_3$	$\dot{x}_0 = k_1 + k_2\, x_0 - 2\, k_3\, {x_0}^2$

Method	OOD R^2	Precision	Recall	F1	TP/FP/FN
SISR	0.993	0.333	0.500	0.400	1/2/1
KineSelect	0.993	0.333	0.500	0.400	1/2/1

(e) BIOMD0000000299 (Leloup1999)

Ground Truth	SISR	KineSelect
$\begin{cases} \dot{x}_0 = k_1 \dfrac{{K_{\mathrm{H},1}}^4}{{K_{\mathrm{H},1}}^4 + {x_2}^4} - \dfrac{k_2\, x_0}{K_{\mathrm{M},2} + x_0} \\ \dot{x}_1 = k_3\, x_0 + k_4\, x_2 - k_5\, x_1 - \dfrac{k_6\, x_1}{K_{\mathrm{M},6} + x_1} \\ \dot{x}_2 = k_7\, x_1 - k_8\, x_2 \end{cases}$	$\begin{cases} \dot{x}_0 = k_1\, x_0 - k_5\, x_0\, x_1 + k_6\, {x_0}^2 \\ \dot{x}_1 = 2\, k_3\, x_0\, x_1 - k_4\, x_1 + k_6\, {x_0}^2 \\ \dot{x}_2 = -k_2\, x_2 + 2\, k_4\, x_1 \end{cases}$	$\begin{cases} \dot{x}_0 = k_3 \dfrac{x_0}{K_{\mathrm{H},3} + x_0} + k_4 \dfrac{{K_{\mathrm{H},4}}^2}{{K_{\mathrm{H},4}}^2 + {x_2}^2} - \dfrac{k_8\, x_0}{K_{\mathrm{M},8} + x_0} \dfrac{x_2}{K_{\mathrm{H},8} + x_2} - k_9\, x_0\, x_1 \\ \dot{x}_1 = k_2\, x_0 - \dfrac{k_6\, x_1}{K_{\mathrm{M},6} + x_1} - k_9\, x_0\, x_1 \\ \dot{x}_2 = k_1\, x_1 - k_5\, x_2 \end{cases}$

Method	OOD R^2	Precision	Recall	F1	TP/FP/FN
SISR	-1.502	0.375	0.375	0.375	3/5/5
KineSelect	-0.407	0.556	0.625	0.588	5/4/3

(f) BIOMD0000000573 (Aguilera2014)

Ground Truth	SISR	KineSelect
$\begin{cases} \dot{x}_0 = k_1 - k_2\, x_0 + k_3 \dfrac{x_0}{K_{\mathrm{H},3} + x_0} - k_4\, x_0\, x_1 \\ \dot{x}_1 = k_5 - k_6\, x_1 - k_7\, x_0\, x_1 \end{cases}$	$\begin{cases} \dot{x}_0 = k_1\, {x_0}^2 - k_3\, x_0 \\ \dot{x}_1 = -k_2\, x_1 + k_4 \end{cases}$	$\begin{cases} \dot{x}_0 = -\dfrac{2\, k_4\, {x_0}^2}{K_{\mathrm{M},4} + x_0} \\ \dot{x}_1 = k_1 + k_2 \dfrac{x_1}{K_{\mathrm{H},2} + x_1} - \dfrac{k_3\, x_1}{K_{\mathrm{M},3} + x_1} \end{cases}$

Method	OOD R^2	Precision	Recall	F1	TP/FP/FN
SISR	0.495	0.750	0.429	0.545	3/1/4
KineSelect	-0.152	0.250	0.143	0.182	1/3/6

D Reproducibility and Implementation Details

We apply each method to the trajectory-to-ODE task through a shared data interface and method-specific fitting routines. This appendix specifies the inputs, search spaces, fitting budgets, and final-model selection used in the experiments.

D.1 Common Experimental Pipeline

Trajectory inputs and derivative targets. For each system, we fit the default trajectory using its recorded observation times and state-variable order. We follow the preprocessing routines of the adopted implementations; Tables 11 and 12 list the method-specific settings. Derivative-based methods estimate state-derivative targets from the fitting observations, whereas ODEFormer receives the trajectory directly. In the noise experiments, these methods re-estimate derivatives from the corrupted fitting trajectory.

Coupled-system adaptation. For scalar regressors, we fit each state equation using the full state vector and assemble the selected state equations into a coupled ODE. LLM-SR also receives time as an input. Reaction-based methods fit a mechanism whose stoichiometry couples the state equations.

