Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adaptive-Batch Trust-Region Semantic Genetic Programming for Distributionally Robust Symbolic Regression

Yutong Zhang
Sichuan University
Yaoran Yang
Sichuan University
Yifan Zhu
Sichuan University
Wentao Zhang
zhang-wt24@mails.tsinghua.edu.cn

Tsinghua University

Research Article

Keywords:

Posted Date: May 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9379299/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Adaptive-Batch Trust-Region Semantic Genetic Programming for Distributionally Robust Symbolic Regression

Yutong Zhang[1†], Yaoran Yang[1†], Yifan Zhu[2], Wentao Zhang[3*]

[1]School of Mathematics, Sichuan University, Chengdu, 610064, China.
[2]College of Computer Science, Sichuan University, Chengdu, 610064, China.
[3]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China.

*Corresponding author(s). E-mail(s): zhang-wt24@mails.tsinghua.edu.cn;
Contributing authors: yutongzhang@stu.scu.edu.cn; yangyaoran@stu.scu.edu.cn; zyfan@stu.scu.edu.cn;
†These authors contributed equally to this work.

Abstract

Symbolic regression has benefited substantially from recent advances in semantic search, neural guidance, and benchmark design. Two issues nevertheless remain insufficiently resolved in genetic programming (GP) for symbolic regression: uncertainty-aware batch sizing during evolution, and robustness-oriented model selection under environment shift. This paper presents *Adaptive-Batch Trust-Region Semantic GP* (ABTR-SGP), a symbolic-regression algorithm that couples three components: a stratified adaptive batch scheduler driven by elite-score uncertainty, a trust-region semantic interpolation operator that bounds per-generation behavioral drift, and a distributionally robust selection criterion built from a smooth worst-group surrogate and an explicit structural-complexity penalty. The resulting method is designed to stabilize search when training data contain environment-specific shortcuts or nuisance correlations.

The paper provides a main-text theoretical treatment. First, the batched robust score is shown to satisfy non-asymptotic concentration bounds under environment-stratified sampling. Second, trust-region semantic projection yields a direct bound on score drift on the monitored batch, and together with the concentration result implies a population-level robust-score stability guarantee.

Third, batched minimization over a finite candidate set is shown to incur at most a local $\mathbf{2}\varepsilon$ suboptimality penalty. These results explain why batch growth must scale with score ambiguity and why semantic step sizes should shrink as the ranking problem becomes sharper.
Empirically, ABTR-SGP is evaluated on a controlled four-task, three-environment symbolic-regression benchmark with appended spurious features and three out-of-distribution (OOD) test protocols. Across 16 paired runs per method, ABTR-SGP achieves the best mean robust OOD RMSE (0.764) and the best Friedman mean rank (2.250, $\boldsymbol{p = 0.0468}$). Relative to StdGP-Full and StdGP-Adaptive, it reduces mean robust OOD RMSE by 19.2% and 19.7%, with one-sided Wilcoxon $\boldsymbol{p = 0.0038}$ and $\boldsymbol{p = 0.0370}$, respectively. Relative to the stronger semantic baselines, ABTR-SGP remains accuracy-competitive while producing substantially smaller models and using fewer fitness-case evaluations. The overall picture is that batching, semantics, and robust validation should be designed jointly rather than in isolation.

1 Introduction

Symbolic regression (SR) occupies a distinctive position within interpretable machine learning because it does not merely fit parameters to a prescribed functional form; rather, it simultaneously searches for the functional form itself, its coefficients, and a parsimonious analytical representation of the underlying relationship [1–4]. This property makes SR particularly attractive in scientific discovery, system identification, and interpretable engineering design, where predictive accuracy alone is not sufficient and the recovered expression must also support inspection, validation, and downstream reasoning.

Recent years have seen substantial progress in the area. On the algorithmic side, neural-guided and hybrid SR systems have significantly expanded the feasible search regime [5–8]. On the evaluation side, benchmark culture has matured from small collections of hand-picked equations to broader and more principled suites emphasizing reproducibility, accuracy–complexity trade-offs, and hard discovery scenarios [4, 9, 10]. In parallel, semantic GP has evolved from the original geometric semantic formulation [11] to practical implementations and refinements involving efficient software support, local search, normalization, cellular topologies, ensemble interpretations, and linear-scaling variants [12–22]. The *Genetic Programming and Evolvable Machines* community has likewise widened its scope toward large-language-model guidance and interpretable control synthesis [24, 25].

