Parsimony, Not the Clip: What Controls the Search in Reinforcement-Learning Symbolic Regression

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi
Transformer Lab
deep@lab.cloud

Abstract

Symbolic regression (SR) recovers a closed-form law from sampled (x, y) data. We study how a reinforcement-learning objective *shapes* this search rather than chasing a leaderboard: a per-equation policy emits prefix-notation expressions, numeric constants are fit by BFGS inside an exactly computed reward (predictive fit minus a parsimony penalty $\lambda \cdot$ complexity), and the policy is post-trained with a DAPO-style decoupled-clip objective. Holding an entropy regularizer fixed, we ask a mechanistic question — which knob governs where the policy lands on the accuracy–parsimony frontier — and contribute three findings. **First**, the parsimony coefficient λ cleanly and monotonically sets the operating point on the frontier, controlling how widely the policy explores the space of expressions and how complex they become. **Second**, the DAPO clip-higher asymmetry does *not*: varying it leaves recovery unchanged and only weakly and inconsistently affects exploration — a negative that persists when the clip is forced to engage and on harder, non-separable problems — and we find that it is the entropy regularizer, not the clip, that sustains exploration. **Third**, on a held-out Feynman benchmark our final model recovers a modest fraction of equations (symbolic recovery 0.205, three-seed mean), in the range of a deep-symbolic-regression baseline and below the genetic-programming incumbent PySR; our single-seed baselines preclude a significance claim, and we instead characterize the model's failure modes (trigonometric laws, high-complexity expressions, and spurious fits that match the data numerically without recovering the true structure). We present this as a mechanistic study of how the training objective shapes symbolic search, with entropy regularization held fixed, not a state-of-the-art result.

1 Introduction

Symbolic regression asks a model to propose a short, readable mathematical expression — a *law* — that could have generated an observed dataset of (x, y) samples drawn from an unknown function. Unlike most machine learning, the output is a symbolic formula rather than a weight matrix, which makes SR a literal instance of AI for scientific discovery: the consumer of the prediction is an exact scorer and, ultimately, a human reading the recovered law. The task has strong incumbents in genetic programming (PySR [1]) and in RL-policy methods (DSR/DSO), so a marginal leaderboard improvement is not the most useful contribution one can make.

We instead ask a *mechanistic* question: given an RL objective with a clipped importance-sampling ratio and an explicit parsimony penalty, which knob varies primarily with how the policy searches the space of expressions — collapsing to trivial under-fitting forms versus drifting into over-complex

overfitting ones — and hence where it lands on the accuracy–parsimony Pareto frontier? The clipping/ratio mechanism is appealing because the LLM-RL literature explicitly frames the decoupled "clip-higher" bound as a way to preserve low-probability-token exploration [2, 3]; whether that story survives transfer to a tiny symbolic vocabulary is an open and falsifiable question. We treat the accuracy–parsimony tradeoff as the central dependent variable and raw recovery as a secondary sanity check, training a small per-equation policy with an exactly computed fit-minus-parsimony reward (no learned judge) and a DAPO-style decoupled-clip objective. Throughout, an entropy bonus (entropy coefficient 0.2) is *held fixed* as an enabling regularizer: all of the λ-versus-clip findings below are conditional on it, and a dedicated entropy-off ablation shows that it is entropy — not the clip — that sustains exploration. The story is therefore honestly three-way (parsimony λ, entropy, clip), not a two-way contest between λ and the clip.

The honest headline is a positive result paired with a hedged negative. The parsimony coefficient λ varies primarily with exploration and frontier position: on the 3-seed means the trend in unique structures and complexity is monotone and statistically significant, satisfying our primary criterion (a statistically significant monotone relationship) for the $\lambda \rightarrow$ exploration/complexity axis. The DAPO clip-higher asymmetry, our originally intended headline lever, does *not* improve recovery and only weakly and non-monotonically modulates exploration ($r = 0.68$ on non-separable problems, $r = 0.21$ on harder Feynman equations; $n = 5$ ε-levels, neither correlation statistically significant, $p \approx 0.2$) — weak, non-monotone, and yielding no recovery benefit, even when the clip is forced to engage on harder, non-separable problems. We report this clip result as a hedged absence-of-recovery-effect, and acknowledge that with $n = 5$ it is underpowered for a strong null. We document it carefully because the field's mechanism claims are usually asserted on large-vocabulary language models, not tested on a grammar of roughly a dozen base symbols.

Our contributions are (all conditional on a fixed entropy regularizer):

1. A two-dimensional $\varepsilon_{\text{high}} \times \lambda$ study of an RL-SR objective in which the parsimony coefficient varies primarily with exploration and accuracy–parsimony frontier position, with a statistically significant monotone trend (on the 3-seed means) for unique structures and complexity (§5, H2).

2. A hedged negative result: the DAPO clip-higher asymmetry does not improve recovery in this setting, and only weakly and non-monotonically modulates exploration ($r = 0.68$ on non-separable problems; $n = 5$, not significant, $p \approx 0.2$) with no recovery benefit. The recovery-null survived a combined stress manipulation (more clip engagement, larger policy and group, harder and non-separable problems) but we did not isolate each factor (§5, H1).

3. A reconciling ablation: our `noclip` variant (our objective minus the clip) collapses, but the clip-free DSR objective explores and recovers normally, so we do not claim clipping is necessary for stable clip-free RL-SR; the clean finding is that the clip *degree* ($\varepsilon_{\text{high}}$) is inert (§5, H3).

4. A faithful competitive and stratified evaluation against single-seed PySR and DSR baselines, with characterized failure modes (trigonometric forms, high complexity, spurious numeric fits, variable-count blow-up) (§5, §6).