Candidate selection and scoring. Internal parameter selection and candidate ranking use the fitting trajectory, with method-specific criteria described below. SISR and KineSelect compute their internal simulation losses against the fitting observations. For final scoring, each method's prediction routine integrates the recovered coupled ODE in the benchmark's original state coordinates and time units. SISR and KineSelect use LSODA (Petzold, 1983) with relative tolerance 10^{-6} and absolute tolerance 10^{-9}. LLM-SR uses the prediction settings in Table 12. We compute trajectory metrics from the saved predictions. ID reconstruction and OOD extrapolation use the initial conditions and time grids in Appendix A.3. We sum expression complexity across state equations (Section 4) and score reaction outputs from Reactive SINDy, SISR, and KineSelect using Appendix A.5. Failure handling and score aggregation follow Appendix A.4.

D.2 Baseline Implementations and Configurations

EPLEX, FE-AFP, FFX, and SINDy use baseline implementations adapted from the ODEFormer codebase (d'Ascoli et al., 2023). The remaining methods use the packages or Python implementations described below. We follow the adopted implementations' default fitting procedures, including their built-in parameter selection where enabled. Tables 11 and 12 list the search spaces and budgets for the eleven baseline configurations. The descriptions below explain their implementation and task adaptations.

Generic symbolic-search methods. We run EPLEX (La Cava et al., 2016) using Ellyn. The default configuration includes arithmetic, exponential, logarithmic, square-root, square, and cube operators. The poly configuration uses arithmetic, square, and cube operators, including division. The two configurations use their respective population sizes, generation limits, and time limits in Table 11.

FE-AFP (Schmidt & Lipson, 2010) uses Ellyn with fitness estimation enabled. Each estimator approximates candidate-program fitness using a small subset of the fitting data. We score estimators by how well they preserve the full-data fitness ranking of a trainer set of candidate programs. Table 11 reports the estimator and trainer population sizes, observations per estimator, and trainer-selection interval.

We retain the expression-construction settings of FFX (McConaghy, 2011) and its built-in parameter selection. The fitting routine ranks returned ODE candidates by trajectory R^2.

PySR (Cranmer, 2023) uses the installed Python package with its Julia search backend. We fit derivative targets with squared error, the configured arithmetic operators, and 50 search iterations. The fitting routine combines component expressions into ODE candidates and ranks them by simulated trajectory R^2.

Sparse system-identification methods. We run SINDy (Brunton et al., 2016) with a polynomial library and sequentially thresholded least squares (STLSQ). The search covers polynomial degrees one through five, together with the threshold, regularization, iteration, and differentiation settings in Table 12.

We use a Python implementation of SINDy-PI (Kaheman et al., 2020). It searches implicit relations with left-hand sides of the form $\dot{x}_j g(\mathbf{x})$, then converts each selected relation to an explicit differential equation. The state library has degree two, and the denominator library has degree one. We search the listed sparsity thresholds.

Reactive SINDy (Hoffmann et al., 2019) uses a local Python implementation of nonnegative sparse regression over vector-valued reaction basis functions. We enumerate mass-action reactions up to order two, including fission and double-conversion candidates, and remove duplicates. The fitter uses L-BFGS-B with nonnegative rate constraints and the regularization grid in Table 12. Absolute and relative rate thresholds remove small fitted terms. Both SINDy-PI and Reactive SINDy normalize library columns and fit derivatives at interior time points.

Neural and language-model methods. We load the authors' pretrained ODEFormer (d'Ascoli et al., 2023) checkpoint without fine-tuning. We use its symbolic-transformer regressor with candidate beam size 50, sampling temperature 0.1, and input rescaling. The regressor ranks candidate systems by trajectory R^2. We return the highest-ranked system.

LLM-SR (Shojaee et al., 2024) runs a separate search for each state equation, using Qwen3-Coder-Flash (Yang et al., 2025) as the language model. Each search receives time and the full state vector and fits program parameters against the corresponding derivative target. We allocate 100 generated samples per state variable, giving a nominal budget of $100d$ samples for a d-dimensional system. Each program has at most eight fitted parameters and a 30-second evaluation limit. We assemble the selected programs into the final ODE.

SISR. We evaluate SISR (Palma Banos et al., 2026) over mechanisms containing 2–12 reactions, with a separate subpopulation for each reaction count. We fit their rate constants by nonnegative least squares. The clean configuration uses 2,000 individuals and 20 generations per subpopulation. The original study selects a mechanism by inspecting the complexity–error Pareto front (Palma Banos et al., 2026). We implement this loss–complexity principle using deterministic α-threshold selection with $\alpha = 1.05$ and $\epsilon = 10^{-12}$ (Appendix B.3). The fixed-archive comparison in Section D.4 evaluates the effect of the final selector.

Table 11: **Generic symbolic-search configurations.** Entries give fixed settings or indicate built-in parameter selection. All time limits are in seconds. The two EPLEX configurations differ in both operators and budget.

