EVOTRACE: LEARNING TO EVOLVE SCIENTIFIC HYPOTHESES FROM TRAJECTORIES OF AGENTIC SYMBOLIC REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Discovering concise and interpretable mathematical laws from observational data has conventionally been addressed through symbolic regression (SR). Recent advances in large language models (LLMs) have enabled this task to be formulated as an agentic search process that iteratively proposes, evaluates, and refines candidate formulas, often achieving superior performance. However, existing LLM-based approaches often fail to learn effectively from accumulated search experience: unsuccessful hypotheses are seldom reused, scalar feedback provides little guidance for structural revision, and exploration repeatedly revisits unproductive regions of the search space. This paper proposes EvoTrace, an agentic SR framework that distills actionable knowledge from search trajectories to drive mechanism-guided hypothesis evolution. By contrasting successful and unsuccessful trajectories, EvoTrace identifies structural modifications associated with performance improvements and uses them to guide subsequent hypothesis generation. It further integrates mechanism-anchored constraints with evaluation feedback to determine whether candidate hypotheses should be retained, revised, or rejected. Experiments across five key space-physics applications, including sunspot activity, plasma pressure, and lunar-tide signals, demonstrate that EvoTrace achieves stronger overall performance than existing LLM-based and non-LLM symbolic regression methods across evaluations of predictive accuracy, formula simplicity, extrapolation capability, and physical interpretability.

1 INTRODUCTION

Many scientific advances begin with the same fundamental challenge: discovering concise mathematical laws that characterize observed phenomena. Symbolic regression (SR) tackles this challenge by searching for explicit expressions that balance predictive accuracy and interpretability (Koza, 1994; Schmidt & Lipson, 2009), transforming numerical observations into candidate hypotheses for scientific examination. Traditional SR mainly follows two directions: genetic programming explores expression spaces through stochastic variation, selection, and constant optimization (Virgolin et al., 2021; Cranmer, 2023), while deep learning methods generate formulas directly from data (Petersen et al., 2021; Biggio et al., 2021; Kamienny et al., 2022). The former is often limited by search efficiency, whereas the latter depends strongly on training-distribution coverage and predefined inductive biases. More recently, LLM-based agents have enabled formula generation, constant optimization, program execution, and numerical evaluation to be coordinated within iterative SR pipelines (Du et al., 2024; Grayeli et al., 2024; Shojaee et al., 2025; Xia et al., 2026).

Despite their flexibility, many LLM-based SR systems retain little structured information beyond candidate expressions and their aggregate fitness scores. Although such scores indicate whether one candidate outperforms another, they provide limited insight into which structural modification produced the improvement, under which evaluation conditions the improvement persists, or whether it reflects scientifically meaningful structure rather than increased model flexibility. For instance, adding a harmonic term may reveal genuine periodicity in the underlying phenomenon, but it may also simply absorb noise. Distinguishing these possibilities requires preserving the relationship between a parent formula and its revision, including the structural edit, evaluation context, and resulting evidence. Yet unsuccessful candidates are often discarded without extracting reusable

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.

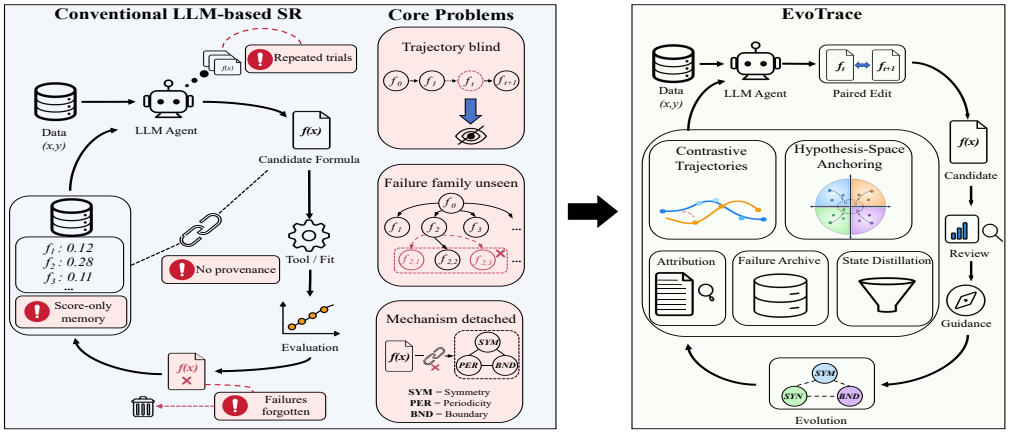

Figure 1: Overview of EvoTrace. Unlike score-driven LLM-based SR, EvoTrace reuses contrastive trajectory evidence and mechanism-anchored constraints to guide iterative hypothesis evolution.

knowledge at the level of formula families, causing similar structural failures to recur. Moreover, domain knowledge, such as physical semantics, periodicity, boundary behavior, dimensional consistency, and complexity, is frequently used only for post hoc validation rather than to guide hypothesis generation and revision. Consequently, the search process lacks the structured memory, causal attribution, and mechanism-level guidance needed to learn systematically from previous exploration.

To this end, we introduce EvoTrace, an agentic symbolic regression framework that learns to evolve scientific hypotheses from accumulated search trajectories, as illustrated in Figure 1. Rather than representing candidates as isolated expressions paired with scalar scores, EvoTrace records structured hypothesis transitions that link parent formulas, proposed structural edits, evaluation contexts, and resulting evidence. Trajectory-level evidence determines which structural changes should be propagated, while mechanism-anchored evidence determines how the admissible hypothesis space should evolve. The first core module, **CoTraS** (**Co**ntrastive **Tra**jectory **S**earch Steering), represents formula evolution as structured transitions from parent expressions to revised candidates. Each transition records the applied structural edit and its outcome under a specified evaluation context. By contrasting successful and unsuccessful transitions evaluated under comparable conditions, CoTraS identifies edits associated with reliable improvements and converts accumulated search experience into actionable guidance for subsequent exploration.

The second core module, **MA-HASE** (**M**echanism-**A**nchored **H**ypothesis-space **A**daptive **S**earch and **E**volution), uses trajectory-derived evidence to adapt the admissible hypothesis space throughout the search. It combines performance with scientific constraints, including physical semantics, periodic structure, boundary behavior, and dimensional consistency, to retain or expand promising formula families while revising or suppressing implausible and repeatedly unsuccessful ones.

EvoTrace further incorporates three auxiliary evidence components that make search experience reusable across iterations. *Paired attribution* links performance changes to specific structural edits; *formula-family failure archival* preserves recurring and reproducible failure patterns; and *scientific search-state distillation* compresses cross-iteration evidence into a reusable search state. Together, these components help retain validated edits, avoid previously observed failures, and concentrate limited evaluation budget on scientifically meaningful regions of the hypothesis space.

We evaluate EvoTrace on five real-world scientific discovery tasks spanning sunspot activity, plasma-sheet pressure, solar differential rotation, solar emission-line contribution functions, and lunar tidal signals. Across these tasks and multiple LLM backbones, EvoTrace achieves stronger overall performance than existing approaches, offering a favorable balance among predictive accuracy, expression simplicity, extrapolation robustness, and physical consistency. Its performance across problems with distinct data characteristics and underlying physical structures demonstrates the broad applicability of trajectory-guided hypothesis evolution. Ablation studies further attribute these gains to the complementary roles of contrastive trajectory steering, mechanism-anchored hypothesis-space evolution, and auxiliary evidence preservation. Our main contributions are summarized as follows:

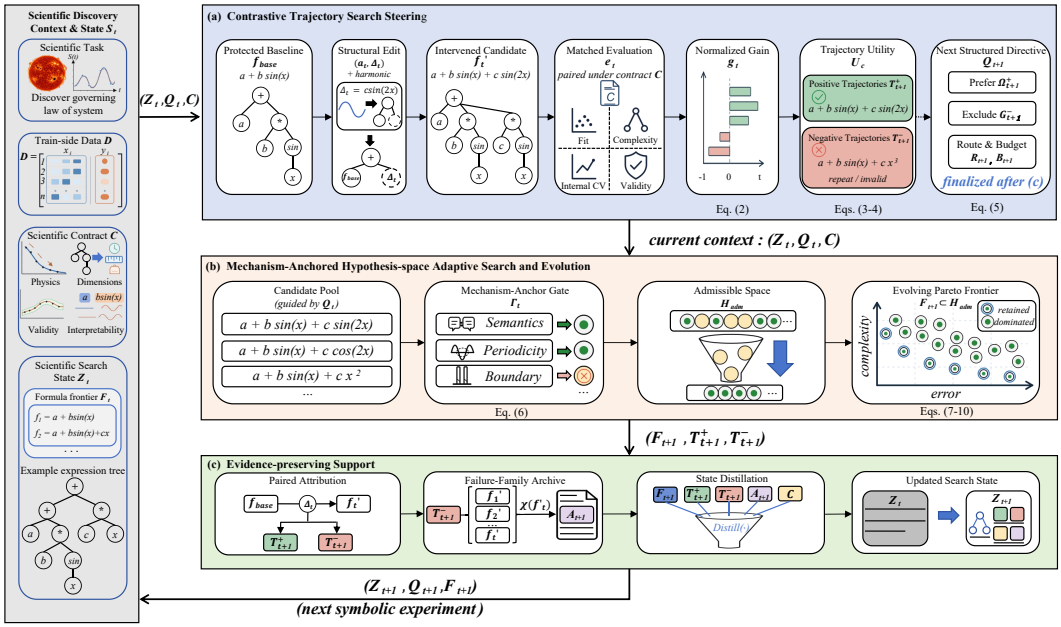

Figure 2: **Overview of EvoTrace.** (a) CoTraS extracts reusable guidance from contrastive search trajectories; (b) MA-HASE evolves the admissible hypothesis space under mechanism anchors; (c) Evidence-preserving Support carries verified search evidence across iterations.

- We propose **EvoTrace**, an agentic symbolic regression framework that distills successful and failed search trajectories into reusable guidance for hypothesis evolution.
- We introduce **CoTraS** and **MA-HASE** to couple contrastive trajectory evidence with mechanism-anchored hypothesis-space evolution.
- Experiments across **five real-world space-physics tasks** and **three LLM backbones** demonstrate strong cross-task and cross-model performance, with ablations confirming the complementary roles of the proposed components.

2 Related Work

Symbolic regression (SR) aims to discover compact mathematical expressions that capture relationships in observed data while preserving predictive accuracy and interpretability. Existing approaches can be broadly categorized into genetic-programming-based methods (Schmidt & Lipson, 2009; Cranmer, 2023; Virgolin et al., 2021), deep-learning-based methods (Petersen et al., 2021; Biggio et al., 2021; Kamienny et al., 2022; Ying et al., 2025), and large-language-model-based methods (Du et al., 2024; Shojaee et al., 2025; Grayeli et al., 2024; Xia et al., 2026). A detailed discussion of representative SR paradigms is provided in Appendix E.

Existing methods have progressively moved from stochastic expression search toward knowledge-guided and agentic hypothesis exploration. However, they do not explicitly organize successful and unsuccessful structural edits as contrastive trajectory evidence, nor use such evidence to adapt the admissible hypothesis space under scientific constraints. EvoTrace addresses this gap by combining contrastive trajectory steering with mechanism-anchored hypothesis evolution, allowing accumulated search experience to directly guide subsequent hypothesis generation, revision, and selection.

3 Method

We propose EvoTrace, an agentic symbolic regression framework that transforms successful and failed structural transitions into reusable search experience and combines this evidence with

mechanism-based constraints to guide iterative hypothesis evolution. Figure 2 summarizes the closed-loop interaction among CoTraS, MA-HASE, and Evidence-preserving Support: CoTraS extracts trajectory-level guidance, MA-HASE enforces mechanism-anchored admissibility, and the distilled state Z_{t+1} carries reusable evidence to the next symbolic experiment.

3.1 CoTraS: Contrastive Trajectory Search Steering

CoTraS provides trajectory-level search control by comparing structural transitions rather than ranking formulas solely by isolated scalar scores. The scientific contract $\mathcal{C}$ fixes the data-visibility boundary, admissible operators, evaluation protocol, search budget, variable semantics, and physical units supplied with the data. It defines the observable context of the search without prescribing a target equation, formula family, or mechanism. The operational contract and search-state definitions are given in Appendix A.1. At iteration t, let Z_t denote the scientific search state and f_t^{base} a protected baseline from the active frontier. We represent each transition as

$$\xi_t = \left(Z_t, f_t^{\mathrm{base}}, a_t, \Delta_t, f_t', \mathbf{e}_t(f_t^{\mathrm{base}}), \mathbf{e}_t(f_t')\right), \tag{1}$$

where a_t is the search action, Δ_t the structural edit, and f_t' the resulting candidate. For a candidate f, we use $\mathbf{e}_t(f) = [\mathcal{L}_{\mathrm{fit}}(f), \mathcal{L}_{\mathrm{cv}}(f), c(f), 1 - q_{\mathrm{num}}(f)]$, where $\mathcal{L}_{\mathrm{fit}}$ and $\mathcal{L}_{\mathrm{cv}}$ are train-side fitting and internal-validation losses, $c(f)$ is normalized AST complexity, and q_{num} measures numerical validity. All dimensions are oriented so that lower values are preferred. Scientific relations inferred from the data are handled by MA-HASE rather than inserted into this empirical evidence vector. The evidence construction, validity predicates, and context-matching rules are detailed in Appendix A.2. CoTraS measures the relative improvement of an edit against its protected baseline:

$$\mathbf{g}_t(\xi_t) = \mathrm{clip}_{[-\rho,\rho]}\left(\mathbf{D}_{\mathcal{C}} \frac{\mathbf{e}_t(f_t^{\mathrm{base}}) - \mathbf{e}_t(f_t')}{|\mathbf{e}_t(f_t^{\mathrm{base}})| + \epsilon}\right), \tag{2}$$

where all vector operations are applied component-wise. $\mathbf{D}_{\mathcal{C}}$ is a diagonal weighting matrix that masks unavailable evidence dimensions and assigns contract-fixed nonnegative weights to the active ones, normalized to sum to one. The constant $\rho > 0$ is the clipping bound for extreme relative changes, and $\epsilon > 0$ is a small stabilizer that prevents division by near-zero baseline values. The resulting gain is summarized by

$$U_{\mathcal{C}}(\xi_t) = \sum_{j=1}^{d} g_{t,j}(\xi_t) - \lambda_{\mathrm{rep}} R_t^{\mathrm{rep}}(\xi_t), \tag{3}$$

where $d = \dim(\mathbf{e}_t)$, $\lambda_{\mathrm{rep}} \geq 0$ controls the strength of the repetition penalty, and $R_t^{\mathrm{rep}}(\xi_t) \geq 0$ measures how strongly the transition revisits previously unproductive structural regions. Contract compliance remains a hard condition; R_t^{rep} is a graded penalty, whereas $\mathrm{Repeat}(\xi_t)$ denotes an explicit canonical repeat. Each transition receives a contrastive label:

$$y_t = \begin{cases} +1, & U_{\mathcal{C}}(\xi_t) \geq \tau_+ \wedge \mathrm{Valid}_{\mathcal{C}}(f_t') \\ & \quad \wedge \neg\, \mathrm{Viol}_{\mathcal{C}}(\xi_t) \wedge \neg\, \mathrm{Repeat}(\xi_t) \wedge \neg\, \mathrm{Fail}(\xi_t), \\ -1, & U_{\mathcal{C}}(\xi_t) \leq \tau_- \vee \mathrm{Repeat}(\xi_t) \\ & \quad \vee\, \mathrm{Fail}(\xi_t) \vee \mathrm{Viol}_{\mathcal{C}}(\xi_t), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Here, $\tau_+ > \tau_-$ are contract-fixed utility thresholds that define supported improvements and harmful transitions, leaving an intermediate non-commitment region between them. $\mathrm{Valid}_{\mathcal{C}}$ covers the static syntax, numerical, dimensional, and complexity checks required to evaluate a transition, while transition-level violations cover the data-visibility, operator, budget, and evaluation rules of $\mathcal{C}$. State-dependent novelty and mechanism-anchor checks are applied later by MA-HASE. At iteration t, the directive Q_t summarizes only evidence accumulated before generating the current transition ξ_t. Accordingly, let $\mathcal{T}_t^{\pm} = \{\xi_\tau : 0 \leq \tau < t,\ y_\tau = \pm 1\}$ denote the positive and negative transitions available at the beginning of iteration t. After the current transition ξ_t is evaluated and labeled by Eq. (4), the trajectory sets are updated as $\mathcal{T}_{t+1}^{\pm} = \mathcal{T}_t^{\pm} \cup \{\xi_t : y_t = \pm 1\}$.

CoTraS aggregates only completed transitions that share the same scientific contract, compatible data-visibility and evaluation protocols, and comparable baseline/edit contexts. Positive trajectories

retain evidence for beneficial structural edits, whereas negative trajectories retain harmful edits, numerical failures, contract violations, and unproductive revisits. After the current transition ξ_t has been evaluated and the updated search state Z_{t+1} has been formed, CoTraS compiles the trajectory evidence available for the next iteration into

$$\mathcal{Q}_{t+1} = \left(\Omega^+_{t+1}, \mathcal{G}^-_{t+1}, \mathcal{R}_{t+1}, \mathcal{B}_{t+1}\right) = \mathrm{Compile}\left(Z_{t+1}, \mathcal{T}^+_{t+1}, \mathcal{T}^-_{t+1}, \mathcal{C}\right), \tag{5}$$

where Ω^+_{t+1} contains structural operations supported for future exploration, $\mathcal{G}^-_{t+1}$ contains temporarily excluded formula families supported by matched failure evidence, $\mathcal{R}_{t+1}$ specifies the next analysis/evaluation route, and $\mathcal{B}_{t+1}$ records the remaining search budget and stopping state. Transitions without sufficient evidence remain uncommitted. Importantly, a positive CoTraS label alone does not promote an edit into Ω^+_{t+1}; promotion to positive search guidance additionally requires the corresponding candidate to pass the active MA-HASE admission rule, as formalized in Sec. B.4. The directive compiler and failure-family update rules are specified in Appendix A.3.

At the beginning of iteration $t+1$, the Parent Agent is conditioned on Z_{t+1}, $\mathcal{Q}_{t+1}$, and $\mathcal{C}$. It selects a protected baseline from the updated frontier and initiates a structural test. MA-HASE combines this trajectory-derived directive with contract-visible scientific evidence to instantiate evidence-qualified scientific constraints and executable candidate edits. Candidates with $\chi(f) \in \mathcal{G}^-_{t+1}$ are excluded from the admitted pool, where $\chi(\cdot)$ denotes the canonical formula-family fingerprint. Consequently, the directive used at iteration $t+1$ depends only on evidence from completed transitions through iteration t, preventing the current transition from influencing the guidance that generated itself.

