

Influence-Guided Symbolic Regression: Scientific Discovery via LLM-Driven Equation Search with Granular Feedback

ICML 2026



**Evgeny
Saveliev**



Samuel Holt



**Nabeel
Seedat**



**David L.
Bentley**



**Jim
Weatherall**



**Mihaela van der
Schaar**

<https://arxiv.org/abs/2604.12060>



van_der_Schaar
\ LAB

vanderschaar-lab.com



Term-Level Credit for Scientific Equation Search

Discovered equation example:

$\text{pause_score} = 0.1430$ (constant baseline term)

$+ 0.0878 \text{signal}_{\text{CpG_methyl}} - 0.2199 \ln(1 + \text{signal}_{\text{CpG_methyl}})$

$+ 0.0103 \ln(1 + \text{down_nuc} - \text{up_nuc})$

$- 0.0745 \ln(1 + \text{signal}_{\text{MNase}}) \dots$

Repeat until: total_budget = 30 expansions.

Dataset & Problem

<Train / Val / Test>

Scientific Context

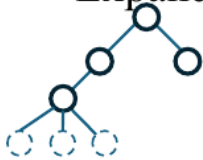
MCTS

Nodes are equations

Selection

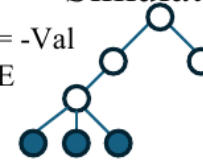


Expand



Simulation

Reward = -Val
MSE



Backprop



1) Term Generation (LLM)

LLM generates candidate basis functions $\{\psi_j(x)\}$; history and context provided

Example

$[1/\sqrt{v}, \log_{10}(v), \text{cbrt}(v), e^{-\sqrt{v}}, v \cdot \log v]$

2) Candidate Evaluation & Influence Computation

Design matrix Φ from all candidates; fit linear model weights w on Train

Compute Δ_j on Validation: $\Delta_j = \text{MSE}(\text{full}) - \text{MSE}(w_j \leftarrow 0)$

3) Term Pruning

Deterministic (default): retain top K best terms by Δ_j

or

LLM: Uses (w_j, Δ_j) and history to select terms; history and context provided

KEEP: $v, v \cdot \log v$
DROP: others

Val MSE
per child



Superior Performance Across Diverse Scientific Tasks

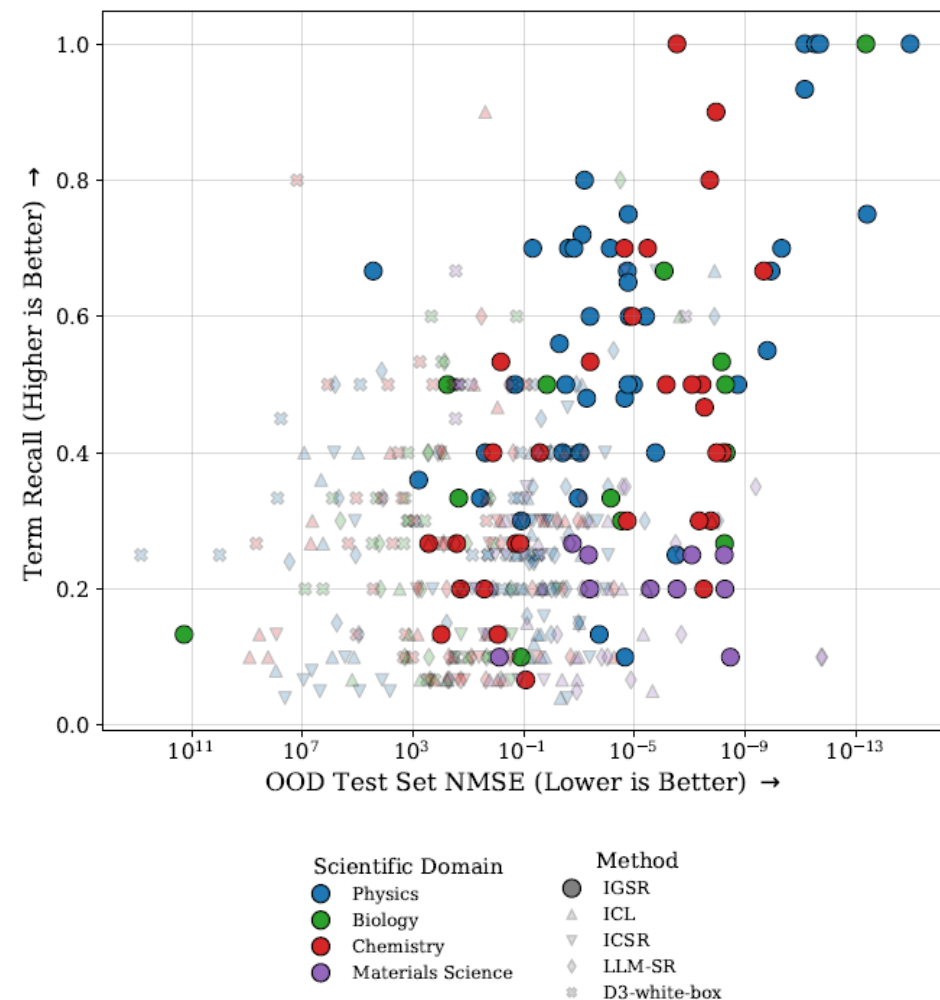
(b) LLM-SRBench Evaluation.