Method	Hyperparameter	Value(s)
EPLEX (default)	Population / generations	200 / 100
	Time limit per state equation (s)	180
	Operators	$+, -, \times, \div, \exp, \log, \sqrt{\cdot}, (\cdot)^2, (\cdot)^3$
EPLEX (poly)	Population / generations	500 / 500
	Time limit per state equation (s)	7,200
	Operators	$+, -, \times, \div, (\cdot)^2, (\cdot)^3$
EPLEX (both)	Maximum / initial expression length	64 / 20
	Difference order / smoothing	2 / none
	Built-in parameter selection	Disabled
	Candidate ranking	Trajectory R^2
FE-AFP	Population / generations	1,000 / 250
	Search time limit per system	3,600
	Operators	$+, -, \times, \div, (\cdot)^2, (\cdot)^3$
	Maximum / initial expression length	64 / 20
	Fitness estimation / rank-based scoring	Enabled / enabled
	Fitness-estimator population	100 estimators
	Observations per estimator	10
	Trainer population	10 programs
	Trainer-selection interval	10 generations
	Difference order / smoothing	2 / none
	Built-in parameter selection	Disabled
	Candidate ranking	Trajectory R^2
FFX	Difference order / smoothing defaults	2 / none
	Built-in parameter selection	Enabled
	Candidate ranking	Trajectory R^2
PySR	Search iterations	50
	Binary operators	$+, -, \times, \div, \mathrm{pow}$
	Unary operator	Negation
	Loss / search processes	Squared error / 1
	Difference order / smoothing defaults	2 / none
	Built-in parameter selection	Enabled
	Candidate ranking	Trajectory R^2

Table 12: **Sparse, neural, language-model, and SISR configurations.** *Searched* denotes values explored by the built-in fitting procedure. SISR population and generation limits apply to each reaction-count subpopulation. The SISR elitism entry describes the clean configuration.

Method	Hyperparameter	Value(s)
SINDy	Polynomial degree (searched)	$1, 2, 3, 4, 5$
	STLSQ threshold (searched)	$0.05, 0.1, 0.15$
	STLSQ regularization (searched)	$0.025, 0.05, 0.075$
	STLSQ iterations (searched)	$20, 100$
	Difference order (searched)	$2, 3, 4$
	Smoothing window (searched)	None, 15
	State normalization	None
SINDy-PI	State / denominator polynomial degree	2 / 1
	Sparsity threshold (searched)	$0.01, 0.025, 0.05, 0.1, 0.2$
	STLSQ iterations	20
	Difference order / smoothing	2 / none
	Minimum denominator magnitude	10^{-8}
Reactive SINDy	Reaction library / maximum order	Enumerated / 2
	Fission / double conversion	Enabled / enabled
	Penalty scale (searched)	$0.001, 0.01, 0.1, 0.5, 1$
	ℓ_1 fraction	1
	Optimizer / maximum iterations	L-BFGS-B / 5,000
	Optimization tolerance	10^{-10}
	Absolute / relative rate threshold	10^{-8} / 10^{-4}
SINDy-PI and Reactive SINDy	Sparsity-selection tolerance	0.05
ODEFormer	Candidate beam size / temperature	50 / 0.1
	Maximum input points	10,000
	Rescaling / candidate ranking	Enabled / trajectory R^2
LLM-SR	Language model	Qwen3-Coder-Flash
	Temperature / samples per prompt	0.7 / 4
	Generated samples per state variable	100
	Samplers / evaluators / program islands	1 / 1 / 10
	Functions per prompt / reset period (s)	2 / 14,400
	Maximum fitted parameters / time input	8 / enabled
	Program-evaluation limit (s)	30
	Prediction integrator / timeout (s)	LSODA / 15
	Prediction rtol / atol	10^{-6} / 10^{-8}
SISR	Reaction-count range	2–12
	Population / generations per subpopulation	2,000 / 20
	Elitism / mutation probability	0.1 / 0.1
	Base search seed	42
	Reaction order / stoichiometric bound	2 / 2
	Final selection	α-threshold
	α / numerical ϵ	1.05 / 10^{-12}
	Candidate integration rtol / atol	10^{-6} / 10^{-9}

D.3 KINESELECT CONFIGURATION

Genetic search. We use the reaction representation and variation operators in Appendix B.1. Each reaction-count subpopulation has a configured population of 2,000 and runs for at most 20 generations. Tournament selection, elitism, and mutation/crossover use the values in Table 13. Early stopping monitors the derivative-fitting loss and ends a subpopulation search after four generations without a relative improvement of at least 10^{-4}.