2 Related Work

RL for symbolic regression. The deep-symbolic-regression / optimization (DSR/DSO) lineage frames SR as an MDP: a policy generates a prefix (pre-order) traversal of an expression tree with

parent and sibling tokens as the observation, a `const` placeholder fit by nonlinear optimization inside the reward, and in-situ constraints applied as additive logit masks [4]. Its operator library closely overlaps with ours, and the decoupled "policy emits skeleton, then BFGS fits constants" pipeline with multi-restart constant fitting is the standard DSR recipe [5, 6, 7]. Protected operators (protected division, $\sqrt{|\cdot|}$, capped exp) are the usual guard against NaN/overflow during fitting [7]. Hayes et al. [4] also caution that average reward is the wrong objective for SR — risk-seeking yields a lower mean reward but a better best expression — and that some equations are simply unrecoverable by all methods. Dynamic-gating and noise-resilient RL variants extend this family [8].

Clipped-ratio LLM-RL and exploration. The clipping/ratio mechanism is the knob we study, and the LLM-RL literature already frames it as exploration control. DAPO's decoupled clip holds ε_{low} fixed and raises $\varepsilon_{\text{high}}$, explicitly motivated as preventing the upper clip from restricting exploration of low-probability tokens [2]. The verifiable-reward exploration literature sharpens this: it is the survival of low-probability tokens, not entropy per se, that sustains exploration, and the lower clip bound is what kills them [3]. A token-level entropy-flow account explains why clip-higher preserves exploration and motivates two-sided objectives [9], and entropy can be targeted to a set level as an alternative dial [10]. DAPO's stability fixes — removing the KL penalty, token-level (not sample-level) loss, critic-free group-relative advantages — transplant directly [2, 11], though sample-level averaging induces a brevity bias that for prefix expressions could masquerade as parsimony pressure [12]. Controlled PPO/GRPO/DAPO comparisons report that clipped objectives are highly sensitive to the exact clip threshold and that RL is high-variance, motivating fine grids and multiple seeds [12, 11]. Crucially, this whole low-probability-token argument was developed on large (10^4–10^5-token) vocabularies; whether it transfers to a grammar of roughly a dozen base symbols is exactly the open question our negative result addresses.

Parsimony and model selection. Multiple studies converge on minimum description length (MDL) as a principled parsimony axis: it measures complexity in the same units as data misfit and includes a Fisher-information term penalizing over-precise fitted constants, which directly attacks constant-based reward hacking [13, 7]. On a controlled benchmark MDL ranks the true expression best while fit-only MSE is worst [7]. Useful complexity definitions beyond raw node count include structural complexity [14], per-operator weights [15], and DAG node counts with common-subexpression sharing [16]. TPSR-style λ-knob methods expose an explicit fit-versus-complexity λ knob, the closest published analogue to our sweep [17, 18].

Benchmarks, metrics, and pitfalls. The field standard is the Feynman database [19] evaluated under SRBench [20], with symbolic-equivalence recovery and per-problem significance testing [4, 16, 18, 17]; subset sizes vary across papers, so a documented, hashed subset is necessary [6]. Published Feynman recovery counts anchor our frontier (e.g. PySR [1] is a strong incumbent) [17, 15, 14]. Several cautions inform our protocol: high R^2 is not recovery, as many regressors overfit to near-perfect accuracy without recovering structure [16, 18]; entropy is a misleading exploration proxy, so unique-structure count is the preferred dependent variable [3, 11]; `sympy` equivalence checks are unreliable and need a numeric fallback [14, 16]; and Feynman is mostly separable and "easy" for structure-first methods, motivating a non-separable stratum so the exploration mechanism is not masked [6]. We position our work in the gap these two literatures leave: the RL-SR lineage has the task, reward, and metrics but uses risk-seeking REINFORCE and reports only end recovery; the

clipped-ratio lineage has the exploration mechanism but, to our knowledge, has not been applied to SR or to a parsimony objective.

3 Method

Policy and MDP. We use a per-equation RL search, training a policy from scratch for each dataset. A small single-layer GRU policy (hidden size 64) emits expressions as token sequences in prefix (Polish) notation over a fixed library $\{+, -, \times, \div, \sin, \cos, \exp, \log, \sqrt{}\}$, variables, and a constant-placeholder token. Following the DSR-style MDP, generation is a pre-order traversal of the expression tree with the parent and sibling tokens as the policy observation; in-situ constraints (length bounds, "children of an operator not all constants," no nested $\log(\exp(\cdot))$) are applied as additive logit masks.

Exact reward. For each proposed structure the constant placeholders are fit by multi-restart BFGS on the fit-train points only, never on the data used to score the reward. The reward is computed exactly — there is no learned judge:

$$r \;=\; \underbrace{\frac{1}{1+\mathrm{NMSE}}}_{\text{fit}} \;-\; \lambda\, L, \tag{1}$$

where the fit term is the NMSE-normalized predictive accuracy on a held-out reward point set and L is the expression complexity (preorder node count). NMSE normalization absorbs cross-dataset scale heterogeneity without rescaling the data, which would corrupt symbolic recovery.

Optimization objective. The policy is post-trained with a DAPO-style decoupled-clip objective: group-relative, critic-free advantages $\hat{A} = r - \mathrm{mean}(r)$ (optionally divided by the group std), token-level (not sample-level) policy-gradient loss, and no KL penalty (there is no reference policy worth staying near for per-equation from-scratch search). The clip is decoupled into a fixed lower bound $\varepsilon_{\text{low}} = 0.20$ and a swept upper bound $\varepsilon_{\text{high}}$; to avoid collapsing the sampling space, we hold ε_{low} fixed and vary only $\varepsilon_{\text{high}}$ to trace the exploration curve. We sweep only $\varepsilon_{\text{high}}$ to match the DAPO "clip-higher" motivation we set out to test; we note that the verifiable-reward literature instead fingers the *lower* clip as what kills low-probability tokens [3], so an ε_{low} sweep is a natural follow-up our grid does not cover. Throughout, an entropy bonus (entropy coefficient 0.2) is held fixed as an enabling regularizer rather than swept; all λ-versus-clip findings are conditional on it, and an entropy-off ablation (§5) shows that entropy — not the clip — is what sustains exploration.