3.2 MA-HASE: Mechanism-Anchored Hypothesis-space Adaptive Search and Evolution

CoTraS determines where the search should proceed; MA-HASE determines which scientific structures remain sufficiently supported. MA-HASE separates deterministic admissibility from evidence-induced scientific constraints. Deterministic checks follow directly from $\mathcal{C}$, whereas scientific constraints are inferred from variable semantics, physical units, train-visible observations, and verified trajectories, and thus remain revisable. Given Z_t, $\mathcal{Q}_t$, and $\mathcal{C}$, MA-HASE compiles

$$\mathcal{M}_t = \mathrm{Compile}_{\mathrm{M}}(\mathcal{C}, Z_t, \mathcal{Q}_t) = (\mathcal{D}_t, \Psi_t, \mathcal{G}^-_t, b_{\mathcal{C}}), \tag{6}$$

where $\mathcal{D}_t$ contains executable structural proposals, Ψ_t is the set of currently active evidence-qualified scientific constraints, $\mathcal{G}^-_t$ is the failure-family archive, and $b_{\mathcal{C}}$ is the AST-complexity ceiling. Each pair $(\psi, \omega) \in \Psi_t$ consists of a falsifiable relation ψ and a positive evidence-derived support weight ω. Each active relation retains its provenance in the search ledger and may describe variable dependence, interactions, periodicity, scaling, or boundary behavior. These evidence-qualified relations serve as mechanism anchors: they propose testable structure rather than a target equation. The construction of $\mathcal{M}_t$, including relation provenance and activation, is detailed in Appendix B.1. When $\Psi_t \neq \varnothing$, candidate support for the active scientific constraints is

$$\Gamma_t(f) = \frac{\sum_{(\psi,\omega)\in\Psi_t} \omega\ \mathrm{Sat}(f;\psi)}{\sum_{(\psi,\omega)\in\Psi_t} \omega} - \lambda_{\mathrm{inc}}\,\mathrm{Incomp}(f, \Psi_t), \tag{7}$$

where $\mathrm{Sat}(f;\psi) \in [0,1]$ measures f's satisfaction of relation ψ using symbolic checks and train-visible numerical probes. The term $\mathrm{Incomp}(f, \Psi_t) \geq 0$ measures incompatibility among active relations, and $\lambda_{\mathrm{inc}} \geq 0$ controls its strength. The support weights ω are computed from provenance and repeated evidence rather than reviewer-assigned. When $\Psi_t = \varnothing$, we define $\Gamma_t(f) = 1$, leaving the scientific-consistency gate inactive until a relation has sufficient evidence for activation.

We distinguish new-proposal admission from incumbent retention. Let $V_{\mathcal{C},t}(f) \in \{0,1\}$ denote the conjunction of the visibility, AST, numerical, dimensional, and complexity checks under the current contract, excluding novelty. The current retention rule is

$$\mathrm{Keep}_{\mathcal{C}}(f \mid \mathcal{M}_t) = \begin{cases} V_{\mathcal{C},t}(f), & \Psi_t = \varnothing, \\ V_{\mathcal{C},t}(f)\,\mathbf{1}[\Gamma_t(f) \geq \gamma_{\mathcal{C}}], & \Psi_t \neq \varnothing, \end{cases} \tag{8}$$

where $\gamma_{\mathcal{C}}$ is the contract-fixed support threshold. For a new proposal, deterministic validity additionally requires structural novelty, $\mathrm{DetValid}_{\mathcal{C},t}(f) = V_{\mathcal{C},t}(f)\,\mathrm{Novel}_t(f)$, and admission is defined as $\mathrm{Adm}_{\mathcal{C}}(f \mid \mathcal{M}_t) = \mathrm{Keep}_{\mathcal{C}}(f \mid \mathcal{M}_t)\,\mathrm{Novel}_t(f)$.

Thus, new proposals and incumbents satisfy the same current validity and scientific-consistency requirements, but novelty is checked only for new proposals. An incumbent, including the protected baseline, is not rejected merely because it has previously been evaluated. Fitting accuracy cannot compensate for a failed validity check or insufficient support for active mechanism anchors. See Appendix B.2 for details. Guided by Ω_t^+ and $\mathcal{D}_t$, MA-HASE generates $\widetilde{\mathcal{P}}_{t+1} = \mathrm{Expand}(\mathcal{F}_t, \Omega_t^+; \mathcal{D}_t)$ and filters new proposals by admission and failure-family exclusion:

$$\mathcal{P}_{t+1} = \left\{ f \in \widetilde{\mathcal{P}}_{t+1} : \mathrm{Adm}_{\mathcal{C}}(f \mid \mathcal{M}_t) = 1,\ \chi(f) \notin \mathcal{G}_t^- \right\}. \tag{9}$$

Before selection, incumbents are re-screened as $\bar{\mathcal{F}}_t = \{f \in \mathcal{F}_t : \mathrm{Keep}_{\mathcal{C}}(f \mid \mathcal{M}_t) = 1\}$, without reapplying novelty. Admitted proposals and retained incumbents are compared under the same evaluation context. Let $\preceq_{\mathcal{C}}$ denote component-wise non-inferiority under contract-defined tolerances, and let $J_{\mathcal{C}}$ index the active empirical dimensions. For $f_i, f_j \in \bar{\mathcal{F}}_t \cup \mathcal{P}_{t+1}$, define

$$\begin{aligned} f_i \rhd_t f_j \iff & e_t(f_i) \preceq_{\mathcal{C}} e_t(f_j) \ \wedge\ \Gamma_t(f_i) \geq \Gamma_t(f_j) \\ & \wedge \left[\exists k \in J_{\mathcal{C}} : e_{t,k}(f_i) < e_{t,k}(f_j) - \delta_k \ \vee\ \Gamma_t(f_i) > \Gamma_t(f_j) \right], \end{aligned} \tag{10}$$

where $\delta_k \geq 0$ is the contract-defined equivalence tolerance for empirical dimension k. f_i dominates f_j only if it is non-inferior in every active empirical dimension, is no worse in scientific-constraint support, and strictly improves either at least one empirical dimension beyond its tolerance or the mechanism-support score. We update the frontier as $\mathcal{F}_{t+1} = \mathrm{ND}_{\rhd_t}(\bar{\mathcal{F}}_t \cup \mathcal{P}_{t+1})$, where $\mathrm{ND}_{\rhd_t}(\cdot)$ returns the nondominated subset under $\rhd_t$. A retained incumbent remains eligible for Pareto selection even when the current proposal is rejected. Retention does not guarantee survival: an incumbent may be removed by dominance or contract-defined frontier-capacity truncation. This separation prevents empirical accuracy from compensating for validity failures while allowing evidence-induced scientific constraints to be strengthened, revised, or withdrawn. Candidate expansion, incumbent re-screening, and the contract-conditioned Pareto update are detailed in Appendix B.3.

At the next iteration, the updated state and trajectory evidence recompile $\mathcal{M}_{t+1}$. CoTraS identifies structural directions worth pursuing or avoiding, while MA-HASE turns them into an evidence-controlled cycle of scientific-constraint induction, candidate generation, deterministic admission, and joint selection. The complete iterative procedure is given in Appendix B.5.

3.3 Evidence-preserving Support

EvoTrace further preserves the evidence required by CoTraS and MA-HASE over long search horizons through three lightweight supporting mechanisms. **Counterfactual attribution** admits a structural edit into positive search guidance only when its paired intervention yields an evidence-supported, valid, and contract-compliant improvement that also passes the active MA-HASE admissibility gate. **Failure-family archival** stores negative transitions using canonical formula-family fingerprints so that recurring failures can be reused by subsequent search cycles. **Scientific search-state distillation** compresses the updated frontier, positive and negative trajectories, failure archive, unresolved hypotheses, and contract-relevant state into Z_{t+1}, which conditions the next symbolic experiment. Formal update rules for these supporting components are provided in Appendix B.4.

4 Experiments

This section presents a comprehensive evaluation of EvoTrace against representative symbolic-regression baselines across five space-physics tasks and three LLM backbones. We first introduce the experimental setup (Sec. 4.1), followed by predictive accuracy and overall performance (Sec. 4.2), task-specific extrapolation and physical consistency (Sec. 4.3), and component ablations (Sec. 4.4).

4.1 Experimental Setup

Benchmarks. We evaluate five real-world space-physics tasks from PhyE2E (Ying et al., 2025): solar differential rotation, sunspot forecasting, plasma-sheet pressure, Fe X 174/175 Å contribution functions, and lunar-tide radial electric field; full task and data details are provided in Appendix C.1.

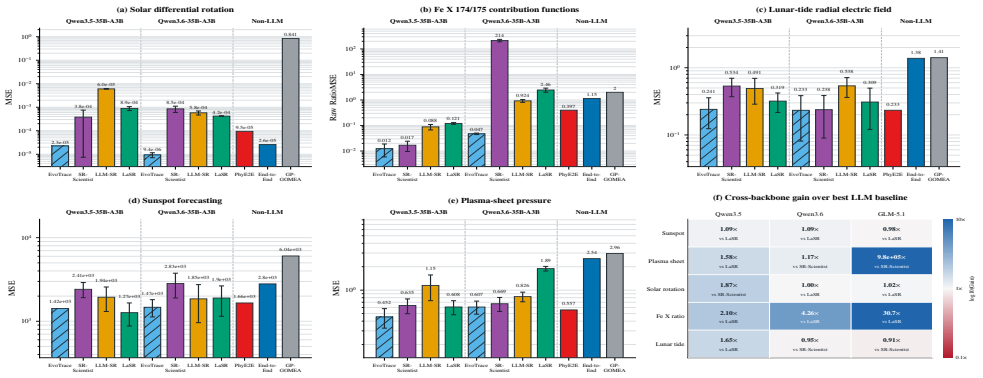

Figure 3: Predictive performance across five space-physics tasks. Panels (a–e) compare EvoTrace with LLM-based baselines under Qwen3.5-35B-A3B and Qwen3.6-35B-A3B, together with representative non-LLM methods; panel (f) reports Gain over the strongest LLM baseline across three backbones. Error bars denote mean $\pm$ sample standard deviation over three independent runs.

Baselines. We compare against 14 non-LLM SR methods and three LLM-based baselines: LLM-SR (Shojaee et al., 2025), LaSR (Grayeli et al., 2024), and SR-Scientist (Xia et al., 2026). All LLM-based methods are evaluated with Qwen3.5-35B-A3B, Qwen3.6-35B-A3B, and GLM-5.1 (Qwen Team, 2026a;b; GLM-5 Team, 2026). For each task–backbone configuration, we independently run each stochastic LLM-based method three times with different random seeds while keeping the data split, inference configuration, evaluator, and search budget fixed. Unless otherwise stated, the quantitative results reported in the main text and tables are averaged over the three runs. The full baseline set is listed in Appendix C.2.

Metrics. We report MSE and, where applicable, R^2, and measure expression simplicity by normalized AST node count. We further evaluate task-specific extrapolation robustness and physical consistency. For the compact summary in Fig. 3(f), we combine accuracy ($\mathcal{A}$), extrapolation ($\mathcal{E}$), physical consistency ($\mathcal{P}$), and AST simplicity ($\mathcal{S}$) as

$$\mathcal{Q} = \mathcal{A}^{0.40}\mathcal{E}^{0.30}\mathcal{P}^{0.20}\mathcal{S}^{0.10}, \qquad \text{Gain} = \frac{\mathcal{Q}_{\text{EvoTrace}}}{\mathcal{Q}_{\text{best LLM baseline}}}. \tag{11}$$

We interpret $\text{Gain} > 1$ as a higher composite score than the strongest LLM baseline under the same task and backbone. We provide the full normalization, unavailable-dimension handling, and weight-sensitivity protocol in Appendix D.3.

Implementation. We use SGLang for Qwen and an authenticated API for GLM-5.1, with temperature 0.7. Non-LLM methods are limited to 32,500 candidates per task and LLM-based methods to 325 calls/candidates; further implementation and budget details are in Appendix C.3–C.4.

4.2 Predictive Accuracy and Overall Performance

As shown in Fig. 3(a–e), under Qwen3.5, EvoTrace achieves a mean rank of **1.4**, ranking first on three tasks and second on the remaining two. Relative to the strongest LLM baseline, it reduces the error by **94.1%**, **27.0%**, **24.6%**, and **25.7%** on differential rotation, Fe X, lunar tide, and plasma-sheet pressure, respectively, while trailing the best sunspot MSE by only **11.8%**. Under Qwen3.6, EvoTrace achieves the lowest primary error among the LLM methods on all five tasks. Run-to-run variation is also small across both backbones; notably, under Qwen3.5, all three random-seed runs recover the same formula on solar differential rotation and sunspot forecasting. These results suggest that EvoTrace is relatively robust to both backbone and stochastic search variation.

Figure 3(f) summarizes the multi-criteria trade-off among predictive accuracy, extrapolation robustness, physical consistency, and AST simplicity. Across the three backbones, **12 of 15** task–backbone combinations achieve $\text{Gain} \geq 1$, including all five tasks under Qwen3.5. The three remaining settings are close to parity: Sunspot under GLM-5.1 reaches $0.98\times$, while lunar tide reaches $0.95\times$

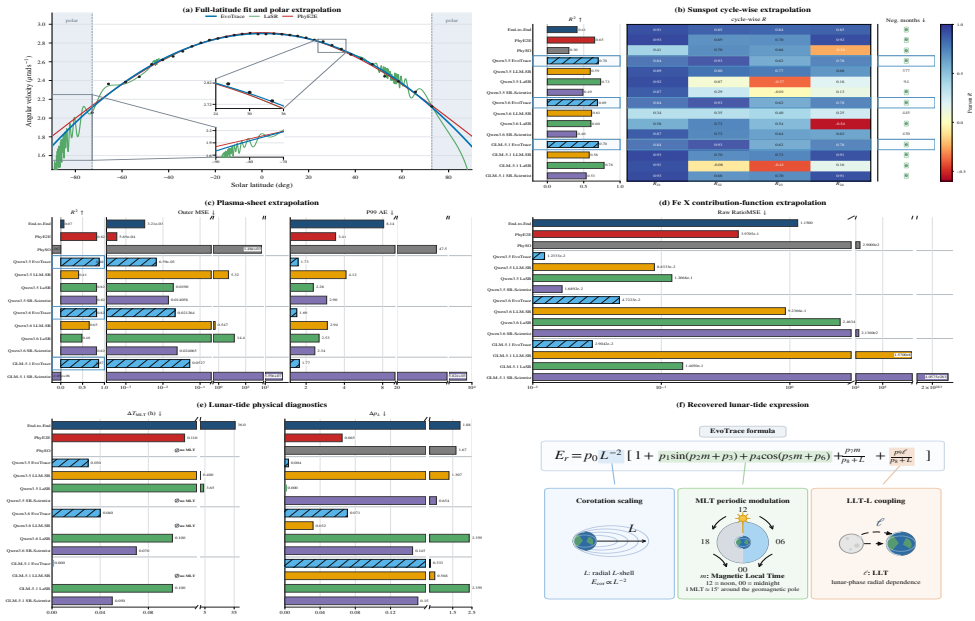

Figure 4: Task-specific extrapolation and physical diagnostics across the five space-physics benchmarks. Panels (a–e) summarize task-level diagnostics; (f) shows the recovered lunar-tide expression and its L^{-2}, MLT-periodic, and LLT-dependent structure.

Table 1: Solar differential rotation on the seven-point sealed holdout. $R_{\mathrm{S}}^{\mathrm{sm.}}$ and $R_{\mathrm{N}}^{\mathrm{sm.}}$ are polar correlations with the Hotta–Kusano simulation; full results are in the Appendix.

Method	MSE ↓	R^2 ↑	AST Complexity ↓	$R_{\mathrm{S}}^{\mathrm{sm.}}$ ↑	$R_{\mathrm{N}}^{\mathrm{sm.}}$ ↑	Δ_{pole}
Representative non-LLM baselines						
AI Feynman	8.38×10^{-5}	0.9980	10	0.9651	0.9713	0.0000
End-to-End	2.59×10^{-5}	0.9994	16	0.9862	0.9900	0.0096
PhyE2E	9.47×10^{-5}	0.9978	7	0.9824	0.9868	0.0000
Qwen3.5-35B-A3B						
EvoTrace	$\mathbf{2.26\times10^{-5}}$	**0.9995**	**17**	0.9861	0.9900	0.0051
LLM-SR	6.05×10^{-3}	0.8573	156	0.9946	0.9738	5.4962
LaSR	8.94×10^{-4}	0.9789	28	0.9328	0.9918	$+\infty$
SR-Scientist	3.81×10^{-4}	0.9910	254	**0.9949**	**0.9997**	5.0571
Qwen3.6-35B-A3B						
EvoTrace	$\mathbf{9.39\times10^{-6}}$	**0.9998**	40	**0.9850**	**0.9875**	0.0459
LLM-SR	5.80×10^{-4}	0.9863	216	0.9845	0.9763	2.2230
LaSR	4.22×10^{-4}	0.9900	**30**	0.8387	0.9814	0.1111
SR-Scientist	8.55×10^{-4}	0.9798	78	0.9457	0.9599	0.7117
GLM-5.1						
EvoTrace	$\mathbf{2.19\times10^{-5}}$	**0.9995**	**24**	**0.9840**	**0.9900**	0.0278
LLM-SR	1.18×10^{-4}	0.9972	115	0.1475	0.9669	5.85×10^{15}
LaSR	6.42×10^{-5}	0.9985	28	0.9689	0.9812	0.0328
SR-Scientist	3.06×10^{-3}	0.9279	79	0.9374	0.9394	0.2653

and $0.91\times$ under Qwen3.6 and GLM-5.1, respectively. Overall, the results indicate that EvoTrace's multi-criteria advantages largely persist across different underlying LLMs.

4.3 Extrapolation and Physical Consistency

Solar differential rotation. We test extrapolation from observed latitudes toward the unobserved poles, assessing whether the recovered expressions retain the established faster-equator/slower-pole profile (Snodgrass, 1983). Table 1 shows that EvoTrace achieves the lowest MSE and highest R^2 within all three LLM-backbone blocks. Fig. 4(a) compares the frozen predictions with the independent Hotta–Kusano simulation (Hotta & Kusano, 2021a;b). Across the three backbones, the

smoothed polar correlations range from 0.9840 to 0.9861 in the south and from 0.9875 to 0.9900 in the north, supporting agreement in polar trends. Unlike the symmetric polar predictions of AI Feynman and PhyE2E, EvoTrace permits a small north–south difference. This is qualitatively compatible with reported hemispheric asymmetry (Hathaway et al., 1996).

Sunspot forecasting. We next evaluate long-horizon extrapolation over Solar Cycles 21–24, covering four unseen cycles and 522 months (WDC–SILSO, Royal Observatory of Belgium, 2026; Clette & Lefèvre, 2016; Hathaway, 2015). As shown in Fig. 4(b), EvoTrace increases avg-R from LaSR's 0.2008 to **0.7915** while reducing negative predictions from **94 months to zero**. These results indicate that EvoTrace preserves cycle fidelity and physical admissibility over extended horizons.