Method	ID Test Set		OOD Test Set		Symbolic Metrics	
	NMSE	Acc _{0.1}	NMSE	Acc _{0.1}	Term Recall	Symb. Acc.
ICL	3.91	2.53	4.60	2.41	3.50	1.35
D3-w.b.	6.89	4.16	6.27	4.34	3.16	1.34
LLM-SR	3.71	2.38	3.92	2.33	3.13	1.29
LaSR	5.53	3.94	4.77	3.33	4.60	1.30
ICSR	4.21	3.34	4.22	2.71	3.70	1.33
IGSR-Agent	<u>2.41</u>	<u>1.50</u>	<u>2.45</u>	<u>1.49</u>	1.98	1.16
IGSR	<u>1.35</u>	<u>1.12</u>	<u>1.77</u>	<u>1.21</u>	1.99	<u>1.17</u>

(a) Benchmark Datasets Evaluation.

Method	LC MSE ↓	LC-C MSE ↓	LC-CR MSE ↓	C-19 MSE ↓	RNAPol MSE ↓	Warf MSE ↓	Avg. Rank ↓
SINDy	335±3.14	0.838±6.35e-3	0.664±5.14e-3	1.43e-7±4.48e-9	— [†]	0.914±0.169	5.50
GPLearn	7.56±1.11	46.8±15.5	46.8±4.91	7.13e-4±5.06e-4	0.0204±5.55e-4	2.53±0.169	7.33
PySR	669±7.86	0.12±0.0697	0.399 ± 0.123	4.23e-7±8.3e-8	— [†]	29.9±0.859	6.17
ZeroShot	2.13e13±4.35e13	4.97e3±3.67e3	2.54e3±2.74e3	1.34e8±2.09e8	1.35e5±1.75e5	5.3e3±1.07e4	10.50
ZeroOptim	0.142±0.119	86.2±27.3	122±6	1.41e-7±1.04e-7	0.0130±2.87e-4	0.861±0.177	5.50
ICL	0.0557±0.0486	21.2±9.8	63.3±16.5	9.35e-8±1.77e-8	0.0119±3.52e-4	0.784±0.193	3.83
D3-w.b.	1.01e4±1.27e4	45±28.9	253±273	7.81e-6±2.48e-7	0.043±0.0366	1.15±0.343	8.00
LLM-SR	33.4±0	42.2±35.6	32.1±48.4	4.53e-4±9.12e-4	1.44±1.91	1.24±0.564	7.00
LaSR	658±7.31	1.71±1.16	3.97±3.21	2.59e-6±8.66e-7	0.0172±6.49e-4	30.1±0.992	6.33
ICSR	0.407±0.244	0.688±0.39	6.1±1.05	1.03e-7±1.6e-8	— [†]	0.497±0.0646	4.17
DyNODE	497±18.4	355±3.63	398±14.2	7.48e-5±5.58e-7	0.0136±6.7e-5	0.593±0.0769	-
RNN	1.33e-3±4.85e-4	6.85e-3±2.2e-3	0.013±0.00552	5.88e-8±1.5e-9	7.28e-3±4.79e-5	0.756±0.139	-
Transformer	0.0375±6.53e-3	0.179±0.0135	0.366±0.0269	4.05e-7±5.05e-8	7.24e-3±3.3e-5	0.804±0.158	-
XGBoost	0.012±9.64e-4	0.155±4.88e-3	0.322±8.84e-3	8.43e-8±5.36e-9	7.58e-3±3.79e-5	1.67±0.473	-
IGSR	5.64e-5±6.79e-5	0.0013±0.0010	0.0141±0.0087	5.01e-8±1.78e-9	0.0111±0.0004	0.565±0.113	1.17