Hypotheses. We test three: **H1** (clip → exploration), that holding ε_{low} fixed and raising $\varepsilon_{\text{high}}$ increases exploration (unique-structure count, complexity-distribution width); **H2** (frontier control), that $(\varepsilon_{\text{high}}, \lambda)$ jointly vary with the accuracy–parsimony operating point, with an intermediate region maximizing recovery and the corners collapsing; and **H3** (clip degree vs. plain RL), that varying the clip degree modulates exploration and recovery. We do *not* hypothesize that clip presence is necessary for stable RL-SR in general: a clip-free DSR-style objective is itself a stable RL-SR method, so any collapse we observe is specific to our particular `noclip` variant (our objective with the clip removed), and the clean, defensible claim is about the clip *degree* ($\varepsilon_{\text{high}}$).

4 Experimental Setup

Data. The dataset is the Feynman Symbolic Regression Database [19] served through PMLB [21] under the SRBench [20] sampling protocol: 119 datasets of 100,000 noiseless i.i.d. rows each. We focus on the low-dimensional ($\leq$ 3-variable) subset of 53 datasets (50 expressible in our library); the sweep uses a tractable 5-equation subset (complexity ≤ 10), and baselines and the final model are scored on a 13-equation stratified subset. To expose the exploration mechanism where Feynman's separability would mask it, we add an 8-equation non-separable synthetic stratum generated over the same operator library and row schema.

Splits and leakage controls. Each equation's rows are deterministically subsampled and split three ways into disjoint roles: 2,000 fit-train points (BFGS fits constants here only), 4,000 reward-heldout points (the RL fit reward; constants are never fit here), and 4,000 final-test points (held-out evaluation the reward never saw). The ground-truth formula is used only by the scorer, never exposed to the policy.

Recovery metrics. We report three notions of success. *Symbolic recovery* requires `sympy` to prove equivalence to the ground truth (allowing fitted constants). *Numeric recovery* is a dense-grid fallback at 10^{-4} relative tolerance (chosen because BFGS-fitted constants are only $\sim 10^{-4}$ accurate), which catches symbolic-prover false negatives. *Accuracy recovery* is held-out $R^2 \geq 0.999$. Symbolic and numeric/accuracy recovery are reported separately so the tolerance choice does not inflate the headline; the accuracy-vs-symbolic gap *is* the spurious-fit rate.

Protocol and compute. Every reported configuration uses 3 seeds; we report 3-seed means, since per-equation RL-SR is genuinely seed-sensitive. Throughout, "$\pm$" denotes the standard deviation across the 3 seeds. Where a per-cell spread is not given explicitly (e.g. stratified or sweep-cell means), it is omitted for space, not because variance is negligible; the headline 0.205 alone carries a 3-seed std of 0.096, so all 3-seed means should be read as carrying comparable per-cell spread. The baselines (PySR, DSR, clip-off), by contrast, are single seed-0 runs, and we attach no cross-method significance to their ordering. The final configuration is $\varepsilon_{\text{high}} = 0.28$, $\lambda = 0$, entropy coefficient 0.2, group size $G = 8$, 400 training steps, maximum length 16. The workload is CPU-bound (GRU + scipy-BFGS + `sympy`); runs used one A10-class GPU per job on a a single cloud GPU node, with total compute of about 43 GPU-hours.

5 Results

Table 1 reports the baseline contrast on the 13-equation subset. All three baselines are single runs (single seed, seed 0). PySR [1] leads at 0.615 symbolic recovery, DSR (risk-seeking REINFORCE, fit-only) recovers 0.308, and the clip-off control falls to 0.000 recovery with mean complexity 2.08. We tie this collapse to the *complexity* 2.08 *and recovery* 0.0 (trivial expressions), *not* to the unique-structure count, which at 10.5 is essentially the same as PySR's 10.2 and so does not by itself read as an exploration collapse. Importantly, the clip-free DSR objective visits 171.7 unique structures and recovers 0.308, so *clipping is not necessary for stable clip-free RL-SR in general* — the collapse is specific to our `noclip` variant (our objective with the clip removed).

Table 1: Baselines on the 13-equation subset, each a **single run (single seed, seed 0)**. Symbolic recovery and exploration (unique structures). We ran no baseline for multiple seeds and attach no significance to the cross-method ordering; note our DAPO seed-0 recovery (0.31) itself exceeds DSR's single 0.308.

Method (single seed, seed 0)	Recovery	Mean complexity	Unique structures
PySR (GP incumbent)	0.615	10.5	10.2
DSR (fit-only REINFORCE)	0.308	5.5	171.7
Clip-off (no clip)	0.000	2.08	10.5

Table 2: Final DAPO model vs. baselines and success gates (3 seeds; baselines are single seed-0 runs). Recovery is a 3-seed mean; "±" (elsewhere) denotes the 3-seed std. We attach no cross-method significance to the baseline ordering (§5).