Plasma-sheet pressure. We next consider multi-condition spatial shift (Wang et al., 2013). As shown in Fig. 4(c), EvoTrace maintains low outer-region and tail errors across backbones, reducing Outer MSE/P99 AE by **53.1%/24.1%** under Qwen3.5 and **11.6%/27.8%** under Qwen3.6 relative to the strongest LLM baselines. Under GLM-5.1, it also avoids SR-Scientist's severe error growth, showing that its gains extend beyond average accuracy to outer-region and tail robustness.

Fe X 174/175 Å contribution functions. We next test whether multi-output formulas preserve the coupled diagnostic relationship between the Fe X 174/175 emission lines (Keenan et al., 2008). As shown in Fig. 4(d), under Qwen3.5 LaSR attains lower line-wise MSE, yet its Raw RatioMSE is 1.2068×10^{-1}, **9.8**$\times$ larger than EvoTrace's 1.2333×10^{-2}; the advantage in ratio recovery persists across the backbones. These results show that EvoTrace better preserves the meaningful coupled diagnostic relation between the two emission lines rather than optimizing their fits in isolation.

Lunar-tide radial electric field. Finally, we examine the physical structure of the expressions recovered for the lunar-tide radial electric field. Under ideal corotation, $v = \Omega_E r \propto L$ and $B \propto L^{-3}$, yielding $E_{\mathrm{cor}} \propto L^{-2}$, while observations reveal lunar-tide modulation and MLT-dependent corotation lag (Xiao et al., 2023; Maus, 2017; Zhang et al., 2024). As shown in Fig. 4(e), the identifiable dominant MLT periods of EvoTrace's frozen expressions deviate from the 12-hour reference by **0.030**, **0.040**, and **0.000** hours across the three backbones, compared with 0.400, 0.070, and 0.050 hours for the best identifiable LLM baseline in each setting. The representative expression in Fig. 4(f) contains an L^{-2} prefactor, MLT harmonic terms, and an LLT-dependent term whose amplitude varies with L. These components provide an interpretable representation, but the additional radial dependence and nonperiodic local-time terms mean that the full expression need not obey an exact L^{-2} law or global periodicity. Full results for all five tasks are provided in Appendix D.

4.4 Ablation Study

We ablate CoTraS, MA-HASE, and Evidence-preserving Support on tasks. Removing CoTraS increases sunspot MSE by $2.81\times$ and Fe X RatioMSE by $11.9\times$, despite lower Fe X line-wise errors, showing that trajectory evidence favors structurally useful revisions over isolated fit improvements. Removing MA-HASE causes severe degradation: differential-rotation MSE increases by $20.8\times$, the polar gap rises from 0.0051 to 1.8273, and Fe X RatioMSE becomes numerically unstable, highlighting the role of mechanism-anchored admissibility. Without Evidence-preserving Support, sunspot avg-R drops from 0.7915 to 0.0198 and the differential-rotation polar gap rises to 1.4765. Together, these results confirm the complementary roles of trajectory steering, mechanism-guided admissibility, and cross-iteration evidence reuse. Results are provided in Appendix D.4.

5 Conclusion

We introduced EvoTrace, an agentic symbolic-regression framework that turns accumulated search trajectories into reusable guidance for hypothesis evolution. By connecting structural edits with evaluation evidence and revisable scientific constraints, EvoTrace provides a structured way to learn from both successful and unsuccessful hypotheses. Experiments across five space-physics tasks and three LLM backbones demonstrate its effectiveness in improving predictive accuracy and task-specific scientific diagnostics across multiple settings. These results highlight the value of preserving and reusing search evidence for agentic scientific discovery. Future work will extend this approach to broader scientific domains and investigate more efficient cross-task reuse of search experience.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript translation, language editing, and organization suggestion. They were also used to assist with LaTeX/code editing. All AI-assisted content was reviewed by the authors, and the reported experimental results, quantitative comparisons, and scientific claims were checked against the underlying experiments and data. The authors take full responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details required to reproduce EvoTrace and its evaluation. Appendices A and B specify the complete CoTraS and MA-HASE procedures, evidence-preserving components, search-state construction, and stopping criteria. Appendix C documents the datasets, preprocessing, data splits, baseline implementations, LLM backbones, inference settings, search budgets, hardware configuration, and task-specific evaluation protocols. Appendix D provides the complete quantitative results, representative discovered expressions, composite-metric construction and sensitivity analysis, and component ablations. These materials complement the algorithmic definitions and experimental protocol described in the main paper. We plan to release the complete EvoTrace implementation upon publication, with the broader aim of providing a reproducible methodological reference and a practical foundation for researchers to further develop trajectory-aware, mechanism-guided approaches to scientific discovery.

REFERENCES

Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurélien Lucchi, and Giambattista Parascandolo. Neural symbolic regression that scales. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 936–945. PMLR, 2021. URL https://proceedings.mlr.press/v139/biggio21a.html.

Bogdan Burlacu, Gabriel Kronberger, and Michael Kommenda. Operon C++: An efficient genetic programming framework for symbolic regression. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*, pp. 1562–1570. Association for Computing Machinery, 2020. doi: 10.1145/3377929.3398099. URL https://doi.org/10.1145/3377929.3398099.

Frédéric Clette and Laure Lefèvre. The new sunspot number: Assembling all corrections. *Solar Physics*, 291:2629–2651, 2016. doi: 10.1007/s11207-016-1014-y.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl, 2023. URL https://arxiv.org/abs/2305.01582.

Mengge Du, Yuntian Chen, Zhongzheng Wang, Longfeng Nie, and Dongxiao Zhang. LLM4ED: Large language models for automatic equation discovery, 2024. URL https://arxiv.org/abs/2405.07761.

GLM-5 Team. GLM-5: From vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Arya Grayeli, Atharva Sehgal, Omar Costilla-Reyes, Miles Cranmer, and Swarat Chaudhuri. Symbolic regression with a learned concept library. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44678–44709, 2024. doi: 10.52202/079017-1419. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4ec3ddc465c6d650c9c419fb91f1c00a-Abstract-Conference.html.

David H. Hathaway. The solar cycle. *Living Reviews in Solar Physics*, 12(4), 2015. doi: 10.1007/lrsp-2015-4.

David H. Hathaway, Peter A. Gilman, John W. Harvey, Frank Hill, Robert F. Howard, Harrison P. Jones, James C. Kasher, John W. Leibacher, John Pintar, and George W. Simon. GONG observations of solar surface flows. *Science*, 272(5266):1306–1309, 1996. doi: 10.1126/science.272.5266.1306.

Hideyuki Hotta and Kanya Kusano. Solar differential rotation reproduced with high-resolution simulation. *Nature Astronomy*, 5:1100–1102, 2021a. doi: 10.1038/s41550-021-01459-0.

Hideyuki Hotta and Kanya Kusano. Solar differential rotation reproduced with high-resolution simulation, 2021b. Data accompanying Hotta and Kusano (2021).

Ying Jin, Weilin Fu, Jian Kang, Jiadong Guo, and Jian Guo. Bayesian symbolic regression, 2019. URL https://arxiv.org/abs/1910.08892.

Pierre-Alexandre Kamienny, Stéphane d'Ascoli, Guillaume Lample, and François Charton. End-to-end symbolic regression with transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10269–10281, 2022. doi: 10.52202/068431-0746. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/42eb37cdbefd7abae0835f4b67548c39-Abstract-Conference.html.

Francis P. Keenan, David B. Jess, Kanti M. Aggarwal, Roger J. Thomas, Jeffrey W. Brosius, and Joseph M. Davila. Emission lines of Fe X in active region spectra obtained with the solar extreme-ultraviolet research telescope and spectrograph. *Monthly Notices of the Royal Astronomical Society*, 389:939–948, 2008. doi: 10.1111/j.1365-2966.2008.13615.x.

John R. Koza. *Genetic Programming II: Automatic Discovery of Reusable Programs*. MIT Press, Cambridge, MA, 1994. ISBN 9780262111898.

Mikel Landajuela, Chak Shing Lee, Jiachen Yang, Ruben Glatt, Claudio P. Santiago, Ignacio Aravena, Terrell N. Mundhenk, Garrett Mulcahy, and Brenden K. Petersen. A unified framework for deep symbolic regression. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33985–33998, 2022. doi: 10.52202/068431-2463. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dbca58f35bddc6e4003b2dd80e42f838-Abstract-Conference.html.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–arnold networks. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/afaed89642ea100935e39d39a4da602c-Abstract-Conference.html.

Stefan Maus. A corotation electric field model of the earth derived from Swarm satellite magnetic field measurements. *Journal of Geophysical Research: Space Physics*, 122(8):8733–8754, 2017. doi: 10.1002/2017JA024221.

Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Cláudio Prata Santiago, Sookyung Kim, and Joanne Taery Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations*, 2021. URL https://mlanthology.org/iclr/2021/petersen2021iclr-deep/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

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

Philipp Scholl, Katharina Bieker, Hillary Hauger, and Gitta Kutyniok. ParFam – (neural guided) symbolic regression via continuous global optimization. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ca98452d4e9ecbc18c40da2aa0da8b98-Abstract-Conference.html.

Parshin Shojaee, Kazem Meidani, Amir Barati Farimani, and Chandan K. Reddy. Transformer-based planning for symbolic regression. In *Advances in Neural Information Processing Systems*, volume 36, pp. 45907–45919, 2023. doi: 10.52202/075280-1990. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8ffb4e3118280a66b192b6f06e0e2596-Abstract-Conference.html.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. LLM-SR: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/28df8e730c054c5331855fd4d5403ba9-Abstract-Conference.html.

Herschel B. Snodgrass. Magnetic rotation of the solar photosphere. *The Astrophysical Journal*, 270: 288–299, 1983. doi: 10.1086/161121.

Trevor Stephens. gplearn: Genetic programming in python, with a scikit-learn inspired api. Software, 2015. URL https://github.com/trevorstephens/gplearn.

Wassim Tenachi, Rodrigo Ibata, and Foivos I. Diakogiannis. Deep symbolic regression for physics guided by units constraints: Toward the automated discovery of physical laws. *The Astrophysical Journal*, 959(2):99, 2023. doi: 10.3847/1538-4357/ad014c. URL https://doi.org/10.3847/1538-4357/ad014c.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631. URL https://doi.org/10.1126/sciadv.aay2631.

Marco Virgolin, Tanja Alderliesten, Cees Witteveen, and Peter A. N. Bosman. Improving model-based genetic programming for symbolic regression of small expressions. *Evolutionary Computation*, 29(2):211–237, 2021. doi: 10.1162/evco_a_00278.

Chih-Ping Wang, Chao Yue, Sorin Zaharia, Xiaoyan Xing, Larry Lyons, Vassilis Angelopoulos, Tsugunobu Nagai, and Anthony T. Y. Lui. Empirical modeling of plasma sheet pressure and three-dimensional force-balanced magnetospheric magnetic field structure: 1. observation. *Journal of Geophysical Research: Space Physics*, 118(10):6154–6165, 2013. doi: 10.1002/jgra.50585.

WDC–SILSO, Royal Observatory of Belgium. International sunspot number, version 2.0. https://www.sidc.be/SILSO/datafiles, 2026. Accessed: 2026-09-18.

Shijie Xia, Yuhan Sun, and Pengfei Liu. SR-Scientist: Scientific equation discovery with agentic ai. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KBN6oUx5uL.

Chao Xiao, Fei He, Quanqi Shi, et al. Evidence for lunar tide effects in earth's plasmasphere. *Nature Physics*, 19(4):486–491, 2023. doi: 10.1038/s41567-022-01882-8.

Jie Ying, Haowei Lin, Chao Yue, Yajie Chen, Chao Xiao, Quanqi Shi, Yitao Liang, Shing-Tung Yau, Yuan Zhou, and Jianzhu Ma. A neural symbolic model for space physics. *Nature Machine Intelligence*, 7(10):1726–1741, 2025. doi: 10.1038/s42256-025-01126-3. URL https://doi.org/10.1038/s42256-025-01126-3.

Zhao Zhang, Wenlong Liu, Dianjun Zhang, and Jinbin Cao. Estimating the corotation lag of the plasmasphere based on the electric field measurements of the Van Allen Probes. *Advances in Space Research*, 73(1):758–766, 2024. doi: 10.1016/j.asr.2023.10.022.

APPENDIX

A CoTraS Method Details **14**

A.1 Scientific Contract and Search State 14

A.2 Trajectory Matching, Filtering, and Evidence Construction 14

A.3 Compilation of Contrastive Search Directives 15

B MA-HASE, Evidence Preservation, and the Full Search Procedure **16**

B.1 Compilation of the Mechanism-Anchored Context 16

B.2 Candidate Admissibility and Mechanism-Consistency Validation 17

B.3 Candidate Expansion and Contract-Driven Selection 18

B.4 Evidence-preserving Support 19

B.5 Complete EvoTrace Algorithm and Stopping Criteria 20

B.6 Structured Runtime Records and a Real Search Trajectory 20

C Experimental Setup and Reproducibility **22**

C.1 Benchmarks, Data Sources, and Data Splits 22

C.2 Baseline Methods and Implementation Details 22

C.3 LLM Inference Backends, Hardware, and Runtime Configuration 24

C.4 Search Budgets, Stopping Rules, and Computational Cost 24

C.5 Task-Specific Evaluation and Candidate-Selection Protocols 25

D Additional Experimental Results and Analyses **26**

D.1 Complete Quantitative Results 26

D.2 Representative Discovered Expressions 27

D.3 Composite-Metric Normalization and Weight Sensitivity 28

D.4 Component Ablation Analysis 34

E Representative Symbolic Regression Paradigms **36**

E.1 Limitations and Future Directions 37

A CoTraS Method Details

A.1 Scientific Contract and Search State

CoTraS conditions every search decision on an explicit scientific contract rather than an unrestricted scalar objective. For problem k, we write

$$\mathcal{C}_k = \left(\mathcal{D}_k^{\mathrm{vis}}, \mathcal{O}_k, \mathcal{E}_k, \mathcal{K}_k, \Sigma_k, \mathcal{U}_k\right), \tag{A.1}$$

where $\mathcal{D}_k^{\mathrm{vis}}$ is the data visible during search, $\mathcal{O}_k$ specifies the admissible operators and expression grammar, $\mathcal{E}_k$ specifies fitting, internal evaluation, and evidence construction, $\mathcal{K}_k$ contains the search and AST-complexity budgets, Σ_k records the supplied variable semantics, and $\mathcal{U}_k$ records the supplied physical units. The contract defines what the search may observe and how candidates are evaluated; it does not contain a target equation, a preferred formula family, or a predeclared scientific mechanism. Sealed test and out-of-distribution targets are excluded from $\mathcal{D}_k^{\mathrm{vis}}$ and are never exposed to the Parent Agent, trajectory labels, mechanism induction, or the stopping rule.

The operator contract permits one symbolic expression per declared output; a multi-output problem is represented as a fixed tuple of such expressions. Static validation rejects malformed signatures, unsafe syntax, unsupported operators or calls, and undeclared variables. Numerical validation fits free constants using only $\mathcal{D}_k^{\mathrm{vis}}$ and rejects candidates that fail to execute or produce non-finite predictions. Dimensional validation uses $\mathcal{U}_k$ to check algebraic compatibility when units are available; it is a deterministic validity test rather than a learned mechanism constraint.

At iteration t, the search state is

$$Z_t = \left(\mathcal{F}_t, \mathcal{L}_t, \mathcal{A}_t, \mathcal{O}_t^{\mathrm{recent}}, \mathcal{H}_t^{\mathrm{unres}}, b_t\right), \tag{A.2}$$

where $\mathcal{F}_t$ is the active admissible frontier, $\mathcal{L}_t$ is the append-only formula ledger, $\mathcal{A}_t$ is the failure and repetition archive, $\mathcal{O}_t^{\mathrm{recent}}$ contains recent evaluator observations, $\mathcal{H}_t^{\mathrm{unres}}$ contains unresolved scientific relations, and b_t is the remaining search budget. Only a bounded projection of Z_t is exposed to the Parent Agent. The full ledger and provenance remain available to deterministic framework components, while the raw conversational transcript is not used as scientific memory.

The protected baseline f_t^{base} is selected from $\mathcal{F}_t$ by a deterministic policy $\pi_{\mathcal{C}}^{\mathrm{base}}$ fixed by the contract. The policy chooses the reference expression for the next paired edit; it does not replace the Pareto update of $\mathcal{F}_t$ and does not guarantee that the baseline will be the final frozen expression.

A.2 Trajectory Matching, Filtering, and Evidence Construction

Paired transitions. For each evaluator-grounded edit, CoTraS stores

$$\xi_t = \left(Z_t, f_t^{\mathrm{base}}, a_t, \Delta_t, f_t', \mathbf{e}_t(f_t^{\mathrm{base}}), \mathbf{e}_t(f_t')\right),$$

together with the baseline identifier, normalized baseline expression, edit type, changed-term identifiers, proposed expression, fitted constants, and pre- and post-edit evidence. A proposal is evaluated against its protected baseline under the same contract. Prose-only recommendations are not treated as positive transitions.

Comparable contexts. Let $h(\mathcal{C})$ denote a hash of the visible data boundary, metric policy, operator grammar, evaluator version, and fixed budget regime. Two transitions are aggregated only when

$$\begin{aligned}\mathrm{Match}(\xi_i, \xi_j) = \mathbb{I}[h(\mathcal{C}_i) = h(\mathcal{C}_j)]\mathbb{I}[r_i = r_j] \\ \times \mathbb{I}[v_i = v_j]\mathbb{I}[d_i = d_j] = 1,\end{aligned} \tag{A.3}$$

where r is the baseline search role, v is the variable set available to the edit, and d is the edit class. Exact baseline identity is required for the paired comparison that creates a single transition; the weaker role-level match above is used only when evidence is aggregated across iterations. This prevents evidence collected under a different visibility, evaluation, or search regime from being reused as a valid contrast.

Evidence dimensions. The empirical evidence vector in the main text is

$$\mathbf{e}_t(f) = [\mathcal{L}_{\text{fit}}(f), \mathcal{L}_{\text{cv}}(f), c(f), 1 - q_{\text{num}}(f)]\,.$$

$\mathcal{L}_{\text{fit}}$ is the primary train-side metric after orienting it as a loss, and $\mathcal{L}_{\text{cv}}$ is the corresponding internal-validation loss when an internal split is enabled. An unavailable dimension is masked rather than imputed from sealed data. The complexity $c(f)$ is the normalized AST node count; for a multi-output candidate it is the sum over output expressions.