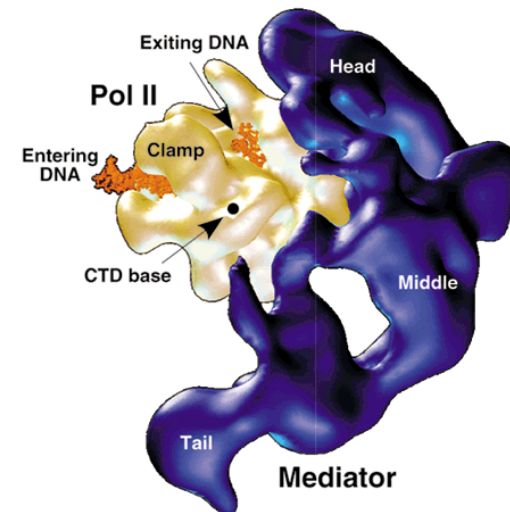
Structural Fidelity vs. Predictive Accuracy



Discovery Confirmed by Wet-Lab Experiment

The Discovered Equation: IGSR converged on the following model. Notably, the system incorporated the new signals (H3K27ac and CpG_methyl) while discarding previous markers like H3K4me3, yet achieving a superior MSE of 0.0102 (compared to 0.0111 in prior experiments). The terms contributing to the novel DNA methylation hypothesis are highlighted in Violet.

$$\begin{aligned}\text{pause_score} = & 0.1430 \quad (\text{constant baseline term}) \\ & + 0.0878 \text{signal}_{\text{CpG_methyl}} - 0.2199 \ln(1 + \text{signal}_{\text{CpG_methyl}}) \\ & + 0.0103 \ln(1 + \text{down_nuc} - \text{up_nuc}) \\ & - 0.0745 \ln(1 + \text{signal}_{\text{MNase}}) \\ & - 0.0923 \mathbb{1}_{\{\text{gene_region}=\text{TSS}\}} + 0.0342 \mathbb{1}_{\{\text{gene_region}=\text{TSS}\}} \cdot \ln(1 + \text{signal}_{\text{MNase}}) \\ & - 0.0732 \ln(1 + \text{signal}_{\text{H3K27ac}}) \\ & + 0.0172 \mathbb{1}_{\{\text{gene_region}=\text{TSS}\}} \cdot \text{signal}_{\text{H3K27ac}} \\ & + 0.0474 \mathbb{1}_{\{\text{exon}\}} \cdot \ln(1 + \text{signal}_{\text{H3K27ac}}) \\ & + 0.0274 \mathbb{1}_{\{\text{seq}_0=G\}} - 0.0143 \sum_{pos \in \{-1,0,1\}} (\mathbb{1}_{\{\text{seq}_{pos}=G\}} + \mathbb{1}_{\{\text{seq}_{pos}=C\}}) \\ & + 0.0112 \ln(1 + \text{gene_length})\end{aligned}$$



Wet-Lab Verification of IGSR Hypothesis

- **IGSR Prediction:** The discovered equation contained methylation terms with negative coefficients, predicting that methylation *suppresses* pausing.
- **Experimental Observation:** Pharmacological depletion of DNA methylation (using 5-azacytidine) caused a marked *increase* in transcriptional pausing.
- **Conclusion:** The wet-lab results support the IGSR-derived hypothesis that intragenic CpG methylation facilitates elongation by suppressing pausing, demonstrating the framework's ability to drive expanding scientific knowledge.