Metric (13-eq subset unless noted)	Result	Reference (single seed)
Symbolic recovery, sympy+numeric fallback (DAPO, ours)	0.205	DSR 0.308; PySR 0.615
Accuracy solution rate ($R^2 \geq 0.999$)	0.282	respectable
Symbolic recovery, sympy-only (exact)	0.103	—
Recovery, 5-eq subset ($\lambda = 0$)	0.60	—

Final DAPO model (competitive sanity). Table 2 places our final DAPO model against the single-seed baselines and our success criteria. On the 13-equation subset DAPO reaches 0.205 symbolic recovery (3-seed mean; per-seed 0.31/0.23/0.08, std 0.096). This is in the same range as DSR's single run (0.308) and below PySR's single run (0.615); we did not run the baselines for multiple seeds, so we attach no significance to this ordering — indeed our own seed-0 (0.31) already exceeds DSR's single 0.308. Its accuracy solution rate ($R^2 \geq 0.999$) is 0.282 while its sympy-only (exact) symbolic recovery rate is 0.103, the sympy+numeric-fallback rate being 0.205; the gap is the spurious-fit rate (below). On the tractable 5-equation subset at $\lambda = 0$, recovery reaches 0.60 (3-seed mean). We frame this as a mechanistic study whose recovery lands in the same range as the single-seed DSR run, not a state-of-the-art claim.

H2: the λ-frontier (positive result). The parsimony coefficient λ is the lever the search varies primarily with. In the broad sweep, on the 3-seed means unique structures fall monotonically $144 \to 67 \to 31 \to 16 \to 3$ and mean complexity falls monotonically $7.7 \to 1.0$ as λ rises through $\{0, 0.005, 0.01, 0.02, 0.05\}$; we restrict the term "monotone" to these two quantities (unique structures and complexity), where the 3-seed means are monotone. Recovery, by contrast, is *flat then collapses* — roughly constant and then dropping to 0.13 (5-eq high-λ subset) at $\lambda = 0.05$, the trivial-underfitting corner; it is not monotone in λ, and the per-seed traces are noisy. Figure 1 shows the recovery-and-complexity trace against λ, and Figure 2 shows the resulting accuracy–parsimony scatter. The two-dimensional view (Figures 3 and 4) shows that movement across the frontier varies primarily with λ, not with $\varepsilon_{\text{high}}$. This monotone $\lambda \to$ exploration/complexity trend holds across 3 seeds and two ε configurations and is our most reproducible result; because λ enters the reward directly, a mechanistic reading is reasonable in our grid, though we still report it as an association.

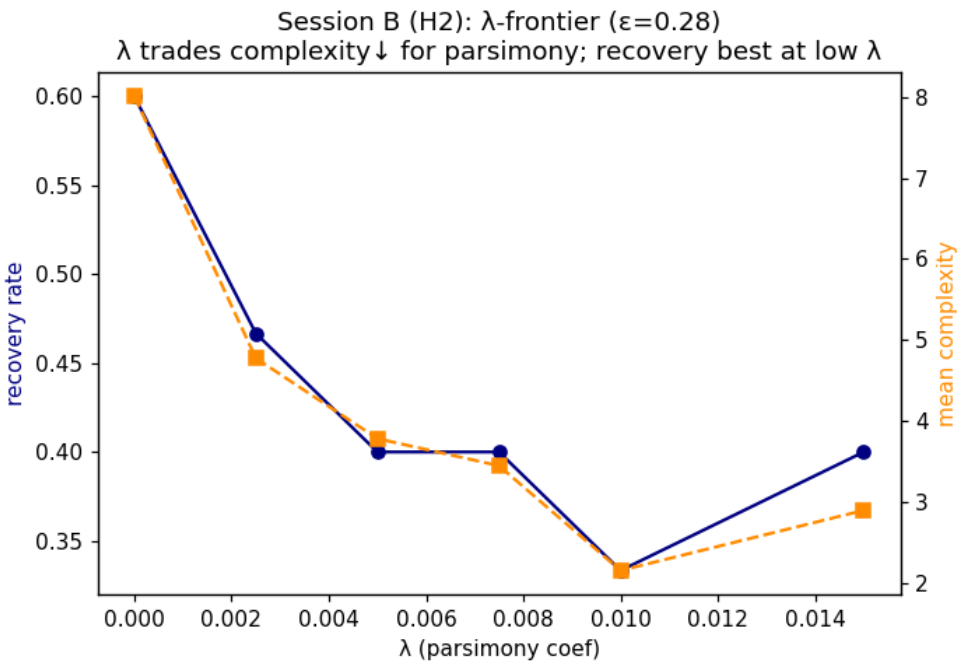

Figure 1: Recovery and mean complexity versus the parsimony coefficient λ. Raising λ monotonically simplifies expressions (complexity, on the 3-seed mean); recovery is flat then collapses to the trivial-underfitting corner (not monotone, with noisy per-seed traces).

H1: the $\varepsilon_{\text{high}}$ null (negative result). In the broad sweep, recovery is approximately flat (0.40 ± 0.03, where $\pm$ is the 3-seed std) across $\varepsilon_{\text{high}} \in [0.20, 0.40]$ at fixed λ. A dedicated re-test exposes the load-bearing role of the entropy bonus (held fixed elsewhere as an enabling regularizer): with the entropy bonus turned off and a wider grid $\varepsilon_{\text{high}} \in [0.20, 0.60]$, exploration collapses and stays low (roughly 5.6–7.3 unique structures) across the entire range, with recovery flat at 0.13 in this entropy-off setting (distinct from the $\lambda = 0.05$ collapse corner of the broad sweep above; Figure 5). With the entropy bonus off, exploration collapses regardless of $\varepsilon_{\text{high}}$ — it is the entropy regularizer, not the clip-higher knob, that sustains exploration. The clip findings in this paper are therefore conditional on entropy being held fixed.