Let $\mathcal{X}_{\mathcal{C}}^{\text{num}}$ be the train-visible numerical probe set and $m_{\mathcal{C}}$ the number of declared outputs. Numerical validity is

$$\begin{aligned} q_{\text{num}}(f) = \mathbb{I}[\text{Exec}_{\mathcal{C}}(f)] &\frac{1}{|\mathcal{X}_{\mathcal{C}}^{\text{num}}|} \\ &\times \sum_{x \in \mathcal{X}_{\mathcal{C}}^{\text{num}}} \mathbb{I}[f(x) \in \mathbb{R}^{m_{\mathcal{C}}},\ \|f(x)\|_\infty < \infty]\,. \end{aligned} \tag{A.4}$$

Thus $q_{\text{num}}(f) = 1$ requires successful execution and finite predictions on every required numerical probe.

The matrix $\mathbf{D}_{\mathcal{C}}$ in Eq. (2) of the main text is diagonal. Its j-th entry is $m_j\omega_j$, where $m_j \in \{0, 1\}$ masks an unavailable evidence dimension and $\omega_j \geq 0$ is a contract-fixed scale. Active scales are normalized to sum to one. The small constant ϵ stabilizes near-zero baselines and clipping at ρ prevents a single near-zero denominator from dominating the transition utility. These quantities are fixed independently of sealed evaluation results.

Validity, failure, violation, and repetition. The candidate-level validity predicate used by CoTraS is

$$\begin{aligned} \text{Valid}_{\mathcal{C}}(f) = \mathbb{I}[\text{ASTValid}_{\mathcal{C}}(f)]&\mathbb{I}[q_{\text{num}}(f) = 1] \\ &\times \mathbb{I}[\text{DimValid}_{\mathcal{C}}(f)]\mathbb{I}[c(f) \leq b_{\mathcal{C}}] \\ &\times \mathbb{I}[\text{MetricFinite}_{\mathcal{C}}(f)], \end{aligned} \tag{A.5}$$

$$\text{Viol}_{\mathcal{C}}(\xi) = \mathbb{I}[\text{ContractViol}_{\mathcal{C}}(\xi)], \tag{A.6}$$

$$\text{Fail}(\xi) = \mathbb{I}[\text{EvalFail}_{\mathcal{C}}(\xi)]. \tag{A.7}$$

$\text{ContractViol}_{\mathcal{C}}$ covers violations of visibility, operator, evaluation, or budget rules. $\text{EvalFail}_{\mathcal{C}}$ covers execution failures, unscored results, invalid candidates, and contract-defined poor outcomes. State-dependent structural novelty is not part of $\text{Valid}_{\mathcal{C}}$; it is applied later by MA-HASE so that CoTraS can still record an attempted repeat as negative trajectory evidence.

Let $\eta(f)$ denote the exact fingerprint of the normalized AST and $\chi(f)$ the canonical formula-family fingerprint. The explicit repetition criterion is

$$\text{Repeat}(\xi_t) = \mathbb{I}[\eta(f'_t) \in \mathcal{X}_{<t} \ \vee\ \chi(f'_t) \in \mathcal{Y}_t^{\text{group}}]\,, \tag{A.8}$$

where $\mathcal{X}_{<t}$ contains previously attempted exact fingerprints and $\mathcal{Y}_t^{\text{group}}$ contains formula-family fingerprints already proposed within the same coordinated review group. By contrast, R_t^{rep} in Eq. (3) is a graded statistic derived from the frequency, similarity, and recency of matching failed exact or family-level fingerprints.

Contrastive labeling. The operational label assignment follows Eq. (4) of the main text. Explicit repeats, contract violations, and evaluation failures yield negative evidence regardless of utility. Among the remaining valid transitions, $\tau_+ > \tau_-$ defines a non-commitment region. After ξ_t is evaluated, $y_t = +1$ appends ξ_t to $\mathcal{T}_{t+1}^+$, $y_t = -1$ appends it to $\mathcal{T}_{t+1}^-$, and $y_t = 0$ leaves it uncommitted. Equivalence tolerances and minimum improvement requirements are fixed by the contract and never use sealed metrics.

A.3 Compilation of Contrastive Search Directives

The compiler in Eq. (5) converts matched trajectory evidence into a bounded, traceable directive rather than a free-form summary. Each retained item stores its supporting transition identifiers, contract hash, evaluator outcome, and expiry iteration; unsupported reviewer statements are excluded.

Preferred operations. For a structural operation o, define

$$s_t^+(o) = \sum_{\xi_\tau \in \mathcal{T}_t^+} \mathbb{I}[\mathrm{op}(\xi_\tau) = o]\, \mathbb{I}[A_\tau^{\mathrm{cf}} = 1]\, \kappa^{t-\tau} \times \max\{U_{\mathcal{C}}(\xi_\tau), 0\}, \qquad 0 < \kappa \leq 1. \tag{A.9}$$

The counterfactual-attribution indicator A_τ^{cf}, defined in Appendix B.4, allows a positive CoTraS transition to influence the next directive only after the candidate also passes the current MA-HASE admission rule. The compiler places an operation in Ω_t^+ only when it has admissible support in the current matched context. Supported operations are edit directions, not instructions to reproduce a previous formula.

Failure-family archive. For each $\xi_\tau \in \mathcal{T}_t^-$, the archive records $\eta(f'_\tau)$, $\chi(f'_\tau)$, the failure reason, contract hash, and iteration index. Exact AST repeats are blocked immediately. A family fingerprint enters $\mathcal{G}_t^-$ only when matched evidence attributes the failure to the structural family rather than to an unrelated execution or context error. Family exclusions may expire or be released by later verified positive evidence under the same contract, preventing one failed parameterization from permanently excluding a useful family.

Analysis and evaluation route. $\mathcal{R}_t$ maps the dominant unresolved evidence deficit to the next admissible analysis. Syntax or execution errors request deterministic diagnosis; repeated poor fits request equation critique; a performance plateau may request exploration of a new structural family; and a data-pattern mismatch may request a train-visible probe. A route may recommend a tool call or reviewer action, but it cannot expose a forbidden split or certify a candidate without evaluator evidence.

Budget and stopping state. $\mathcal{B}_t$ is derived from the remaining budget b_t and contract-fixed stopping criteria. It records the remaining outer iterations, remaining Parent calls, review allowance, success threshold, complexity ceiling, and trigger cooldown. Early stopping is possible only when an active admissible candidate satisfies the contract-owned train-side stopping rule; otherwise search continues until the budget is exhausted.

The serialized directive retains the four components of Eq. (5):

$$\mathrm{Serialize}(\mathcal{Q}_t) = \left\{ \begin{array}{l} \texttt{preferred_operations}, \\ \texttt{excluded_families}, \\ \texttt{evaluation_route}, \\ \texttt{budget_state} \end{array} \right\}. \tag{A.10}$$

Each field stores the evidence identifiers needed to trace a directive back to evaluator output. The resulting $\mathcal{Q}_t$ steers the next structural experiment but does not replace MA-HASE admission or Pareto selection and does not use sealed evaluation results.

B MA-HASE, Evidence Preservation, and the Full Search Procedure

B.1 Compilation of the Mechanism-Anchored Context

MA-HASE receives the search direction $\mathcal{Q}_t$ from CoTraS but does not treat that direction, or any reviewer statement, as a scientific conclusion. Using only the observable state, contract metadata, train-visible evidence, and verified trajectories, it compiles the context in Eq. (6):

$$\mathcal{M}_t = \mathrm{Compile}_{\mathrm{M}}(\mathcal{C}, Z_t, \mathcal{Q}_t) = \left(\mathcal{D}_t, \Psi_t, \mathcal{G}_t^-, b_{\mathcal{C}}\right). \tag{B.1}$$

$\mathcal{D}_t$ contains executable structural proposals; Ψ_t is the set of currently active evidence-qualified scientific relations; $\mathcal{G}_t^-$ is the failure-family archive; and $b_{\mathcal{C}}$ is the AST-complexity ceiling. Each $(\psi, \omega) \in \Psi_t$ pairs a falsifiable relation ψ with a positive evidence-derived support weight ω. Deterministic admission checks remain owned by $\mathcal{C}$ and are not generated by the mechanism reviewer.

Evidence-induced scientific relations. A proposed relation ψ must be expressible as a falsifiable claim about the supplied variables, such as dependence, interaction, periodicity, scaling, monotone or bounded behavior, or a boundary trend. The reviewer may form such a proposal from variable semantics, physical units, train-visible observations, and trajectory evidence, but the proposal remains inactive until the framework can attach both provenance and an executable falsification procedure. Its supporting evidence identifiers are stored in the ledger rather than inside the compact pair (ψ, ω).

The support weight $\omega > 0$ is computed from stored provenance and repeated support of the relation under matched contexts; it is not assigned directly by the reviewer. A relation enters Ψ_t only if its provenance is resolvable, its falsification procedure can be executed using permitted information, and its activation confidence exceeds a contract-fixed threshold. Unsupported or contradictory relations remain in $\mathcal{H}_t^{\text{unres}}$ and cannot constrain candidate admission. This activation rule prevents a single model-generated statement from becoming a mechanism anchor by assertion.

Representation of structural proposals. The executable proposal set is represented as

$$d = (h, s, o, \Delta, E, \ell) \in \mathcal{D}_t, \tag{B.2}$$

where h is the proposed scientific rationale, s identifies its variable scope, o is the operator family, Δ is the concrete structural edit, E contains supporting evidence identifiers, and ℓ is its validity horizon. The proposal can instantiate a relation in Ψ_t, but neither the rationale nor the candidate equation is treated as an evaluation result.

Reviewer-visible evidence. The mechanism reviewer receives only a bounded provenance-preserving packet:

$$\begin{aligned} I_t^{\text{rev}} = \big(&\mathcal{C}^{\text{compact}}, f_t^{\text{base}}, \mathcal{F}_t^{\text{compact}}, \mathcal{Q}_t, \\ &\mathcal{O}_t^{\text{recent}}, \mathcal{A}_t^{\text{compact}}, \mathcal{H}_t^{\text{unres}}, \mathcal{X}_{<t}, b_t\big). \end{aligned} \tag{B.3}$$

The compact contract includes the allowed variables, their semantics and units, operator grammar, visible-data declaration, and remaining budget. Sealed targets, external evaluation labels, hidden Parent reasoning, and the raw conversation transcript are excluded.

The reviewer returns typed relation hypotheses and executable edit proposals. Each executable action is normalized into canonical fields such as `base_formula_id`, `relation_hypothesis`, `falsification_probe`, `structural_edit`, `candidate_equation`, and `evidence_ids`. Constants remain free and are fitted independently by the framework. The reviewer cannot assign trusted metrics, set its own confidence weight, waive deterministic admission checks, or select the final expression.

Schema validation and bounded repair. Before evaluation, the framework checks that a proposal refers to the protected baseline, follows the output schema, uses only declared variables and operators, and contains the required number of symbolic outputs. The exact AST fingerprint $\eta(f)$ must differ from the protected baseline and previously attempted exact variants, while $\chi(f)$ must instantiate the declared family-level edit. A malformed proposal may receive only the bounded repair allowed by the contract; any repaired expression is revalidated from the beginning and still requires independent fitting and evaluation.

B.2 Candidate Admissibility and Mechanism-Consistency Validation

MA-HASE distinguishes current validity from new-proposal novelty and support for evidence-induced scientific relations. For each formula f, define

$$\begin{aligned} V_{\mathcal{C},t}(f) &= \mathbb{I}[\text{Visible}_{\mathcal{C}}(f)]\mathbb{I}[\text{ASTValid}_{\mathcal{C}}(f)] \\ &\quad \times \mathbb{I}[q_{\text{num}}(f) = 1]\mathbb{I}[\text{DimValid}_{\mathcal{C}}(f)]\mathbb{I}[c(f) \le b_{\mathcal{C}}], \\ \text{DetValid}_{\mathcal{C},t}(f) &= V_{\mathcal{C},t}(f)\,\text{Novel}_t(f). \end{aligned} \tag{B.4}$$

The predicate $V_{\mathcal{C},t}$ applies to both new proposals and incumbents. The predicate $\text{DetValid}_{\mathcal{C},t}$ additionally requires novelty and is used only for new-proposal admission. These deterministic checks do not encode a preferred scientific relation.

Scientific-constraint support. When $\Psi_t \neq \varnothing$, the support score in Eq. (7) of the main text is

$$\Gamma_t(f) = \frac{\sum_{(\psi,\omega)\in\Psi_t} \omega \, \mathrm{Sat}(f;\psi)}{\sum_{(\psi,\omega)\in\Psi_t} \omega} - \lambda_{\mathrm{inc}} \, \mathrm{Incomp}(f, \Psi_t).$$

Here, $\mathrm{Sat}(f;\psi) \in [0,1]$ is produced by the executable falsification procedure attached to ψ. The procedure may combine symbolic inspection with train-visible numerical probes, but it must use the same observable boundary as the search. The term $\mathrm{Incomp}(f, \Psi_t) \geq 0$ penalizes candidate-level incompatibility with the jointly active relations, while $\lambda_{\mathrm{inc}} \geq 0$ controls the strength of this penalty. Both Sat and Incomp are evaluator-produced; the reviewer cannot directly set them.

When $\Psi_t = \varnothing$, we define $\Gamma_t(f) = 1$, leaving the scientific-consistency gate inactive. As in Eq. (8) of the main text, incumbent retention is defined by

$$\mathrm{Keep}_{\mathcal{C}}(f \mid \mathcal{M}_t) = \begin{cases} V_{\mathcal{C},t}(f), & \Psi_t = \varnothing, \\ V_{\mathcal{C},t}(f)\, \mathbb{I}[\Gamma_t(f) \geq \gamma_{\mathcal{C}}], & \Psi_t \neq \varnothing. \end{cases}$$

New-proposal admission additionally requires novelty: $\mathrm{Adm}_{\mathcal{C}}(f \mid \mathcal{M}_t) = \mathrm{Keep}_{\mathcal{C}}(f \mid \mathcal{M}_t)\, \mathrm{Novel}_t(f)$. The threshold $\gamma_{\mathcal{C}}$ is fixed before sealed evaluation. Both new proposals and incumbents must satisfy current validity and mechanism-support requirements, but only new proposals undergo the novelty check. Prior evaluation, including identity with the protected baseline, does not itself disqualify an incumbent. A low empirical loss cannot compensate for failed validity or mechanism-support checks, and a reviewer statement cannot activate the scientific gate without the required provenance and executable evidence.

Visibility, numerical validity, and dimensions. $\mathrm{Visible}_{\mathcal{C}}(f) = 1$ only when every datum, statistic, and probe used to generate, fit, or evaluate f is permitted by the contract. The numerical condition $q_{\mathrm{num}}(f) = 1$ requires successful restricted execution, the correct output shape, and finite predictions on all required train-visible probes; fitted constants and required selection metrics must also be finite. Dimensional validity propagates the supplied units through the normalized AST: additions and subtractions require compatible units, multiplicative operations compose units algebraically, and elementary functions that require dimensionless arguments are checked accordingly. Missing unit metadata disables only the corresponding dimensional test; it never supplies a learned scientific relation.

AST validity, complexity, and novelty. The AST validator permits only the grammar declared by $\mathcal{C}$ and rejects control flow, assignment, imports, undeclared calls, and unsafe parameter access. For new proposals, structural novelty is

$$\begin{aligned} \mathrm{Novel}_t(f) = \mathbb{I}[\eta(f) \neq \eta(f_t^{\mathrm{base}})] \mathbb{I}[\eta(f) \notin \mathcal{X}_{<t}] \\ \times \mathbb{I}[\chi(f) \notin \mathcal{Y}_t^{\mathrm{group}}], \end{aligned} \tag{B.5}$$

where $\eta(f)$ is the exact normalized-AST fingerprint and $\chi(f)$ is the canonical formula-family fingerprint. The exact-fingerprint conditions prevent constant-only or exact structural repeats, while the group-level family condition prevents coordinated proposals from repeatedly spending budget on the same declared structural family. This novelty criterion is applied only to new proposals; incumbent re-screening uses $\mathrm{Keep}_{\mathcal{C}}$ without reapplying novelty.

B.3 Candidate Expansion and Contract-Driven Selection

Candidate expansion implements Eq. (9). The preferred operations Ω_t^+ allocate proposal effort, while $\mathcal{D}_t$ supplies evidence-conditioned structural edits. Complete proposals are statically validated, constant-fitted, independently evaluated, and then filtered by the current admission rule:

$$\begin{aligned} \widetilde{\mathcal{P}}_{t+1} &= \mathrm{Expand}(\mathcal{F}_t, \Omega_t^+; \mathcal{D}_t), \\ \mathcal{P}_{t+1} &= \left\{ f \in \widetilde{\mathcal{P}}_{t+1} \;\middle|\; \mathrm{Adm}_{\mathcal{C}}(f \mid \mathcal{M}_t) = 1,\ \chi(f) \notin \mathcal{G}_t^- \right\}. \end{aligned} \tag{B.6}$$

The exact-fingerprint novelty rule and the family-level exclusion $\mathcal{G}_t^-$ serve different purposes: the former prevents direct repetition, whereas the latter excludes a broader structural family only when matched negative evidence supports that exclusion.

Every incumbent is re-screened using the current retention rule:

$$\bar{\mathcal{F}}_t = \{f \in \mathcal{F}_t : \mathrm{Keep}_{\mathcal{C}}(f \mid \mathcal{M}_t) = 1\}. \tag{B.7}$$

This is necessary because Ψ_t is revisable: newly supported relations may strengthen the scientific-consistency gate, while contradicted or expired relations may be removed. Re-screening applies current validity and mechanism-support checks without rejecting incumbents solely because they have already been evaluated. It uses search-visible evidence and never refits incumbents on sealed data.

Let $J_{\mathcal{C}}$ index the active empirical evidence dimensions and $\delta_j \geq 0$ the contract-defined equivalence tolerance. The empirical non-inferiority relation in Eq. (10) is

$$\mathbf{e}_t(f_i) \preceq_{\mathcal{C}} \mathbf{e}_t(f_j) \quad \Longleftrightarrow \quad e_{t,j}(f_i) \leq e_{t,j}(f_j) + \delta_j, \quad \forall j \in J_{\mathcal{C}}.$$

For two formulas in $\bar{\mathcal{F}}_t \cup \mathcal{P}_{t+1}$, dominance additionally requires

$$\Gamma_t(f_i) \geq \Gamma_t(f_j)$$

and at least one strict improvement: either $e_{t,k}(f_i) < e_{t,k}(f_j) - \delta_k$ for some $k \in J_{\mathcal{C}}$, or $\Gamma_t(f_i) > \Gamma_t(f_j)$. This is exactly the operational meaning of Eq. (10). When $\Psi_t = \varnothing$, both candidates have $\Gamma_t = 1$, so dominance reduces to the empirical evidence dimensions.