Kinetic candidates and fitting. For each species, we construct K candidates from the 0.05 and 0.50 quantiles of its positive fitting observations. Thus, systems share the candidate-generation rule while their numerical K values depend on the observed species trajectories. Hill candidates use exponents $1, 2, 4$. We set the regularization coefficient to $\lambda = 10^{-6}$ in Eq. (7).

Archive and final selection. We simulate the 32 candidates with the lowest derivative loss from each subpopulation to form the archive described in Appendix B.1. FPMS selects the final model using fidelity criteria $\mathcal{L}_i^{\mathrm{der}}$ and $\mathcal{L}_i^{\mathrm{sim}}$, and parsimony criteria C_i^{tree} and C_i^{spec}, as defined in Section 5.3. Appendix B.3 gives normalization and tie-breaking, and Appendix B.1 specifies the derivative-based fallback.

Table 13: **KineSelect runtime configuration.** Population size and simulation-candidate limits apply to each reaction-count subpopulation. The K-candidate quantiles use each species' fitting observations.

Hyperparameter	Value
Genetic search	
Reaction-count range	2–12
Population / maximum generations	2,000 / 20
Elitism / tournament size	0.1 / 20
Mutation / crossover rate	0.3 / 0.7
Base search seed	42
Early-stopping patience / relative improvement	4 generations / 10^{-4}
Maximum reactant / product order	2 / 2
Maximum stoichiometry per species	2
Kinetic selection	
Reactant-dependent kinetics	Mass action, Michaelis–Menten
Regulatory terms	None, linear, Hill activation/inhibition
Species-specific K quantiles	0.05, 0.50
Hill exponents	$1, 2, 4$
Archive and final selection	
Simulation candidates per subpopulation	32
Final selector	FPMS

D.4 Settings for Selection Studies and Noise Experiments

KFI and FPMS ablations. We evaluate four configurations of KineSelect by enabling or disabling KFI and FPMS. Without KFI, all reaction laws are fixed to mass action, and nonnegative least squares fits their rate constants. Without FPMS, final models use α-threshold selection with $\alpha = 1.05$. Within each KFI setting, the two selectors reuse the same saved candidate archive and fitted parameters. For the mass-action-only configuration, we reintegrate the selected models with their saved rate constants using LSODA (Petzold, 1983), relative tolerance 10^{-6}, and absolute tolerance 10^{-9}.

FPMS component ablations. We apply the four settings in Table 14 to KineSelect. The base uses simulation loss $\mathcal{L}_i^{\mathrm{sim}}$ and expanded ODE expression-tree complexity C_i^{tree}. We add derivative loss $\mathcal{L}_i^{\mathrm{der}}$, specialized kinetic-component count C_i^{spec}, or both. Each setting reuses the same candidates and fitted parameters, with normalization and tie-breaking from Appendix B.3.

Table 14: **FPMS component-ablation settings.** All settings use $U_i = (1 - Q_i)(1 - P_i)$. Tildes denote the transformations and normalization in Appendix B.3.

Setting	Fidelity score Q_i	Parsimony score P_i
Base	$\widetilde{\mathcal{L}}_i^{\mathrm{sim}}$	$\widetilde{C}_i^{\mathrm{tree}}$
Base + derivative loss	$\frac{\widetilde{\mathcal{L}}_i^{\mathrm{der}} + \widetilde{\mathcal{L}}_i^{\mathrm{sim}}}{2}$	$\widetilde{C}_i^{\mathrm{tree}}$
Base + specialized kinetic-component count	$\widetilde{\mathcal{L}}_i^{\mathrm{sim}}$	$\frac{\widetilde{C}_i^{\mathrm{tree}} + \widetilde{C}_i^{\mathrm{spec}}}{2}$
Full FPMS	$\frac{\widetilde{\mathcal{L}}_i^{\mathrm{der}} + \widetilde{\mathcal{L}}_i^{\mathrm{sim}}}{2}$	$\frac{\widetilde{C}_i^{\mathrm{tree}} + \widetilde{C}_i^{\mathrm{spec}}}{2}$

Observation-noise experiments. We rerun FFX, SINDy, PySR, SINDy-PI, ODEFormer, Reactive SINDy, SISR, and KineSelect on fitting trajectories corrupted at the noise levels in Appendix A.3. All methods receive the same corrupted observations for each system and noise level, with noise seed 2026 and one realization per level. The recovered systems are evaluated against the clean ID reconstruction and OOD extrapolation trajectories. For SISR, the noisy runs use elitism 0.4 in the two-reaction subpopulation and 0.1 in the remaining subpopulations; the clean configuration uses 0.1 throughout.

D.5 Reproduction Resources

We will provide an anonymized archive containing the benchmark and code for running and evaluating the experiments. Software dependencies and environment specifications will accompany the code.