H1 stress test: clip engaged, still null. A natural objection is that the clip was never active enough to matter (prior runs had a low clip-active fraction). We forced engagement (more inner epochs and a larger group), raising the mean clip-active fraction from 0.013 to 0.173 — which is more engagement but still mostly inactive (the clip is active on only ∼17% of tokens) — and re-ran at scale on harder Feynman equations *and* the non-separable synthetic stratum (3 seeds). This stress test changed several factors *together*: clip engagement (inner epochs $2 \to 6$), policy size (hidden $64 \to 128$), group size ($8 \to 12/16$), problem difficulty, and separability. Even so, the correlation between $\varepsilon_{\text{high}}$ and unique structures is weak on harder Feynman ($r = 0.21$) and only moderate but saturating on the synthetic stratum ($r = 0.68$); with $n = 5$ ε-levels neither correlation is statistically significant ($p \approx 0.2$), so we read both as weak and non-monotone, not as a degree effect. Recovery shows no ε trend (flat ∼0.33 on harder Feynman, ∼0 on synthetic); see Figure 6. The recovery-null survived this combined manipulation, but we did not isolate each factor, so we do

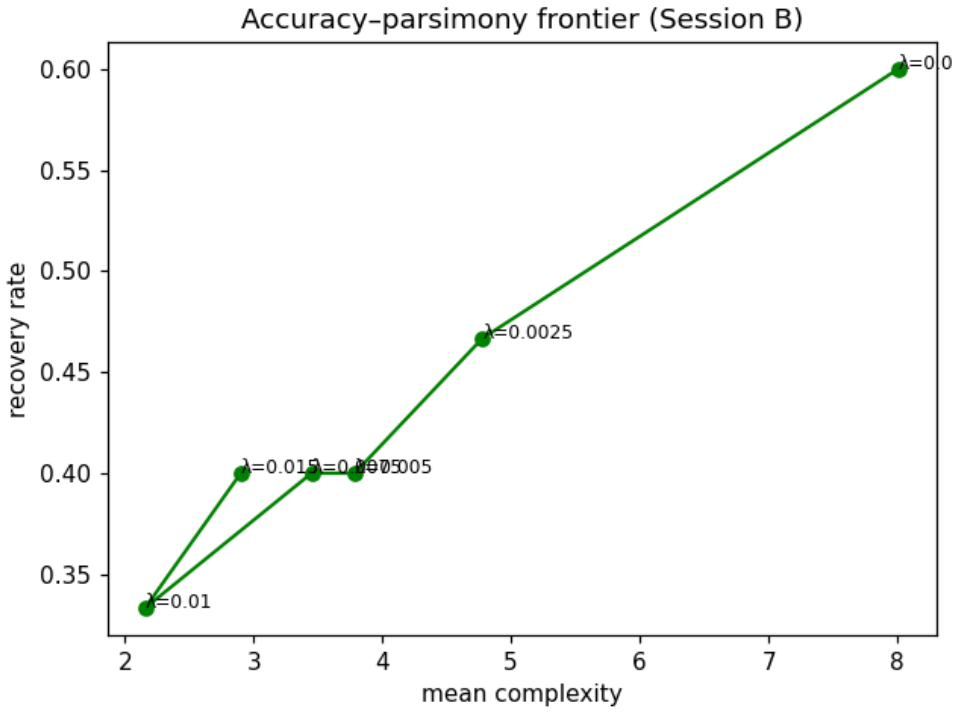

Figure 2: The accuracy–parsimony frontier traced by the λ sweep: each point is a configuration's (accuracy, complexity) operating point.

not claim independent robustness to scale, difficulty, or separability as separable factors.

H3: clip presence vs. degree. The two results reconcile as follows. Our `noclip` variant (our objective with the clip removed) collapses recovery to 0.000 and complexity to 2.08 (Table 1). But this collapse is specific to that variant: DSR — a clip-free objective — explores 171.7 structures and recovers 0.308, so clipping is *not* necessary for stable clip-free RL-SR in general. We therefore do *not* claim that clip presence is necessary; we soften H3 accordingly. The clean, defensible finding is that the clip *degree* ($\varepsilon_{\text{high}}$) is inert: varying it yields no recovery benefit. (We also note the clip-off collapse should be read off complexity 2.08 and recovery 0.0, not the unique-structure count 10.5, which roughly matches PySR's 10.2.)

Difficulty strata. Recovery breaks down sharply by difficulty; the strata are thin (a ∼13-equation × 3-seed pool), so these rates should be read as low-denominator estimates, not precise fractions. By variable count, 1/2/3-variable recovery is 0.33/0.56/0.07 (worst: 3 variables, combinatorial blow-up). By ground-truth complexity, $\leq$ 8/9–14/$\geq$ 15 recovery is 0.33/0.25/0.00 (every complexity-$\geq$ 15 equation fails). By operator family, polynomial-rational / exp-log / trigonometric recovery is 0.39/0.11/0.00 — trigonometric equations are never recovered (0/4), even simple ones.

Calibration and spurious fits. Held-out fit is a near-necessary gate for recovery but not sufficient. Among equations reaching $R^2 \geq 0.999$, recovery is 0.73; among those below 0.999 it is 0.00. However, 5 of the high-R^2 equations (27%) are spurious numeric fits — high accuracy, wrong structure. This

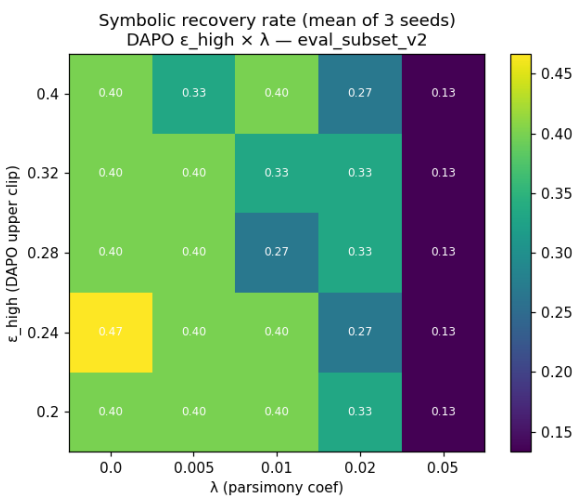

Figure 3: Recovery over the $\varepsilon_{\text{high}} \times \lambda$ grid: variation is organized by λ (vertical), not $\varepsilon_{\text{high}}$.