The corresponding frontier update is

$$\mathcal{F}_{t+1} = \mathrm{ND}_{\triangleright_t}\left(\bar{\mathcal{F}}_t \cup \mathcal{P}_{t+1}\right), \tag{B.8}$$

where $\mathrm{ND}_{\triangleright_t}(\cdot)$ returns the nondominated subset under $\triangleright_t$. If the nondominated set exceeds the contract-owned frontier capacity, deterministic truncation uses the declared empirical ordering, AST complexity, and a stable identifier. All evaluated formulas remain in the append-only ledger even if they are removed from the active frontier. Passing retention makes an incumbent eligible for this update even when the new proposal is rejected; it does not guarantee survival after dominance filtering or capacity truncation.

Frontier maintenance and final formula selection are separated. At termination,

$$f^\star = \mathrm{SelectFinal}_{\pi_{\mathcal{C}}}(\mathcal{F}_T), \tag{B.9}$$

where $\pi_{\mathcal{C}}$ is fixed before sealed evaluation and uses only search-visible empirical evidence, current validity and mechanism support, and complexity information. Novelty is not rechecked during final selection. Sealed and post-hoc scientific diagnostics are computed only after $f^\star$ has been frozen.

B.4 Evidence-preserving Support

Evidence-preserving Support does not introduce another search objective. It retains the provenance needed to decide whether an improvement can be attributed to a structural edit, whether a failure should constrain future proposals, and which parts of the accumulated state must survive context compression.

Counterfactual evidence attribution. For a protected-baseline intervention $\iota_t = (f_t^{\mathrm{base}}, \Delta_t, f_t')$, define

$$A_t^{\mathrm{cf}}(\iota_t) = \mathbb{I}[y_t = +1]\ \mathrm{Adm}_{\mathcal{C}}(f_t' \mid \mathcal{M}_t). \tag{B.10}$$

A positive CoTraS transition therefore influences the next compiled preferred-operation set only when the edited candidate also passes MA-HASE admission for that transition. The attribution record stores the baseline and candidate identifiers, normalized expressions, explicit edit, contract hash, empirical evidence, contrastive label, active relation identifiers, Γ_t, and deterministic admission outcome. It supports a local claim about the tested structural edit without turning the paired comparison into a general causal statement.

Failure-family archival. For each negative transition, the exact-fingerprint archive is updated by

$$\mathcal{A}_{t+1}[\eta(f_t')] = \mathrm{UpdateArchive}\left(\mathcal{A}_t[\eta(f_t')], \chi(f_t'), y_t, \mathrm{Reason}(\xi_t), h(\mathcal{C}), t\right). \tag{B.11}$$

Each entry retains the exact fingerprint, formula-family fingerprint, failure reason, evidence identifiers, contract provenance, occurrence count, recency, and expiry state. A family-level exclusion is promoted into $\mathcal{G}_{t+1}^-$ only when the updated matched evidence attributes repeated failure to that family. Later verified positive evidence may release the exclusion while retaining its historical record.

Scientific search-state distillation. After frontier and archive updates,

$$Z_{t+1} = \text{Distill}\left(\mathcal{C}, \mathcal{F}_{t+1}, \mathcal{T}^{+}_{t+1}, \mathcal{T}^{-}_{t+1}, \mathcal{A}_{t+1}, \mathcal{H}^{\text{unres}}_{t+1}, b_{t+1}\right). \tag{B.12}$$

The distilled state preserves contract-ranked active formulas, compact train-side evidence, active and unresolved scientific relations, verified edits, recent failure reasons, attempted fingerprints, and remaining budget. Full expressions, metrics, relation provenance, and event traces remain in the ledger; only the bounded projection supplied to subsequent model calls is compressed.

The archive and compact state retain explicit provenance:

$$\mathcal{A}_t[\eta] = \left\{ \begin{array}{l} \texttt{family}, \texttt{reasons}, \texttt{evidence_ids}, \\ \texttt{count}, \texttt{last_seen}, \texttt{expiry} \end{array} \right\}, \tag{B.13}$$

and

$$Z_t^{\text{compact}} = \left\{ \begin{array}{l} \texttt{contract}, \\ \texttt{active_candidates}, \\ \texttt{active_relations}, \\ \texttt{unresolved_relations}, \\ \texttt{recent_observations}, \\ \texttt{failure_labels}, \\ \texttt{budget} \end{array} \right\}. \tag{B.14}$$

Evidence-preservation ablation. The *w/o Evidence-preserving Support* condition disables reusable counterfactual-attribution records, durable candidate summaries, active relation provenance, and the historical failure archive in subsequent Parent and reviewer contexts. The append-only ledger and trace remain enabled only for evaluation and audit, while the evaluator, visible-data boundary, metric policy, deterministic validity checks, search budget, and final selector are unchanged. The ablation therefore isolates cross-iteration evidence reuse.

B.5 Complete EvoTrace Algorithm and Stopping Criteria

Algorithm 1 integrates CoTraS, MA-HASE, and Evidence-preserving Support. CoTraS labels paired transitions and compiles search directives; MA-HASE induces evidence-qualified scientific relations, evaluates candidate support, and updates the admissible frontier; the support layer carries verified evidence across iterations.

Early stopping uses only train-side evidence. A formula may trigger stopping only after satisfying the current validity and mechanism-support requirements, surviving frontier selection, and meeting the contract-owned train-side success criterion within the complexity budget. Novelty is checked when a new proposal is admitted, not when a retained incumbent is considered for stopping. If no formula becomes stop-eligible, search continues until the available budget is exhausted. At termination, the contract-owned selector chooses from the final admissible frontier; the selected structure and constants are frozen before sealed evaluation.

B.6 Structured Runtime Records and a Real Search Trajectory

The mathematical objects above are implemented as typed runtime records rather than unrestricted agent dialogue. A review packet contains the contract identifier, protected baseline, active frontier, recent failure summaries, active and unresolved relation records, and evaluator-produced train-visible evidence. The reviewer emits

$$\mathcal{O}_t^{\text{rev}} = \left(\begin{array}{c} \texttt{summary} \\ \texttt{findings} \\ \texttt{relation_hypotheses} \\ \texttt{recommended_actions} \\ \texttt{evidence} \\ \texttt{risk_flags} \end{array} \right), \tag{B.15}$$

where a relation hypothesis must identify a falsifiable relation and supporting evidence, and each recommended action names an explicit structural edit. The framework normalizes executable proposals according to Appendix B.1, validates them, fits their constants, and evaluates them independently.

Algorithm 1 EvoTrace search with CoTraS and MA-HASE

Require: Scientific contract $\mathcal{C}$, grammar $\mathcal{G}$, train-visible data $\mathcal{D}^{\mathrm{vis}}$, budget b_0
Ensure: Frozen formula or formula tuple $f^\star$

Initialize ledger $\mathcal{L}_0$, archive $\mathcal{A}_0$, frontier $\mathcal{F}_0$, unresolved relations $\mathcal{H}_0^{\mathrm{unres}}$, and $\mathcal{T}_0^+ = \mathcal{T}_0^- = \varnothing$
$Z_0 \leftarrow \mathrm{Distill}(\mathcal{C}, \mathcal{F}_0, \varnothing, \varnothing, \mathcal{A}_0, \mathcal{H}_0^{\mathrm{unres}}, b_0)$
$\mathcal{Q}_0 \leftarrow \mathrm{Compile}(Z_0, \varnothing, \varnothing, \mathcal{C})$
for $t = 0, 1, \ldots$ while budget remains **do**
    Select a protected baseline from $\mathcal{F}_t$ using $\pi_{\mathcal{C}}^{\mathrm{base}}$
    $\mathcal{M}_t \leftarrow \mathrm{Compile}_{\mathrm{M}}(\mathcal{C}, Z_t, \mathcal{Q}_t)$, activating only evidence-qualified relations
    $\widetilde{\mathcal{P}}_{t+1} \leftarrow \mathrm{Expand}(\mathcal{F}_t, \Omega_t^+; \mathcal{D}_t)$
    Select the next proposal $f_t' \in \widetilde{\mathcal{P}}_{t+1}$ under the contract-owned proposal order
    Construct and evaluate ξ_t; compute $\mathbf{e}_t(f_t')$, y_t, $\Gamma_t(f_t')$, and $\mathrm{Adm}_{\mathcal{C}}(f_t' \mid \mathcal{M}_t)$
    Set $\mathcal{P}_{t+1} \leftarrow \{f_t'\}$ if $\mathrm{Adm}_{\mathcal{C}}(f_t' \mid \mathcal{M}_t) = 1$ and $\chi(f_t') \notin \mathcal{G}_t^-$; otherwise set $\mathcal{P}_{t+1} \leftarrow \varnothing$
    $\bar{\mathcal{F}}_t \leftarrow \{f \in \mathcal{F}_t : \mathrm{Keep}_{\mathcal{C}}(f \mid \mathcal{M}_t) = 1\}$ ▷ No novelty check for incumbents
    $\mathcal{F}_{t+1} \leftarrow \mathrm{ND}_{\triangleright_t}(\bar{\mathcal{F}}_t \cup \mathcal{P}_{t+1})$
    Apply deterministic frontier truncation if capacity is exceeded
    Update $\mathcal{T}_{t+1}^+$, $\mathcal{T}_{t+1}^-$, counterfactual records, relation provenance, and $\mathcal{A}_{t+1}$
    Distill the updated evidence into Z_{t+1}
    $\mathcal{Q}_{t+1} \leftarrow \mathrm{Compile}(Z_{t+1}, \mathcal{T}_{t+1}^+, \mathcal{T}_{t+1}^-, \mathcal{C})$
    if a formula in $\mathcal{F}_{t+1}$ is stop-eligible under $\mathcal{C}$ **then**
        $T \leftarrow t + 1$; **break**
    end if
end for
If the budget is exhausted, set T to the final completed iteration
$f^\star \leftarrow \mathrm{SelectFinal}_{\pi_{\mathcal{C}}}(\mathcal{F}_T)$
Freeze $f^\star$; then compute sealed, external, and post-hoc scientific diagnostics
return $f^\star$

The runtime decision layer is stored separately:

$$\mathcal{O}_t^{\mathrm{dec}} = \begin{pmatrix} \texttt{mode} \\ \texttt{should_inject} \\ \texttt{should_stop} \\ \texttt{reason} \end{pmatrix}. \tag{B.16}$$

These fields are implementation controls rather than optimization objectives. In particular, `should_inject` becomes effective only after evaluator evidence and MA-HASE admission support the intervention, while `should_stop` cannot override the contract-owned stopping rule.

A real lunar-tide trajectory. We illustrate the control flow with a recorded Qwen3.5-35B-A3B run on the lunar-tide radial electric-field problem. After the first outer iteration, the protected baseline was

$$f_{\mathrm{base}}(L) = \left(\alpha_0 + \frac{\alpha_1}{L}\right)[1 - s(L)] + \alpha_4 s(L), \tag{B.17}$$

$$s(L) = \{1 + \exp[-\alpha_3(L - \alpha_2)]\}^{-1}. \tag{B.18}$$

It attained training MSE 0.662755, $R^2 = 0.559577$, and AST complexity 30. The review packet also contained a previously attempted threshold variant marked as a duplicate, so the reviewer received both the current incumbent and explicit evidence that a nearby structural direction had already been exhausted.

The packet exposed train-side probe results rather than sealed targets. A logged radial probe score ranked $\tanh(L)$ at 0.4968, above the identity (0.3714) and quadratic (0.2914) alternatives. A single-harmonic probe showed a larger MLT signal ($R^2 = 0.1303$) than LLT ($R^2 = 0.0012$), while the estimated interaction contribution was $R^2 = 0.0009$. These observations supplied provenance for a periodic relation hypothesis and the structural action `add_one_periodic`. The framework then independently fitted and evaluated

$$f_{\mathrm{rev}}(L, m) = \beta_0 + \beta_1 \tanh(\beta_2 L) + \beta_3 \sin(\beta_4 m + \beta_5). \tag{B.19}$$

The candidate attained training MSE 0.527079, $R^2 = 0.649738$, and AST complexity 18, corresponding to a 20.5% reduction in training MSE and a 40.0% reduction in AST nodes relative to the protected baseline.

The candidate was still not frozen by the reviewer. Its evaluator-produced evidence determined its CoTraS label and MA-HASE admission outcome, and only then could the verified intervention support the periodic-addition operation in the *next* directive Ω^{+}_{t+1}. The repeated threshold construction remained negative revisit evidence because its exact fingerprint was already known.

A subsequent parent-guided edit introduced an inverse-radial correction,

$$f_{\text{next}}(L, m) = \gamma_0 + \gamma_1 \tanh[\gamma_2(L - \gamma_3)] + \frac{\gamma_4}{L + \gamma_5} + \gamma_6 \sin(\gamma_7 m + \gamma_8), \tag{B.20}$$

reducing training MSE to 0.520108; a later compact refinement reached 0.506723. The final frozen expression was selected only by the contract-owned selector from the admissible frontier, using search-visible evidence. The minimum training-loss candidate was therefore not automatically treated as the final scientific model.

This trace makes the division of authority explicit: the reviewer proposes falsifiable relations and structural edits; deterministic tools fit and evaluate executable expressions; MA-HASE decides whether relations have enough evidence to become active anchors and whether candidates remain admissible; Evidence-preserving Support determines which verified interventions can affect future search; and $\pi_{\mathcal{C}}$ freezes the final expression. All numerical values above are training-side or internal-validation evidence. None uses the sealed holdout.

C Experimental Setup and Reproducibility

This section expands the experimental protocol summarized in the main paper. Its primary purpose is to make explicit which information is visible during search, how budgets are charged, how formulas are frozen, and which diagnostics are computed only after model selection. Unless stated otherwise, the same task split, preprocessing, evaluator, and sealed holdout are used for all methods compared on a given benchmark.

C.1 Benchmarks, Data Sources, and Data Splits

We use the five real-world space-physics benchmarks introduced by PhyE2E (Ying et al., 2025). Table 2 summarizes the variables exposed to symbolic regression and the evaluation regime. Search uses only the train-visible portion declared by the scientific contract $\mathcal{C}$. Sealed holdouts, chronological future intervals, and external simulation diagnostics are excluded from the search state, trajectory labels, review packets, stopping decisions, and final model selection.

Preprocessing and visibility. We preserve the released benchmark coordinate conventions used by the corresponding task interfaces. Any normalization or coordinate transform used by symbolic search is therefore fixed before the run and applied identically across methods. Constants are fitted only on search-visible data. Once a formula is selected, its structure and fitted constants are frozen before any sealed, future, or external diagnostic is evaluated. In particular, the Hotta–Kusano polar comparison, the sunspot 1976–2019 future interval, the Fe X high-temperature holdout, and the lunar-tide sealed holdout do not enter the search state.

C.2 Baseline Methods and Implementation Details

Compared methods. The non-LLM comparison set contains AI Feynman (Udrescu & Tegmark, 2020), BSR (Jin et al., 2019), End-to-End (Kamienny et al., 2022), GP-GOMEA (Virgolin et al., 2021), gplearn (Stephens, 2015), KAN (Liu et al., 2025), NeSymReS (Biggio et al., 2021), Operon (Burlacu et al., 2020), ParFam (Scholl et al., 2025), PhySO (Tenachi et al., 2023), PySR (Cranmer, 2023), TPSR (Shojaee et al., 2023), uDSR (Landajuela et al., 2022), and PhyE2E (Ying et al., 2025). The LLM-based comparison set contains LLM-SR (Shojaee et al., 2025), LaSR (Grayeli et al., 2024), and SR-Scientist (Xia et al., 2026). EvoTrace and all LLM baselines are evaluated with

Table 2: Benchmark variables and sealed evaluation regimes. "Search variables" refer to the coordinates exposed to symbolic regression after the benchmark's released preprocessing. Sealed and external evaluation data are never available to the search procedure.

Task	**Search variables**	**Target**	**Sealed / external evaluation regime**
Solar differential rotation	Heliographic latitude λ (equivalently the transformed angular coordinate used by the formula interface)	Solar angular velocity $\Omega(\lambda)$	Common seven-point sealed holdout. Polar behavior is additionally assessed after freezing using southern and northern polar subsets and the independent Hotta–Kusano simulation (Hotta & Kusano, 2021a;b).
Sunspot forecasting	Time coordinate x, linearly scaled to $[0, 8]$ for symbolic search	Monthly International Sunspot Number, Version 2.0 (WDC–SILSO, Royal Observatory of Belgium, 2026; Clette & Lefèvre, 2016)	218-point random sealed holdout plus a 522-month chronological extrapolation interval from 1976.540 to 2019.958, spanning Solar Cycles 21–24.
Plasma-sheet pressure	Radial coordinate r, azimuth θ, and physical-condition index $c = (Kp, Pd)$	Plasma-sheet pressure	Twelve physical conditions with a common sealed evaluation protocol; outer-region and high-error-tail diagnostics are computed from the frozen predictions.
Fe X 174/175 Å contribution functions	Normalized temperature T and electron-density coordinate n_e	Paired contribution functions G_{174} and G_{175}	Common 650-point sealed high-temperature holdout. Line-wise error, positivity, and the direct G_{175}/G_{174} ratio diagnostic are evaluated only after the two expressions are frozen.
Lunar-tide radial electric field	McIlwain L-shell and transformed local-time coordinates $m = 10\,\mathrm{MLT}/24$ and $\ell = 10\,\mathrm{LLT}/24$	Radial electric field E_r	Common 10,000-point sealed holdout; MLT-period and radial-scaling diagnostics are computed from the frozen expression.

the same three backbone settings: Qwen3.5-35B-A3B (Qwen Team, 2026a), Qwen3.6-35B-A3B (Qwen Team, 2026b), and GLM-5.1 (GLM-5 Team, 2026).

Common evaluation wrapper. The comparison changes the search strategy, not the downstream scoring rule. For a given task, every method receives the same search-visible inputs and the same sealed evaluation split. Numeric constants are fitted without access to the sealed split, and the resulting expression is frozen before final evaluation. We recompute reported predictive metrics with a common evaluator. Structural simplicity is also recomputed uniformly for every method as the node count of the normalized abstract syntax tree (AST); for the paired Fe X task, joint complexity is the sum of the two normalized-AST node counts. Thus, differences in parser-specific tokenization or prefix notation do not enter the complexity comparison.

Method-specific search logic is otherwise preserved. Conventional SR methods retain their own generation, variation, and internal optimization mechanisms; LLM baselines retain their method-specific prompting and search procedures; and EvoTrace alone uses CoTraS, MA-HASE, and evidence-preserving support. Task-interface adaptations are limited to exposing the benchmark variables, enforcing safe executable syntax, applying the common visibility boundary, and charg-

ing the common budget described in Appendix C.4. This prevents EvoTrace-specific trajectory or mechanism information from leaking into a baseline.

C.3 LLM Inference Backends, Hardware, and Runtime Configuration

The two Qwen backbones are served locally with SGLang, whereas GLM-5.1 is accessed through an authenticated API. Local Qwen inference is executed on a four-GPU workstation with $4\times$ NVIDIA RTX 4090 GPUs. Formula-proposal calls use temperature 0.7 for EvoTrace and the LLM baselines. Auxiliary validation or structured-repair calls use the fixed deterministic temperature recorded in the run manifest and cannot directly supply a final formula. EvoTrace's Parent output cap is 8,000 tokens; baseline-specific output caps are retained where required by their documented implementations.

Each run is associated with a manifest that pins the backbone identifier, inference backend, decoding configuration, random seed, task configuration, and evaluator configuration used to generate the reported record. The same manifest also records deterministic repair settings and the constant-fitting configuration. These run-level records are the source of truth for stochastic reproduction; the paper reports only settings shared across runs rather than collapsing method-specific decoding choices into a single artificial configuration.

Structured-output compression. EvoTrace does not replay the complete natural-language interaction history at every iteration. Instead, its compression protocol follows the state objects defined in the method. The Parent Agent returns the protected-baseline identifier, the structural action a_t, the explicit edit Δ_t, and an executable candidate. CoTraS stores paired evaluator evidence and compiles trajectory history into $\mathcal{Q}_t = (\Omega_t^+, \mathcal{G}_t^-, \mathcal{R}_t, \mathcal{B}_t)$; MA-HASE retains the active mechanism anchors and validator outcomes in $\mathcal{M}_t$; and the state distiller constructs the bounded state Z_{t+1} from the active frontier, recent contrastive evidence, failure-family references, unresolved hypotheses, and remaining budget. Full prompts, tool outputs, and rejected candidates remain in the append-only audit ledger but are reintroduced into the active context only through identifier-based retrieval.

Compression therefore removes redundant transcript tokens while preserving the fields required for candidate reconstruction, admissibility checking, trajectory labeling, provenance, and final selection. It changes neither the numerical evaluator nor the candidate/call budget. The real runtime record in Appendix B.6 illustrates how these typed objects are carried across iterations without exposing the sealed holdout.

C.4 Search Budgets, Stopping Rules, and Computational Cost

Category-wise budgets. Budgets are fixed within method categories. Conventional non-LLM SR methods evaluate at most $32{,}500$ candidate expressions per task. Every LLM-based method is capped at 325 backbone calls and at most 325 executable candidate expressions. For agentic methods, the allowance is distributed over at most 13 outer iterations with at most 25 LLM interactions per iteration, so the same run cannot exceed the 325-call ceiling. A run may terminate earlier. All LLM invocations charged to the search procedure, including Parent/reviewer or structured-repair calls when present, consume this call budget; deterministic parsing, constant fitting, numerical evaluation, and validator execution do not consume an LLM call.

These category-specific limits are intentional. A conventional evolutionary candidate and an LLM inference have substantially different unit costs, so a single global expression count would not provide a meaningful access constraint. The protocol instead fixes access within each category and uses the same LLM call/candidate ceiling for EvoTrace and its LLM baselines. It does not claim equality of FLOPs between conventional SR and LLM inference.

Stopping and finalization. Agentic search may stop before budget exhaustion only when a formula in the updated frontier satisfies the contract-defined train-side success criterion and the current validity and mechanism-support requirements. Novelty is checked when admitting a new proposal and is not rechecked for a retained incumbent. Otherwise, search continues until the available budget is exhausted. Stopping does not itself choose the reported formula: after termination, $\pi_{\mathcal{C}}$ selects from the final frontier using only search-visible fitting, internal-validation, and complexity evidence,

as defined in Appendix B.3. The selected structure and constants are then frozen before sealed evaluation.

Cost reporting. Because the two Qwen models are executed locally whereas GLM-5.1 is served by a remote API, raw wall-clock time is not used as a cross-backend fairness criterion. The comparable resource accounting used for the headline experiments is therefore the candidate budget for conventional SR and the LLM-call/candidate budget for LLM-based methods. Structured state compression is an efficiency mechanism inside this common budget, not an additional source of calls: earlier evidence can be reused without retransmitting the complete trajectory, but the 325-call and 325-candidate upper bounds remain unchanged.

C.5 Task-Specific Evaluation and Candidate-Selection Protocols

Separation of search and evaluation. All model selection uses search-visible fitting and internal-validation evidence only. The sealed holdout is opened only after a formula (or Fe X formula pair) has been frozen. Reported task-specific extrapolation and physical diagnostics therefore evaluate a selected hypothesis; they do not provide feedback to CoTraS, MA-HASE, the stopping rule, or $\pi_{\mathcal{C}}$. Train-side structural probes may instantiate the same *type* of relation when that check was declared by the scientific contract, but they use only permitted data and are distinct from the sealed diagnostic values reported below.

Predictive accuracy and complexity. For a frozen expression f, MSE is the arithmetic mean of squared prediction error on the designated evaluation split, and R^2 is reported where meaningful for the benchmark. Complexity is the node count of the normalized AST, with each operator, elementary function, variable, and fitted numeric constant represented as an AST node under the common normalizer. For Fe X, the reported joint complexity is the sum over the two frozen line expressions.

Solar differential rotation. The primary sealed predictive metrics are MSE and R^2 on the common seven-point holdout. For the external polar diagnostic, $R_{\mathrm{S}}^{\mathrm{sm.}}$ and $R_{\mathrm{N}}^{\mathrm{sm.}}$ are Pearson correlations between the smoothed frozen prediction and the Hotta–Kusano simulation over the southern and northern polar subsets, respectively. We additionally report

$$\Delta_{\mathrm{pole}} = \left| \hat{\Omega}(-90^\circ) - \hat{\Omega}(+90^\circ) \right|. \tag{C.1}$$

Non-finite or undefined polar predictions are treated as invalid for the corresponding diagnostic. Neither the polar correlations nor Δ_{pole} is used to choose the frozen formula.

Sunspot forecasting. For Solar Cycle $k \in \{21, 22, 23, 24\}$, R_k denotes the Pearson correlation between the frozen monthly prediction and the observed International Sunspot Number within that cycle. We summarize

$$\text{avg-}R = \frac{1}{4} \sum_{k=21}^{24} R_k, \qquad N_{\mathrm{neg}} = \sum_{t \in \mathcal{H}_{\mathrm{future}}} \mathbb{I}[\hat{y}_t < 0], \tag{C.2}$$

where $\mathcal{H}_{\mathrm{future}}$ is the 522-month chronological extrapolation interval. The future interval is sealed throughout search; cycle-wise correlation and the negative-month count are post-selection diagnostics.

Plasma-sheet pressure. Outer MSE is the MSE restricted to the benchmark-defined radial outer subset of the sealed evaluation set. P99 AE is the empirical 99th percentile of $|\hat{y} - y|$ pooled over the designated sealed evaluation records. These diagnostics distinguish average predictive quality from degradation in the radially shifted region and the high-error tail. The same frozen formula is used for the overall MSE, Outer MSE, and P99 AE calculation.

Fe X 174/175 Å contribution functions. For each line $\lambda \in \{174, 175\}$, we report the line-wise MSE and the positivity rate

$$\mathrm{Pos}_\lambda = \frac{100}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{G}_{\lambda,i} > 0]. \tag{C.3}$$

The coupled diagnostic is computed in direct ratio space on the same 650-point high-temperature holdout:

$$\mathrm{RawRatioMSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{G}_{175,i}}{\hat{G}_{174,i}} - \frac{G_{175,i}}{G_{174,i}}\right)^2. \tag{C.4}$$

No log transform is applied to this ratio metric. The formula pair is frozen before any of these sealed values are computed. During search, the Fe X selector operates only on the permitted train-side evidence and the contract-defined pairwise admissibility certificate; the sealed RatioMSE is not a selection objective.

Lunar-tide radial electric field. Let $\hat{T}_{\mathrm{MLT}}$ denote the identifiable dominant MLT period of the frozen expression and let $\hat{p}_L$ denote its effective radial power-law exponent under the diagnostic procedure. We report

$$\Delta T_{\mathrm{MLT}} = |\hat{T}_{\mathrm{MLT}} - 12\ \mathrm{h}|, \qquad \Delta p_L = |\hat{p}_L + 2|. \tag{C.5}$$

If no identifiable MLT-periodic component is present, ΔT_{MLT} is reported as unavailable rather than imputed. These quantities are post-selection diagnostics and do not enter trajectory labels. The real search trace in Appendix B.6 illustrates the distinct selection stage: the lunar-tide run freezes its reported model using repeated k-fold internal validation and the one-standard-error rule rather than choosing the expression with the smallest training MSE.

Finally, the cross-task score $\mathcal{Q}$ and Gain shown in the main paper are summary statistics only; they are never used for search or model selection. Their normalization, extreme-value handling, and weight-sensitivity analysis are reported separately in Appendix D.3.

D Additional Experimental Results and Analyses

D.1 Complete Quantitative Results

The complete quantitative results are presented in the same task order as the task-specific analysis in the main paper: solar differential rotation, sunspot forecasting, plasma-sheet pressure, Fe X contribution functions, and lunar-tide radial electric field. The complete task-level results are reported in Tables 3–7. All complexity values are measured consistently as the node count of the normalized abstract syntax tree (AST); for Fe X, the reported joint AST complexity is the sum of the AST node counts of the two frozen line expressions. Exact task-specific metric definitions and candidate-selection protocols are given in Appendix C.5.

Solar differential rotation. Table 3 reports predictive accuracy on the seven-point sealed holdout, AST complexity, and post-selection polar diagnostics across method groups and LLM backbones. Relative to the lowest-MSE LLM baseline under each backbone, EvoTrace reduces MSE by **94.1%**, **97.8%**, and **65.9%** under Qwen3.5, Qwen3.6, and GLM-5.1, respectively. Under Qwen3.6, its 9.39×10^{-6} MSE is also **63.7%** lower than End-to-End. Its smoothed polar correlations range from 0.9840 to 0.9861 in the south and from 0.9875 to 0.9900 in the north. The predicted north–south polar differences are 0.0051, 0.0459, and 0.0278, respectively. These differences quantify the north–south asymmetry of the frozen predictions; they are not errors relative to a reference asymmetry magnitude.

Sunspot forecasting. Table 4 reports the complete predictive, cycle-wise, and extrapolation diagnostics across all method groups and LLM backbones. Under Qwen3.5, EvoTrace raises avg-R from LaSR's 0.2008 to 0.7915, a **3.94**$\times$ increase, while eliminating all 94 negative-prediction months. Across all three backbones, EvoTrace maintains avg-$R = 0.7915$ and produces no negative predictions over the full 522-month extrapolation interval.

Plasma-sheet pressure. Table 5 reports the full predictive, outer-region, tail-error, and AST-complexity results across the 12 physical conditions. Under Qwen3.5, EvoTrace reduces overall MSE, Outer MSE, and P99 AE by **25.7%**, **53.1%**, and **24.1%**, respectively, relative to the strongest LLM baseline for each metric. Compared with PhyE2E, it further lowers overall MSE by **18.9%** and P99 AE by **49.3%**. Under Qwen3.6, the corresponding MSE and P99-AE reductions over the strongest LLM baselines are **9.3%** and **27.8%**.

Table 3: Solar differential rotation: predictive accuracy on the common seven-point sealed holdout, AST complexity, and post-selection polar diagnostics. $R_{\mathrm{S}}^{\mathrm{sm.}}$ and $R_{\mathrm{N}}^{\mathrm{sm.}}$ are smoothed Pearson correlations with the Hotta–Kusano simulation over the southern and northern polar subsets, respectively. Δ_{pole} is the absolute difference between the predicted southern and northern polar values; it is reported descriptively and is not ranked. For the other metrics, bold and underlined entries indicate the best and second-best distinct values within each method block, including ties.

Method	MSE ↓	R^2 ↑	AST Complexity ↓	$R_{\mathrm{S}}^{\mathrm{sm.}}$ ↑	$R_{\mathrm{N}}^{\mathrm{sm.}}$ ↑	Δ_{pole}
Without LLMs						
AIFeynman	$\underline{8.38\times10^{-5}}$	<u>0.9980</u>	10	0.9651	0.9713	0.0000
BSR	3.83×10^{1}	-901.4639	24	0.9333	0.8563	0.2776
E2E	$\mathbf{2.59\times10^{-5}}$	**0.9994**	16	**0.9862**	**0.9900**	0.0096
GP-GOMEA	8.41×10^{-1}	-18.8259	10	0.9626	-0.9835	13.5097
gplearn	2.01×10^{-3}	0.9526	**4**	<u>0.9824</u>	0.9868	0.0000
KAN	3.37×10^{-1}	-6.9441	8	-0.9862	-0.9900	0.0256
NeSymReS	9.26×10^{-5}	0.9978	<u>7</u>	0.9818	0.9863	0.0056
ParFam	1.97×10^{32}	-4.64×10^{33}	93	0.2618	-0.9791	7137.1142
PhySO	3.55×10^{-4}	0.9916	12	-0.3478	<u>0.9875</u>	878.6902
PySR	7.08×10^{1}	-1669.1699	22	0.8474	0.9843	1.3033
TPSR	1.30×10^{1}	-305.9169	18	0.9751	-0.9982	0.1788
PhyE2E	9.47×10^{-5}	0.9978	<u>7</u>	<u>0.9824</u>	0.9868	0.0000
Qwen3.5-35B-A3B						
EvoTrace	$\mathbf{2.26\times10^{-5}}$	**0.9995**	**17**	0.9861	0.9900	0.0051
LLM-SR	6.05×10^{-3}	0.8573	156	<u>0.9946</u>	0.9738	5.4962
LaSR	8.94×10^{-4}	0.9789	<u>28</u>	0.9328	<u>0.9918</u>	$+\infty$
SR-Scientist	<u>3.81×10^{-4}</u>	<u>0.9910</u>	254	**0.9949**	**0.9997**	5.0571
Qwen3.6-35B-A3B						
EvoTrace	$\mathbf{9.39\times10^{-6}}$	**0.9998**	<u>40</u>	**0.9850**	**0.9875**	0.0459
LLM-SR	5.80×10^{-4}	0.9863	216	<u>0.9845</u>	0.9763	2.2230
LaSR	<u>4.22×10^{-4}</u>	<u>0.9900</u>	**30**	0.8387	<u>0.9814</u>	0.1111
SR-Scientist	8.55×10^{-4}	0.9798	78	0.9457	0.9599	0.7117
GLM-5.1						
EvoTrace	$\mathbf{2.19\times10^{-5}}$	**0.9995**	**24**	**0.9840**	**0.9900**	0.0278
LLM-SR	1.18×10^{-4}	0.9972	115	0.1475	0.9669	5.85×10^{15}
LaSR	<u>6.42×10^{-5}</u>	<u>0.9985</u>	<u>28</u>	<u>0.9689</u>	<u>0.9812</u>	0.0328
SR-Scientist	3.06×10^{-3}	0.9279	79	0.9374	0.9394	0.2653

Fe X 174/175 Å contribution functions. Table 6 reports the complete line-wise and coupled diagnostic results on the sealed high-temperature holdout. Under Qwen3.5, LaSR attains a lower mean line-wise MSE, but its Raw RatioMSE is 1.2068×10^{-1}, **9.8**× larger than EvoTrace's 1.2333×10^{-2}. EvoTrace also reduces RatioMSE by **27.0%** relative to SR-Scientist in the same block. The advantage persists under the other backbones, with reductions of **94.9%** (**19.6**×) under Qwen3.6 and **80.2%** (**5.0**×) under GLM-5.1 relative to the strongest competing LLM result.

Lunar-tide radial electric field. Table 7 reports the complete predictive and physical-diagnostic results across all three LLM backbones. EvoTrace reduces MSE over the strongest LLM baseline by **24.6%**, **2.1%**, and **1.2%** under Qwen3.5, Qwen3.6, and GLM-5.1, respectively. Its MLT-period deviations are only 0.030, 0.040, and 0.000 hours; under Qwen3.5, 0.030 hours is a **13.3**× reduction from the next-best identifiable period deviation of 0.400 hours.

D.2 Representative Discovered Expressions

Tables 8–12 report representative frozen expressions recovered by EvoTrace and the comparison methods. The tasks are presented in the same order as the task-specific extrapolation and physical-consistency analysis in the main paper: solar differential rotation, sunspot forecasting, plasma-sheet pressure, Fe X contribution functions, and lunar-tide radial electric field.

Solar differential rotation. Table 8 lists the frozen expressions used for the full-latitude and polar-extrapolation evaluation.

Table 4: Full results for sunspot forecasting on the 218-point sealed holdout and the 522-month chronological extrapolation interval spanning Solar Cycles 21–24. We report predictive accuracy, AST complexity, cycle-wise correlations, their mean avg-R, and the number of negative extrapolation months. Bold and underlined values denote the best and second-best non-tied results, respectively, within each model block.

Method	MSE↓	R^2 ↑	AST Complexity↓	avg-R ↑	R_{21} ↑	R_{22} ↑	R_{23} ↑	R_{24} ↑	Neg. months↓
Without LLMs									
AIFeynman	1.1634×10^{4}	-1.4612	11	0.2139	0.2099	0.2336	0.2741	0.1379	522
BSR	1.1717×10^{4}	-1.4786	**7**	0.2828	0.2775	0.4020	0.3619	0.0898	207
End-to-End	2.7999×10^{3}	0.4077	16	**0.8652**	0.9122	0.8544	**0.8395**	0.8547	0
GP-GOMEA	6.0449×10^{3}	-0.2788	11	0.2219	0.1793	0.2365	0.2984	0.1734	0
GPLearn	4.6941×10^{3}	0.070	55	-0.0687	-0.1372	-0.4445	0.2683	0.0386	0
KAN	4.2365×10^{3}	0.1038	10	-0.1313	0.2684	0.2283	-0.7245	-0.2972	0
NeSymReS	9.7852×10^{3}	-1.0700	10	0.4170	0.1912	0.6422	0.7536	0.0810	0
Operon	2.4858×10^{3}	0.4742	91	-0.3631	0.4196	-0.4201	-0.8191	-0.6329	26
ParFam	1.3013×10^{12}	-2.7529×10^{8}	109	-0.2896	-0.2833	-0.3011	-0.3567	-0.2173	0
PhySO	3.3317×10^{3}	0.2952	14	0.3838	0.4145	**0.7032**	0.6601	-0.2425	0
PySR	3.5422×10^{3}	0.2507	10	0.1657	0.4205	-0.1398	-0.1206	0.5029	0
TPSR	8.6525×10^{3}	-0.8304	28	-0.1995	-0.6967	0.2905	0.2053	-0.5970	227
uDSR	2.7713×10^{6}	-585.2436	52	0.3157	0.3569	0.3675	0.3427	0.1957	522
PhyE2E	$\mathbf{1.6559\times10^{3}}$	**0.6497**	34	0.8099	**0.9330**	0.6888	0.7012	**0.9164**	0
Qwen3.5-35B-A3B									
EvoTrace	1.4222×10^{3}	0.6992	28	**0.7915**	0.8353	**0.9311**	0.6240	**0.7756**	0
LLM-SR	1.9401×10^{3}	0.5896	66	0.7856	0.8931	0.8029	**0.7710**	0.6755	377
LaSR	$\mathbf{1.2673\times10^{3}}$	**0.7319**	**22**	0.2008	**0.9162**	0.0710	-0.3683	0.1842	94
SR-Scientist	2.4125×10^{3}	0.4897	27	0.3178	0.8709	0.2857	-0.0109	0.1256	0
Qwen3.6-35B-A3B									
EvoTrace	$\mathbf{1.4222\times10^{3}}$	**0.6992**	28	**0.7915**	0.8353	**0.9311**	0.6240	**0.7756**	0
LLM-SR	1.8548×10^{3}	0.6076	153	0.3355	0.3444	0.3457	0.3992	0.2528	445
LaSR	1.8980×10^{3}	0.5985	**13**	0.3270	0.5847	0.7311	0.5351	-0.5429	0
SR-Scientist	2.8296×10^{3}	0.4014	24	0.7134	**0.8717**	0.7202	**0.6396**	0.6222	430
GLM-5.1									
EvoTrace	1.4222×10^{3}	0.6992	28	0.7915	0.8353	**0.9311**	0.6240	0.7756	0
LLM-SR	2.0071×10^{3}	0.5754	64	**0.8179**	**0.9298**	0.7048	**0.7318**	0.9053	0
LaSR	$\mathbf{1.0412\times10^{3}}$	**0.7797**	**22**	0.1445	0.9139	-0.0835	-0.4105	0.1581	0
SR-Scientist	2.2046×10^{3}	0.5336	35	0.8073	0.9270	0.6840	0.7044	**0.9137**	0

Sunspot forecasting. Table 9 lists the frozen expressions used for the sealed-holdout and long-horizon forecasting evaluation.