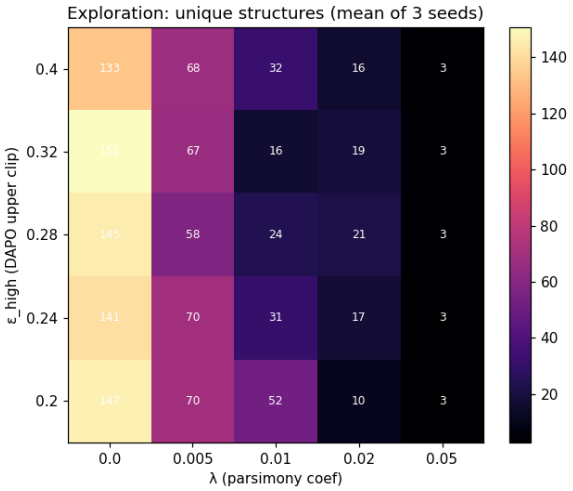

Figure 4: Unique structures over the same grid: again λ drives the gradient, with no clear $\varepsilon_{\text{high}}$ trend.

is precisely the accuracy-vs-symbolic gap (accuracy rate 0.282 vs. symbolic rate 0.103): the model is over-confident on structure, and $\lambda = 0$ (chosen to maximize raw recovery) removes the parsimony brake that would trade some recovery for fewer spurious fits.

6 Discussion and Limitations

Our central question is answered, with an honest split. On the 3-seed means the parsimony coefficient λ shows a statistically significant monotone trend in exploration (unique structures) and complexity, and varies primarily with where the policy lands on the accuracy–parsimony frontier (H2, positive; this is the relationship that meets our primary criterion of a significant monotone effect). The DAPO clip-higher asymmetry $\varepsilon_{\text{high}}$, by contrast, does not improve recovery and only weakly and non-monotonically modulates exploration ($r = 0.68$ on non-separable problems, $r = 0.21$ on harder Feynman equations; $n = 5$ ε-levels, neither correlation statistically significant, $p \approx 0.2$) — weak, non-monotone, and yielding no recovery benefit (H1, a hedged absence-of-recovery-effect, underpowered for a strong null); exploration is sustained primarily by λ together with the entropy regularizer, not by the clip ratio. The clip is not useless — our `noclip` variant collapses to trivial expressions — but a clip-free DSR objective is itself stable, so we do not claim clipping is necessary; the clean finding is that the clip *degree* does not modulate recovery. We read this as evidence that the low-probability-token argument developed on large-vocabulary language models [2] may not transfer cleanly to a grammar of roughly a dozen base symbols with no long noise tail.

On competitive recovery, our final DAPO model (0.205, 13-eq, 3-seed mean) is in the same range as the single-seed DSR run (0.308) and below the single-seed PySR run (0.615); we ran no baseline for multiple seeds and attach no significance to this ordering. The model has clear failure modes: trigonometric blindness (0/4 trig equations, the policy substitutes exp/polynomial forms), an expressivity/search ceiling (every complexity-≥ 15 equation fails), spurious numeric fits (5 cases of high R^2 without structural recovery), and combinatorial blow-up with variable count (0.07 at 3 variables vs. 0.56 at 2). These are concrete and characterized rather than mysterious, and

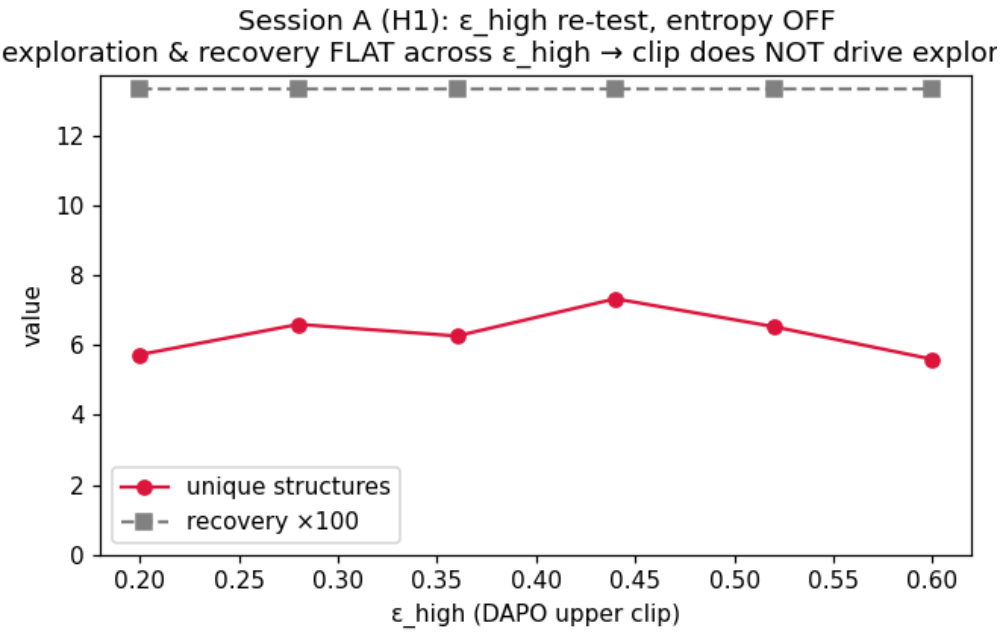

Figure 5: $\varepsilon_{\text{high}}$ re-test with the entropy bonus off: exploration and recovery are flat across the entire $\varepsilon_{\text{high}}$ range. The DAPO clip-higher knob is inert here.

they define the future-work agenda: trigonometric token priors, turning parsimony on to suppress spurious fits, and a larger policy/budget for high-complexity and high-arity equations.