Plasma-sheet pressure. Table 10 summarizes representative frozen expression families across the physical-condition setting.

Fe X 174/175 Å contribution functions. Table 11 reports the frozen expressions underlying the line-wise and coupled-ratio diagnostics.

Lunar-tide radial electric field. Table 12 lists the frozen expressions used for the predictive and physical-diagnostic analysis.

Physical interpretation of the lunar-tide expression. The EvoTrace expression contains an explicit L^{-2} prefactor consistent with the ideal-corotation reference, together with additional L-dependent terms that allow departures from this radial scaling. Its sinusoidal terms encode MLT-periodic modulation, while the ℓ-dependent term introduces an LLT contribution whose amplitude varies with L. Together, these components provide a compact and physically interpretable hypothesis for the radial, local-time, and lunar-phase dependence of the recovered lunar-tide contribution.

D.3 Composite-Metric Normalization and Weight Sensitivity

The composite score in Fig. 3(f) is a compact summary of heterogeneous task diagnostics; it is not used for search, stopping, or task-specific model selection. Raw task metrics remain the primary

Table 5: Full results for plasma-sheet pressure prediction across 12 physical conditions. We report sealed-holdout MSE and R^2, AST complexity, Outer MSE, and P99 absolute error (P99 AE). Bold and underlined values denote the best and second-best results, respectively, within each model block.

Method	MSE↓	R^2 ↑	AST Complexity↓	Outer MSE↓	P99 AE↓
Without LLMs					
BSR	<u>1.79</u>	<u>0.330</u>	**9**	0.138	<u>5.35</u>
End-to-End	2.54	0.072	30	<u>3.21×10^{-3}</u>	8.14
GP-GOMEA	2.96	−0.127	16	0.0174	8.53
KAN	2.88	−0.093	<u>14</u>	0.311	8.50
Operon	2.97	−0.130	100	0.0256	8.46
ParFam	6.88	−2.32	54	10.1	7.52
PhySO	5.02×10^{2}	−267	17	1.46×10^{3}	47.5
PySR	2.79×10^{7}	-1.49×10^{7}	20	173	1.86×10^{4}
TPSR	30.6	−15.0	28	3.16	22.5
PhyE2E	**0.557**	**0.821**	19	**5.69×10^{-4}**	**3.41**
Qwen3.5-35B-A3B					
EvoTrace	**0.452**	**0.880**	<u>34</u>	**0.006590**	**1.73**
SR-Scientist	0.635	<u>0.822</u>	48	<u>0.014058</u>	2.90
LLM-SR	1.15	0.408	97	5.32	4.12
LaSR	<u>0.608</u>	0.819	**24**	0.0190	<u>2.28</u>
Qwen3.6-35B-A3B					
EvoTrace	**0.607**	**0.820**	<u>39</u>	**0.021264**	**1.69**
SR-Scientist	<u>0.669</u>	<u>0.818</u>	45	<u>0.024065</u>	<u>2.34</u>
LLM-SR	0.826	0.652	116	0.547	2.94
LaSR	1.89	0.484	**25**	14.4	2.53
GLM-5.1					
EvoTrace	**0.476**	**0.866**	<u>58</u>	**0.0527**	**1.77**
SR-Scientist	<u>1.60×10^{6}</u>	<u>-1.05×10^{6}</u>	**29**	<u>5.59×10^{7}</u>	<u>5.82×10^{3}</u>

evidence throughout the paper. The construction below is deliberately ordinal so that metrics with different physical units and extreme numerical ranges can be combined without allowing one outlier to dominate the summary.

Rank-percentile normalization. Let $g = (t, b)$ denote a task–backbone comparison group, where t indexes the task and b the LLM backbone. For atomic metric k, let $\mathcal{M}_{gk}^{\mathrm{fin}}$ be the methods with a finite, applicable value and let $M_{gk} = |\mathcal{M}_{gk}^{\mathrm{fin}}|$. After orienting metric k so that rank one is best, a finite result of method m receives average rank r_{gmk} under ties and utility

$$u_{gmk} = \begin{cases} 1, & M_{gk} = 1, \\ 1 - \dfrac{r_{gmk} - 1}{M_{gk} - 1}, & M_{gk} > 1. \end{cases} \tag{D.1}$$

Thus, among finite applicable results, the best and worst values map to one and zero, respectively. Smaller-is-better metrics are ranked in ascending order and larger-is-better metrics in descending order. Because the mapping depends only on order, it is invariant to monotone rescaling and prevents catastrophic values spanning many orders of magnitude from compressing the remaining methods into a narrow numerical interval.

Table 6: Full results for Fe X 174/175 Å contribution-function extrapolation on the 650-point sealed high-temperature holdout. We report line-wise and mean MSE, Raw RatioMSE, joint AST complexity, and positivity rates for both emission lines. Bold and underlined values denote the best and second-best results, respectively, within each model block.

Method	$\mathrm{MSE}_{175}\downarrow$	$\mathrm{MSE}_{174}\downarrow$	Mean MSE↓	Raw RatioMSE↓	Joint AST Complexity↓	$\mathrm{Positivity}_{175}\uparrow$	$\mathrm{Positivity}_{174}\uparrow$
Without LLMs							
PhyE2E	$\underline{2.6187\times10^{-7}}$	$\underline{2.5152\times10^{-7}}$	$\underline{2.5670\times10^{-7}}$	$\mathbf{3.9705\times10^{-1}}$	**24**	**100.0%**	**100.0%**
End-to-End	1.2041×10^{-5}	1.1963×10^{-5}	1.2002×10^{-5}	$\underline{1.1500\times10^{0}}$	74	**100.0%**	**100.0%**
GP-GOMEA	7.7037×10^{-3}	7.7044×10^{-3}	7.7041×10^{-3}	2.0000×10^{0}	40	0.0%	0.0%
PhySO	$\mathbf{2.1803\times10^{-10}}$	$\mathbf{1.5937\times10^{-9}}$	$\mathbf{9.0587\times10^{-10}}$	2.9000×10^{2}	$\underline{36}$	**100.0%**	**100.0%**
PySR	1.5073×10^{0}	1.5073×10^{0}	1.5073×10^{0}	6.0000×10^{2}	48	**100.0%**	**100.0%**
TPSR	4.4493×10^{-4}	4.4445×10^{-4}	4.4469×10^{-4}	3.1600×10^{0}	70	**100.0%**	**100.0%**
Qwen3.5-35B-A3B							
EvoTrace	$\underline{2.5191\times10^{-9}}$	$\underline{2.1584\times10^{-8}}$	$\underline{1.2052\times10^{-8}}$	$\mathbf{1.2333\times10^{-2}}$	$\underline{40}$	**100.0%**	**100.0%**
LLM-SR	1.3415×10^{4}	2.3562×10^{4}	1.8489×10^{4}	8.8333×10^{-2}	318	0.0%	0.0%
LaSR	$\mathbf{2.2250\times10^{-12}}$	$\mathbf{2.3572\times10^{-9}}$	$\mathbf{1.1797\times10^{-9}}$	1.2068×10^{-1}	71	**100.0%**	**100.0%**
SR-Scientist	3.7182×10^{-3}	2.5036×10^{-2}	1.4377×10^{-2}	$\underline{1.6892\times10^{-2}}$	**35**	**100.0%**	**100.0%**
Qwen3.6-35B-A3B							
EvoTrace	$\underline{2.4467\times10^{-10}}$	$\underline{1.7174\times10^{-9}}$	$\underline{9.8102\times10^{-10}}$	$\mathbf{4.7223\times10^{-2}}$	$\underline{88}$	**100.0%**	**100.0%**
LLM-SR	7.2299×10^{-2}	4.5615×10^{-2}	5.8957×10^{-2}	$\underline{9.2366\times10^{-1}}$	303	98.2%	**100.0%**
LaSR	2.6577×10^{-5}	1.1203×10^{-5}	1.8890×10^{-5}	2.4634×10^{0}	**65**	23.1%	0.0%
SR-Scientist	$\mathbf{2.3869\times10^{-11}}$	$\mathbf{1.2485\times10^{-10}}$	$\mathbf{7.4359\times10^{-11}}$	2.1360×10^{2}	189	**100.0%**	**100.0%**
GLM-5.1							
EvoTrace	$\mathbf{6.8344\times10^{-11}}$	$\mathbf{3.3869\times10^{-10}}$	$\mathbf{2.0352\times10^{-10}}$	$\mathbf{2.9042\times10^{-2}}$	108	**100.0%**	**100.0%**
LLM-SR	2.1020×10^{3}	$\underline{1.3312\times10^{-5}}$	1.0510×10^{3}	1.5700×10^{8}	531	$\underline{0.2\%}$	0.0%
LaSR	$\underline{2.4467\times10^{-10}}$	2.4690×10^{0}	1.2345×10^{0}	$\underline{1.4650\times10^{-1}}$	$\underline{67}$	0.0%	**100.0%**
SR-Scientist	4.5170×10^{-1}	7.5017×10^{-4}	$\underline{2.2623\times10^{-1}}$	4.0575×10^{283}	**63**	**100.0%**	**100.0%**

A method-specific numerical failure, non-finite value, or missing value caused by an invalid/unavailable prediction receives utility zero for that atomic metric. By contrast, if a diagnostic is not defined for the task itself, the metric is removed for every method before aggregation. No finite raw value is clipped or replaced by an arbitrary large constant. A reported dash for an unidentifiable physical component is therefore treated as a zero-utility method-specific failure only when that component is required by the task's declared diagnostic.

Task-level dimension scores and Gain. Within each task–backbone group, atomic utilities are averaged into four predefined dimensions: predictive accuracy ($\mathcal{A}$), extrapolation robustness ($\mathcal{E}$), physical consistency ($\mathcal{P}$), and structural simplicity ($\mathcal{S}$). Atomic metrics have equal weight within a dimension. For example, Fe X accuracy uses the two line-wise MSE utilities, its extrapolation term uses Raw RatioMSE, and its physical term uses the two positivity rates; the lunar-tide physical term uses ΔT_{MLT} and Δp_L. Structural simplicity is computed from normalized-AST complexity as defined in Appendix C.5.

Let $\mathcal{D}_t \subseteq \{A, E, P, S\}$ be the dimensions defined for task t. If a dimension is not applicable to the task, it is omitted for all methods and the remaining default weights are renormalized,

$$\widetilde{w}_{t,d} = \frac{w_d}{\sum_{d' \in \mathcal{D}_t} w_{d'}}, \qquad d \in \mathcal{D}_t. \tag{D.2}$$

The task-level composite score is then

$$\mathcal{Q}_{gbm} = \prod_{d \in \mathcal{D}_t} \left(\mathcal{U}^{(d)}_{gbm} + \epsilon \right)^{\widetilde{w}_{t,d}}, \qquad \epsilon = 10^{-6}, \tag{D.3}$$

where $g = (t, b)$ and $\mathcal{U}^{(d)}_{gbm} \in [0, 1]$ is the corresponding dimension score. Equivalently, we write $\mathcal{Q}_{tbm}$ when the task/backbone indices are shown explicitly. The default weights are

$$\mathbf{w}_0 = (w_A, w_E, w_P, w_S) = (0.40, 0.30, 0.20, 0.10).$$

They are fixed uniformly across methods and backbones and are not adjusted to individual tasks except for the deterministic renormalization in Eq. (D.2). They encode the declared priority ordering used in the main paper: predictive accuracy is primary, extrapolation robustness is next, and physical consistency and simplicity provide complementary evidence.

Table 7: Full results for lunar-tide radial electric-field prediction on the common 10,000-point sealed holdout. We report MSE, R^2, AST complexity, MLT-period deviation ΔT_{MLT}, and radial-scaling deviation Δp_L. Bold and underlined values denote the best and second-best results, respectively, within each model block.

Method	MSE↓	R^2 ↑	AST Complexity↓	ΔT_{MLT} (h)↓	Δp_L ↓
Without LLMs					
PhyE2E	**0.2331**	**0.8311**	43	**0.110**	<u>0.065</u>
BSR	3.564	-1.583	<u>10</u>	17.3	2.00
End-to-End	1.382	-0.0017	39	36.0	1.88
GP-GOMEA	1.413	-0.0246	**9**	36.0	1.26
KAN	1.492	-0.0817	31	18.4	1.83
NeSymReS	4.693×10^{9}	-3.402×10^{9}	15	8.87	1.49
Operon	20.80	-14.08	94	33.6	2.65
ParFam	<u>0.9218</u>	<u>0.3318</u>	50	36.0	0.204
PhySO	1.616	-0.1713	13	–	1.67
PySR	2.097	-0.5199	20	<u>6.89</u>	**0.000**
TPSR	26.47	-18.19	28	36.0	2.05
uDSR	1.268×10^{5}	-9.190×10^{4}	132	36.0	2.01
Qwen3.5-35B-A3B					
EvoTrace	**0.2406**	**0.8256**	<u>25</u>	**0.030**	<u>0.004</u>
LLM-SR	0.4912	0.6439	64	<u>0.400</u>	1.307
LaSR	<u>0.3189</u>	<u>0.7688</u>	**13**	3.65	**0.000**
SR-Scientist	0.5342	0.6127	34	–	0.654
Qwen3.6-35B-A3B					
EvoTrace	**0.2328**	**0.8312**	48	**0.040**	<u>0.071</u>
LLM-SR	0.5382	0.6099	78	–	**0.032**
LaSR	0.3090	0.7760	**9**	0.100	2.359
SR-Scientist	<u>0.2378</u>	<u>0.8276</u>	<u>21</u>	<u>0.070</u>	0.145
GLM-5.1					
EvoTrace	**0.2355**	**0.8293**	34	**0.000**	<u>0.333</u>
LLM-SR	0.7343	0.4677	61	–	0.568
LaSR	0.3090	0.7760	**10**	0.100	2.359
SR-Scientist	<u>0.2384</u>	<u>0.8272</u>	<u>15</u>	<u>0.050</u>	**0.151**

Note. "–" denotes that no identifiable MLT-periodic component is present. For Qwen3.5 SR-Scientist, Δp_L is averaged over its three frozen repetitions, while ΔT_{MLT} is unavailable because one repetition contains no MLT dependence.

For each task–backbone cell in Fig. 3(f), the reported value is

$$\mathrm{Gain}_{tb} = \frac{\mathcal{Q}_{tb,\mathrm{EvoTrace}}}{\max_{m\in\mathcal{B}_{tb}} \mathcal{Q}_{tbm}}, \tag{D.4}$$

where $\mathcal{B}_{tb}$ is the set of available LLM baselines under the same task and backbone. This definition is the one underlying the 15 task–backbone Gain cells discussed in the main text.

The small ϵ in Eq. (D.3) is only a numerical floor after rank normalization; it does not modify any raw metric. Consequently, a Gain can become numerically large when the strongest baseline has zero rank-percentile utility on one or more weighted dimensions. Such a Gain should therefore be interpreted as a relative composite-score ratio, not as a multiplicative reduction in prediction error.

Table 8: Representative frozen expressions for solar differential rotation. Let λ denote heliographic latitude in degrees and $x = \lambda\pi/180 + \pi/2$. Coordinate notation is method-specific: $\theta = x$ in the EvoTrace row, whereas $\theta = x - \pi/2$ in the E2E and GPLearn rows. The expressions correspond to frozen models used in predictive and polar-extrapolation evaluation.

Method	Frozen expression
EvoTrace	$\Omega(\theta) = 0.031126381670\theta^4 - 0.185118023330\theta^3 - 0.129536209990\theta^2$ $+ 1.286715916\theta + 1.737669583$ $+ 0.0101828097\sin(3.497750600 9\theta) + 0.0047172294\sin(6.995501201 8\theta)$
PhyE2E	$\Omega(x) = 1.80703 + 1.10364\sin x$
E2E	$\Omega(\theta) = 2.901 - 0.488\left(-\theta + 0.00533 + \dfrac{0.786}{-56.308\theta - 375.424}\right)^2$
GPLearn	$\Omega(\theta) = \cos\theta + 1.934$
SR-Scientist	$\Omega(x) = \sum_{j=1}^{9} a_j \sin(jkx)$
LaSR	$r(x) = -\dfrac{1.0521629091}{\sin(x^{1.2294879693})} - \dfrac{0.0310426831}{x - 0.7785261516}$, $q(x) = \dfrac{0.8967028834}{\cos\left[0.0655377429 - \sin(\sin r(x))\right]}$, $\Omega(x) = \exp\left[\left(\sqrt{\sqrt{1.2092842111\sin x}}\right)^{q(x)}\right]$
LLM-SR	$\Omega(x) = a + bx + c\,\mathrm{asinh}(dx) + e\ \mathrm{sgn}(x)\lvert x\rvert^f$ $+ ge^{-h(x-i)^2} + j\tanh[k(x-\ell)]$

Table 9: Representative frozen expressions for monthly sunspot-number forecasting. The time coordinate x is linearly scaled to $[0, 8]$. The displayed expressions are the frozen models used for the sealed-holdout and chronological-extrapolation evaluation.