6.1 Threats to Validity

Internal. Per-equation RL-SR is highly seed-sensitive: overall recovery is 0.31/0.23/0.08 across seeds (mean 0.205, std 0.096, ∼47% relative), so all headline numbers are 3-seed means with the spread and no single-seed claim is made. Spurious numeric fits (27% of high-R^2 cases) mean accuracy can outrun structure; we mitigate by reporting symbolic recovery separately from accuracy/numeric recovery. We also stress-tested the clip null against the most obvious confound (the clip being barely active) by forcing engagement; the recovery-null survived, but the stress test changed clip engagement, policy size, group size, difficulty, and separability *together*, so we did not isolate each factor and do not claim independent robustness to any one of them. The correlations underlying the clip result rest on $n = 5$ ε-levels and are not statistically significant ($p \approx 0.2$), so the clip null is necessarily underpowered for a strong claim.

External. The study uses a small single-layer GRU policy on a low-dimensional, noiseless Feynman subset, with a CPU-scale compute budget; it may not generalize to larger policies, higher-dimensional or noisy equations, or substantially larger search budgets. The synthetic non-separable stratum partially guards against Feynman's separability masking the mechanism, but it is small (8 equations). Cross-equation amortization (a single policy conditioned on the dataset and shared across equations) was out of scope here, so the claims rest on per-equation search.

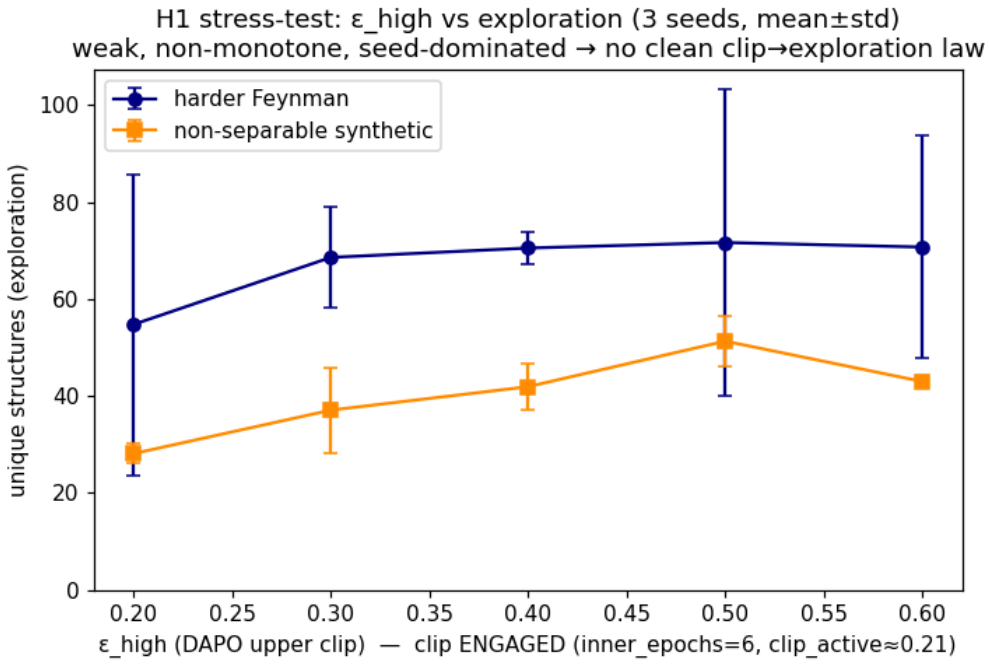

Figure 6: H1 stress test with the clip forced to engage on harder and non-separable problems: $\varepsilon_{\text{high}} \rightarrow$ exploration is weak/saturating and gives no recovery benefit.

Construct. Symbolic recovery is operationalized via `sympy` equivalence with a 10^{-4} dense-grid numeric fallback; the tolerance is set by BFGS constant-fit slack and could in principle admit a structurally-wrong-but-close expression, so symbolic and numeric recovery are reported separately. Exploration is measured by unique-structure count and complexity-distribution width rather than policy entropy, following the literature's caution that entropy is a misleading exploration proxy.

7 Availability

Artifacts are available from the authors on request.[1]

8 Conclusion and Future Work

We studied how an RL objective shapes the search in symbolic regression (with an entropy regularizer held fixed) and found a mechanistic split: on the 3-seed means the parsimony coefficient λ shows a statistically significant monotone trend in exploration and complexity and varies primarily with accuracy–parsimony frontier position, while the DAPO clip-higher asymmetry does not improve recovery — only weakly and non-monotonically modulating exploration (neither correlation significant; $n = 5$, $p \approx 0.2$) with no recovery benefit, even when the clip is forced to engage on harder, non-separable problems (a combined, non-isolated manipulation). Our `noclip` variant collapses, but a clip-free DSR objective is itself stable, so we do not claim clipping is necessary; the clean finding is that the clip degree does not modulate recovery. Recovery (0.205, 13-eq, 3-seed mean) lands in

[1]Artifacts (task code, configs, sweep grids, and run records) are available from the authors on request; nothing further is publicly hosted.

the same range as the single-seed DSR run and below the single-seed PySR run (no significance attached), with well-characterized failure modes. Ranked next steps: (1) trigonometric token priors to address the systematic 0/4 trig failure; (2) a small positive λ to suppress spurious numeric fits at a modest recovery cost; (3) a larger policy and search budget for high-complexity and high-arity equations; and (4) revisiting amortized cross-equation transfer once the per-equation mechanism is settled.

References

[1] Miles Cranmer. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. arXiv:2305.01582 [astro-ph.IM], 2023. URL https://arxiv.org/abs/2305.01582.

[2] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv:2503.14476 [cs.LG], 2025. URL https://arxiv.org/abs/2503.14476.

[3] Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. Low-probability Tokens Sustain Exploration in Reinforcement Learning with Verifiable Reward. arXiv:2510.03222 [cs.LG], 2025. URL https://arxiv.org/abs/2510.03222.

[4] Conor F. Hayes, Felipe Leno Da Silva, Jiachen Yang, T. Nathan Mundhenk, Chak Shing Lee, Jacob F. Pettit, Claudio Santiago, Sookyung Kim, Joanne T. Kim, Ignacio Aravena Solis, Ruben Glatt, Andre R. Goncalves, Alexander Ladd, Ahmet Can Solak, Thomas Desautels, Daniel Faissol, Brenden K. Petersen, and Mikel Landajuela. Deep Symbolic Optimization: Reinforcement Learning for Symbolic Mathematics. arXiv:2505.10762 [cs.LG], 2025. URL https://arxiv.org/abs/2505.10762.

[5] Jacob F. Pettit, Chak Shing Lee, Jiachen Yang, Alex Ho, Daniel Faissol, Brenden Petersen, and Mikel Landajuela. DisCo-DSO: Coupling Discrete and Continuous Optimization for Efficient Generative Design in Hybrid Spaces. arXiv:2412.11051 [cs.LG], 2024. URL https://arxiv.org/abs/2412.11051.

[6] Giorgio Morales and John W. Sheppard. Decomposable Neuro Symbolic Regression. arXiv:2511.04124 [cs.LG], 2025. URL https://arxiv.org/abs/2511.04124.

[7] Ali Soltani, Gabriel Kronberger, Fabricio Olivetti de Franca, Mattia Billa, and Alessandro Lucantonio. A Comparative Study of Model Selection Criteria for Symbolic Regression. arXiv:2605.11233 [cs.LG], 2026. URL https://arxiv.org/abs/2605.11233.

[8] Chenglu Sun, Shuo Shen, Wenzhi Tao, Deyi Xue, and Zixia Zhou. Noise-Resilient Symbolic Regression with Dynamic Gating Reinforcement Learning. arXiv:2501.01085 [cs.LG], 2025. URL https://arxiv.org/abs/2501.01085.

[9] Huimin Xu, Shuai Zhao, Xiaobao Wu, and Anh Tuan Luu. Understanding and Preventing Entropy Collapse in RLVR with On-Policy Entropy Flow Optimization. arXiv:2605.11491 [cs.LG], 2026. URL https://arxiv.org/abs/2605.11491.

[10] Chen Wang, Zhaochun Li, Jionghao Bai, Hexuan Deng, Ge Lan, and Yue Wang. SCOPE-RL: Stable and Quantitative Control of Policy Entropy in RL Post-Training. arXiv:2510.08141 [cs.LG], 2025. URL https://arxiv.org/abs/2510.08141.

[11] Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The Art of Scaling Reinforcement Learning Compute for LLMs. arXiv:2510.13786 [cs.LG], 2025. URL https://arxiv.org/abs/2510.13786.

[12] Yongsheng Lian. Comparative Analysis and Parametric Tuning of PPO, GRPO, and DAPO for LLM Reasoning Enhancement. arXiv:2512.07611 [cs.AI], 2025. URL https://arxiv.org/abs/2512.07611.

[13] Harry Desmond. (Exhaustive) Symbolic Regression and model selection by minimum description length. arXiv:2507.13033 [astro-ph.IM], 2025. URL https://arxiv.org/abs/2507.13033.

[14] Madhav R. Muthyala, Farshud Sorourifar, You Peng, and Joel A. Paulson. SyMANTIC: An Efficient Symbolic Regression Method for Interpretable and Parsimonious Model Discovery in Science and Beyond. arXiv:2502.03367 [cs.LG], 2025. URL https://arxiv.org/abs/2502.03367.

[15] Elizabeth S. Z. Tan, Adil Soubki, and Miles Cranmer. SymTorch: Symbolic Distillation of Neural Networks. arXiv:2602.21307 [cs.LG], 2026. URL https://arxiv.org/abs/2602.21307.

[16] Paul Kahlmeyer, Joachim Giesen, Michael Habeck, and Henrik Voigt. Scaling Up Unbiased Search-based Symbolic Regression. arXiv:2506.19626 [cs.LG], 2025. URL https://arxiv.org/abs/2506.19626.

[17] Michael Scherk and Boyuan Chen. SymMatika: Structure-Aware Symbolic Discovery. arXiv:2507.03110 [cs.LG], 2025. URL https://arxiv.org/abs/2507.03110.

[18] Henrik Voigt, Paul Kahlmeyer, Kai Lawonn, Michael Habeck, and Joachim Giesen. Analyzing Generalization in Pre-Trained Symbolic Regression. arXiv:2509.19849 [cs.LG], 2025. URL https://arxiv.org/abs/2509.19849.

[19] Silviu-Marian Udrescu and Max Tegmark. AI Feynman: a Physics-Inspired Method for Symbolic Regression. arXiv:1905.11481 [physics.comp-ph], 2019. URL https://arxiv.org/abs/1905.11481.

[20] William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabrício Olivetti de França, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason H. Moore. Contemporary Symbolic Regression Methods and their Relative Performance. arXiv:2107.14351 [cs.NE], 2021. URL https://arxiv.org/abs/2107.14351.

[21] Randal S. Olson, William La Cava, Patryk Orzechowski, Ryan J. Urbanowicz, and Jason H. Moore. PMLB: A Large Benchmark Suite for Machine Learning Evaluation and Comparison. arXiv:1703.00512 [cs.LG], 2017. URL https://arxiv.org/abs/1703.00512.