Method	Frozen expression
EvoTrace	$S(x) = 97.231234 - 38.909231\sin(9.651343x + 1.809742)$ $- 63.095916\sin(8.775094x + 7.304316)$ $- 17.810229x + 2.712819x^2$
PhyE2E	$S(x) = \dfrac{N(x)}{D(x)}$, $N(x) = 106.04710\sin^2(4.36189x - 0.09037)$, $D(x) = 0.63754\sin^2(0.48342x + 3.15311)$ $+ 0.11800\cos^2(1.20028x - 0.18912) + 0.42032$
BSR	$S(x) = -\sin(0.0805x - 148.657)$
End-to-End	$S(x) = 1375.295\left[1 - 0.0232\sin(0.578x - 1070.873)\right]^2$ $- 1294.074$
SR-Scientist	$S(x) = 415.67387x - 123.35406x^2 + 8.47674x^3$ $- 340.79033\sin(0.64718x) - 63.21312\cos(8.66897x)$
LaSR	$S(x) = 80.1512 + 28.1574\cos x$ $- 74.7569\cos[8.71918x + \sin(x + 0.66313) - 0.60392]$
LLM-SR	$S(x) = 458.0199 + 682.7331x - 15.4841x^2 - 41.2222e^{-2130.2280\lvert x\rvert}$ $+ 64.4171\sin(8.7190x - 1.8551) - \dfrac{4619.0431}{1 + 0.38078x^2}$ $+ 2258.5164e^{-0.14339x^2} - 682.6527\lvert x\rvert^{0.92548}$ $+ 2005.4816e^{-0.64959x^2}$

Weight-sensitivity protocol. The default weights in the headline figure are kept fixed independently of this sensitivity audit. We additionally evaluate an equal-weight setting,

Table 10: Representative frozen expression families for plasma-sheet pressure prediction across 12 physical conditions. Here $r = x_0$, $\theta = x_1$, and $c = (Kp, Pd)$ indexes the physical condition. For methods using condition-specific expressions, the displayed formulas are representative frozen members of the corresponding evaluated family.

Method	Frozen expression or expression family
EvoTrace	$P_c(r,\theta) = a_{7,c}\exp(B_c(r)M_c(r,\theta)) + a_{8,c},$ $B_c(r) = a_{0,c}e^{-a_{1,c}r},$ $M_c(r,\theta) = a_{2,c} + a_{3,c}\theta + a_{4,c}\theta^2 + a_{5,c}r\theta + a_{6,c}r^2$
PhyE2E	$P_c(r,\theta) = A_c + \dfrac{B_c}{(r-d_c)^2[\cos(\omega_c\theta)+b_c]}$
BSR	$P_c(r,\theta) = \dfrac{C_{1,c}}{r(\theta+C_{2,c})+r}$
End-to-End	$P_c(r,\theta) = C_{1,c}\left(C_{2,c} - \dfrac{1}{U_c(\theta)V_c(r,\theta)}\right)^2 + C_{10,c},$ $U_c(\theta) = C_{3,c}\theta + C_{4,c},$ $V_c(r,\theta) = C_{5,c}r + C_{6,c}\sin(C_{7,c}\theta + C_{8,c}) + C_{9,c}$
SR-Scientist	$P_{3,1.5}(r,\theta) = \dfrac{a_0}{r^2} + \dfrac{a_1\cos\theta + a_2\sin\theta}{r^3} + \dfrac{a_3\cos 2\theta + a_4\sin 2\theta}{r^4} + \dfrac{a_5\cos 3\theta + a_6\sin 3\theta}{r^5} + \dfrac{a_7\cos 4\theta + a_8\sin 4\theta}{r^6} + \dfrac{a_9\cos 5\theta}{r^7}$
LaSR	$P_{2,3}(r,\theta) = \dfrac{N(r,\theta)}{D(r,\theta)},$ $N(r,\theta) = \dfrac{\cos(1.579532^{\,r^{-2.883567\theta}})}{r} + \sqrt{\exp(2.297967/r)},$ $D(r,\theta) = [\sin\theta + 2.390056]r - 0.932885$
LLM-SR	$P_{3,1.5}(r,\theta) = 4.83704 + 1.76014\bar{r} + 2.56495\bar{\theta} - 0.09144\bar{r}^2 - 0.18583\bar{\theta}^2 - 0.51250\bar{r}\bar{\theta} + 123.00810e^{-5.74393(\bar{r}^2+0.01003\bar{\theta}^2)} + 89.09458e^{1.92065\rho}\sin(0.03337\bar{r} + 0.00808\bar{\theta} + 6.24656) + 4.80883\dfrac{\bar{r}+\bar{\theta}}{\lvert\bar{r}+\bar{\theta}\rvert + 10^{-12}} - 11.10668\arctan(\bar{r}+\bar{\theta}),$ $\bar{r} = \mathrm{clip}(r,-100,100), \quad \bar{\theta} = \mathrm{clip}(\theta,-100,100),$ $\rho = \sqrt{\bar{r}^2+\bar{\theta}^2}$

$(0.25, 0.25, 0.25, 0.25)$, and eight symmetric one-at-a-time perturbations. For the latter, one default weight is changed by $+0.10$ or -0.10, and the remaining three are rescaled proportionally so that all weights remain non-negative and sum to one. The same task-level applicability and renormalization rules are applied under every setting.

To summarize sensitivity without replacing the 15 task-level Gain values, we form a backbone-level audit score by first averaging each dimension over the tasks for which that dimension is defined and then applying the candidate weight vector. This auxiliary backbone-level aggregation is used only for the sensitivity table below; it is not used in Fig. 3 or in any model-selection decision. For each weight setting we report: (i) the number of backbones on which EvoTrace is top ranked under this audit score, (ii) the median aggregate Gain over the strongest baseline across the three backbones, and (iii) the mean Kendall τ between the resulting method ordering and the default-weight ordering, computed separately within each backbone and then averaged. Ranges summarize the eight one-at-a-time perturbations. Across the ten evaluated settings, EvoTrace remains top ranked under the auxiliary cross-task score for all three backbones. Reweighting changes the margin more than the ordering: the median aggregate Gain remains above $1.52\times$, while the mean Kendall correlation with the default ordering never falls below 0.889. Together with the task-level Gain cells in the main paper, this indicates that the qualitative cross-backbone conclusion is not tied to the particular $0.40/0.30/0.20/0.10$ weighting choice.

Table 11: Representative frozen expression pairs for Fe X 174/175 Å contribution-function extrapolation. T and n_e denote the normalized temperature and electron-density coordinates. The paired outputs are shown together because the physical diagnostic depends on their predicted 174/175 relationship.

Method	Frozen expression or expression pair
EvoTrace	$\hat{G}_{175}(T,n_e) = e^{1.1386-1.0685T^2}[1-0.5741(n_e/5)^{-0.1089}]$, $\hat{G}_{174}(T,n_e) = e^{1.0171-1.0348T^2}$
PhyE2E	$G(T,n_e) = C_1\dfrac{n_e+C_2}{n_e+C_3}(C_4T+1)^{C_5}$ $\times \exp[C_6(C_7T+1)^{C_8}]$
BSR	$G(T,n_e) = 2\times10^{-6}T$
End-to-End	$G(T,n_e) = \left(-0.106+\dfrac{0.0413}{-1.45\times10^{10}n_e+0.426}\right)$ $\times\left[-0.0328+15.566e^{-9.861Q(T,n_e)^2}\right]$, $Q(T,n_e) = -9.235T+1-\dfrac{0.197}{2.86\times10^{11}n_e+26.889}$
SR-Scientist	$\hat{G}_{174} = \dfrac{a_0T^{a_1}n_e^{a_2}e^{-a_3T}}{1+a_4T^{a_5}}$, $\hat{G}_{175} = e^{b_0+b_1\log T-b_2T-b_3T^2}[1+b_4\log n_e+b_5(\log n_e)^2]$
LaSR	$\hat{G}_{174} = \dfrac{A(T)-0.00908848/T}{\sin(\sin((n_e/T)^{0.167721}))}$, $A(T) = \left[\dfrac{\cos(\sin(\log(0.726675T+0.234034)))}{0.966902}\right]^{30.876185}$, $\hat{G}_{175} = B(T)\sin[\sin(\sin(\sin(n_e^{0.589417}-0.262822)))]$, $B(T) = 0.0519518\dfrac{\sin T}{T}+\left(\dfrac{0.765664}{T}\right)^{e^T-5.755614/T}$
LLM-SR	$\hat{G}_{174} = \dfrac{P_2(T,n_e)e^{A(T,n_e)}e^{B(T,n_e)}\lvert T\rvert^{3.65913}\lvert n_e\rvert^{-0.10734}}{D_2(T,n_e)}$ $+0.01932\log(1+\lvert T\rvert+\lvert n_e\rvert)$, $A(T,n_e) = \mathrm{clip}[-(4.24910T^2-0.000194n_e^2+0.07845Tn_e),-20,0]$, $B(T,n_e) = \mathrm{clip}(4.06665T+0.21466n_e,-10,10)$, $D_2(T,n_e) = \max(1+2.09165\lvert T\rvert+0.20708\lvert n_e\rvert,10^{-6})$, $P_2(T,n_e) = -7.65949+7.81405T+0.72170n_e-1.44520Tn_e$ $+12.99467T^2+0.003869n_e^2$, $\hat{G}_{175} = \dfrac{N_3(T,n_e)}{1-1.83140T+0.02453n_e+0.88689T^2-0.001970n_e^2}$ $+1.26525e^{-0.29625(T+n_e)}$, $N_3(T,n_e) = -1.26505+2.65516T+0.08188n_e-1.65723T^2$ $-0.004481n_e^2-0.05395Tn_e+0.25347T^3+0.000190n_e^3$ $+0.03717T^2n_e-0.002120Tn_e^2$

D.4 Component Ablation Analysis

We further examine the contribution of the three major components of EvoTrace: CoTraS, MAHASE, and Evidence-preserving Support. The ablations are performed on representative tasks that expose different aspects of symbolic discovery, including long-horizon temporal behavior, spatial extrapolation, and coupled physical relationships.

Effect of CoTraS. Removing CoTraS increases sunspot MSE from 1422.16 to 3998.17 and differential-rotation MSE from 2.26×10^{-5} to 1.54×10^{-4}. On Fe X, the individual line-wise MSEs become smaller, but RatioMSE deteriorates from 1.2333×10^{-2} to 1.4650×10^{-1}. This contrast indicates that local numerical fit alone does not guarantee recovery of the coupled physical

Table 12: Representative frozen expressions for lunar-tide radial electric-field prediction. We use $L = x_0$, $m = 10\,\mathrm{MLT}/24$, and $\ell = 10\,\mathrm{LLT}/24$. The expressions correspond to the frozen models used in the predictive and physical-diagnostic evaluation.

Method	Frozen expression
EvoTrace	$E_r = 4.039885L^{-2}\Big[1 + 8.913179\sin(-0.002451m + 3.075693) + 1.389836\cos(1.256127m - 3.206327) - \dfrac{0.004346m}{L - 1.027866} - \dfrac{0.002571\ell}{L - 1.027866}\Big]$
PhyE2E	$E_r = C\dfrac{N(m,\ell)}{D(L,m)}$, $N(m,\ell) = [a - b\cos(k_{\mathrm{LLT}}\ell)][d\cos(k_{\mathrm{MLT}}m) - e]$, $D(L,m) = [f - g\cos(hm - q)](uL^2 - vL + w)$
BSR	$E_r = \exp\{\cos[\cos(0.00523\,\mathrm{LLT}^3\mathrm{MLT}^3)]\}$
End-to-End	$z = 0.00333 + \dfrac{1}{1 - 9/(9x_4 + 0.001)}$, $Q = (0.010 - 0.060L)[3.75\,\mathrm{LLT} + 3.75\,\mathrm{MLT} + 32.4z^2 - 0.003]$, $E_r = 0.729(1 - Q^{-1})^3 + 0.002$
SR-Scientist	$E_r = a_0 L^{a_1}[a_2 + a_3\cos(a_4 m) + a_5\sin(a_6 m)]$
LaSR	$E_r = \dfrac{10.25917}{\exp[L + \cos(1.247887m)]}$
LLM-SR	$N = 0.4721 + 6.6162L - 1.4596m + 0.2735\ell + 0.4534Lm - 0.0188m\ell - 0.2194L\ell + 0.0226Lm\ell$, $D = 1 - 4.3491L + 2.0053m + 0.0101\ell + 0.7438L^2 - 0.4073m^2 + 0.00043\ell^2 + 0.9868L^3 + 0.02249m^3$, $E_r = N/D$

Note. The End-to-End expression retains x_4 verbatim because its mapping to the three reported physical coordinates is not specified in the original source expression.

Table 13: Sensitivity of the auxiliary cross-task ordering to the composite weights. The analysis uses the same frozen task-level metrics as the headline composite-score pipeline and does not affect model selection.

Weights	Top-ranked	Median Gain	Mean Kendall τ
Default $(.40, .30, .20, .10)$	$3/3$	$1.647\times$	1.000
Equal $(.25, .25, .25, .25)$	$3/3$	$1.522\times$	1.000
One-at-a-time ± 0.10	$3/3$	$[1.527, 1.777]\times$	$[0.889, 1.000]$

relationship; trajectory-level evidence helps distinguish structurally useful edits from changes that improve only isolated fitting objectives.

Effect of MA-HASE. Without MA-HASE, differential-rotation MSE rises to 4.71×10^{-4}, while Δ_{pole} increases from 0.0051 to 1.8273. The effect is even more pronounced for Fe X, where RatioMSE increases from 1.2333×10^{-2} to 3.0253×10^{90}. These results show that mechanism-anchored admissibility is important not only for predictive fitting but also for preventing structurally unstable hypotheses from surviving the search.

Effect of Evidence-preserving Support. Removing Evidence-preserving Support reduces the sunspot avg-R from 0.7915 to 0.0198 and introduces 75 negative-prediction months. For differential rotation, Δ_{pole} increases from 0.0051 to 1.4765. Because the evaluator, search budget, candidate selector, and numerical validation remain unchanged in this ablation, the degradation isolates the contribution of reusing attributed evidence and historical failure information across iterations.

Table 14: Component ablation on representative scientific tasks.

Configuration	Sunspot	Differential rotation	Fe X
	MSE / avg-R / Neg. months	MSE / Δ_{pole}	MSE_{175} / MSE_{174} RatioMSE / AST Complexity
EvoTrace	**1422.1558 / 0.7915 / 0**	2.26×10^{-5} **/ 0.0051**	2.5191×10^{-9} **/** 2.1584×10^{-8} 1.2333×10^{-2} **/ 40**
w/o CoTraS	3998.1668 / -0.5388 / 0	1.54×10^{-4} / 0.0054	2.4466×10^{-10} / 4.0270×10^{-10} 1.4650×10^{-1} / 72
w/o MA-HASE	4284.5591 / -0.2894 / 0	4.71×10^{-4} / 1.8273	6.6712×10^{-10} / 1.7174×10^{-9} 3.0253×10^{90} / 44
w/o Both	3959.8955 / -0.2052 / 390	3.72×10^{-5} / 0.0081	3.6412×10^{-2} / 2.1626×10^{-1} 1.3657×10^{-1} / 35
w/o Evidence-preserving Support	3766.2288 / 0.0198 / 75	4.19×10^{-4} / 1.4765	3.6413×10^{-2} / 2.1655×10^{-1} 1.4214×10^{-1} / 30

Combined effect. The *w/o Both* setting further demonstrates that trajectory steering and mechanism-guided admissibility are not interchangeable. Removing both substantially degrades sunspot behavior and Fe X coupled diagnostics even when some individual task metrics remain competitive. Overall, the ablations support a complementary division of labor: CoTraS extracts reusable trajectory-level search evidence, MA-HASE constrains the admissible hypothesis space, and Evidence-preserving Support carries verified successful and failed evidence across iterations.

E Representative Symbolic Regression Paradigms

This section provides additional background on the three major symbolic regression paradigms summarized in the main paper: genetic-programming-based, deep-learning-based, and large-language-model-based methods.

Genetic-programming-based methods. Genetic-programming-based symbolic regression searches over expression structures through evolutionary variation, fitness-based selection, and numerical constant optimization. Early work by Schmidt & Lipson (2009) demonstrated that evolutionary search can recover compact analytical relationships directly from observational data. PySR (Cranmer, 2023) improves the scalability and practical efficiency of evolutionary symbolic regression through parallelized search, expression simplification, and numerical constant fitting. GP-GOMEA (Virgolin et al., 2021) further introduces gene-pool optimal mixing to exploit dependencies among expression components during structural search. These methods provide substantial flexibility in exploring symbolic structures, but candidate refinement is still primarily driven by stochastic variation and scalar fitness values, with limited mechanisms for retaining structured information about why particular edits succeed or fail.

Deep-learning-based methods. Deep-learning-based symbolic regression learns data-driven priors over mathematical expressions and uses them to generate candidate equations from numerical observations. DSR (Petersen et al., 2021) trains a recurrent policy using rewards derived from equation quality, casting symbolic regression as a reinforcement-learning problem. NeSymReS (Biggio et al., 2021) formulates equation discovery as sequence generation and uses a Transformer to infer symbolic expressions from numerical samples. End-to-End symbolic regression (Kamienny et al., 2022) further develops a Transformer-based framework for directly predicting equations from tabular observations. For space-physics applications, PhyE2E (Ying et al., 2025) incorporates physical priors and variable decomposition to improve equation discovery under domain-specific constraints. These approaches enable efficient amortized inference, although their behavior can depend strongly on the coverage of the training distribution and on the inductive biases encoded during model design and training.

Large-language-model-based methods. Recent work uses large language models to incorporate scientific knowledge, natural-language reasoning, and executable program generation into symbolic regression. LLM4ED (Du et al., 2024) iteratively refines equations using previously explored can-

didates and their empirical performance. LLM-SR (Shojaee et al., 2025) represents equations as executable programs and combines LLM-based proposal generation with evolutionary optimization. LaSR (Grayeli et al., 2024) extracts reusable textual concepts from high-performing hypotheses and feeds them back into subsequent search. More recent systems move toward agentic scientific discovery: SR-SCIENTIST (Xia et al., 2026) integrates data analysis, code execution, and equation evaluation within an iterative search loop. These developments expand symbolic regression beyond independent candidate generation toward iterative, knowledge-guided hypothesis exploration.

E.1 Limitations and Future Directions

We identify three directions for extending EvoTrace. First, our experiments currently use three open-weight models; whether our findings generalize to frontier closed-source, dense, or domain-specialized models remains an open question, which we plan to study systematically.

Second, EvoTrace relies on budget-based stopping rules rather than certified convergence. Principled convergence diagnostics and adaptive stopping criteria are a natural extension of our search-state analysis.

Third, unlike classical symbolic-regression methods, EvoTrace incurs additional token overhead from repeated LLM calls for generation, comparison, and revision.

Despite these limitations, EvoTrace already delivers strong and stable results on a wide range of symbolic-regression tasks, and we hope its design inspires further exploration of agentic approaches to scientific discovery.