Despite this progress, two methodological tensions remain underdeveloped.

The first tension is computational. Full-batch GP remains expensive because every individual is scored on the entire training set. Mini-batching is therefore appealing, but unlike in gradient-based learning, naive batching in GP directly perturbs the induced ranking over individuals. In other words, batching does not merely inject stochastic gradient noise; it can invert the elite order that drives reproduction and survivor selection. The recent survey of Rosenfeld and Vanneschi identifies dynamic

and adaptive batch sizing as a particularly underexplored direction in GP [23]. Yet adaptive batching is only useful if it is tied to rank uncertainty rather than used as a generic computational shortcut.

The second tension is statistical. Semantic operators make GP search smoother in behavior space, but semantic smoothness by itself does not imply robustness. A semantic step that is attractive on a small or unrepresentative batch can still amplify environment-specific shortcuts and degrade performance under shift. This issue is especially acute when the data are heterogeneous across environments or when shortcut features correlate with the target on the training mixture but break under OOD perturbations. In such settings, selecting a final model solely by average validation error is often insufficient; the selection rule itself should express a preference for stable behavior across environments.

This paper addresses precisely the intersection of these two issues. We study environment-aware symbolic regression in which search is conducted with adaptive stratified batches, semantic interpolation is explicitly trust-region constrained, and the final model is selected by a distributionally robust criterion that penalizes both worst-group error and unnecessary structural complexity.

Contributions.

1. We propose *Adaptive-Batch Trust-Region Semantic GP* (ABTR-SGP), a symbolic-regression algorithm that jointly combines stratified adaptive batching, trust-region semantic interpolation, and distributionally robust validation selection.
2. We provide a main-text theoretical treatment: a non-asymptotic concentration bound for the batched robust score, a direct stability bound for trust-region semantic updates, and a local suboptimality guarantee for batched selection over a finite candidate set.
3. We introduce a controlled multi-environment SR benchmark with appended spurious features and three OOD protocols, designed to isolate robustness from mere grammar coverage.
4. We show empirically that ABTR-SGP attains the best mean robust OOD performance among the tested methods while substantially reducing semantic-GP model size and evaluation cost.

2 Related work

2.1 Symbolic regression systems and benchmark culture

Modern SR spans classical evolutionary search, neural proposal models, and hybrid systems. Foundational GP-based SR remains attractive because it directly exposes symbolic structure [1], while physically motivated equation-discovery systems have demonstrated the value of structural priors and analytical recovery in scientific domains [2, 3]. More recently, large-scale pretraining and neural guidance have improved raw search efficiency and recovery rates [5–8]. At the same time, the field has become increasingly benchmark-driven. The SRBench effort emphasized common evaluation protocols and broad reproducibility [4]; subsequent work sharpened

the evaluation lens further by introducing absolute Pareto fronts for accuracy–complexity benchmarking [9], multi-level SR settings [26], and more demanding scientific equation-discovery suites for large language models [10]. These developments make robustness-oriented model selection more important, not less, because average in-distribution accuracy alone is no longer a sufficient discriminator among methods.

2.2 Semantic GP and semantic search operators

Semantic GP arose from the observation that program behavior, rather than syntax alone, can be a more meaningful search space for supervised learning. The geometric semantic formulation of Moraglio et al. showed that suitably designed operators induce a unimodal error landscape in function space [11]. Subsequent work made this family more practical and expressive, including efficient implementations [12], local-search integration [13], normalized and standardized random semantic programs [14], cellular semantic topologies [15, 21], ensemble-theoretic interpretations [16], linear-scaling variants [17], macro-semantic operators for regression-oriented feature construction [18], semantic boosting [19], and constraint-aware semantic backpropagation [20]. The present paper differs from these lines in a specific way: semantics are treated not merely as a direction of descent, but as a *trust-region object* whose magnitude must be coupled to batch uncertainty.

2.3 Batching and stochastic evaluation in GP

Batch training has long been implicit in parts of evolutionary computation, but only recently has the GP literature begun to organize the topic systematically. Rosenfeld and Vanneschi survey partial-data fitness estimation across GP and identify adaptive batch-size selection as a particularly open problem [23]. The issue is fundamentally sharper in GP than in gradient-based optimization because fitness estimates determine an induced order over individuals rather than a single noisy gradient direction. That distinction motivates rank-sensitive batching rules instead of static mini-batch schedules. ABTR-SGP adopts exactly this perspective by making batch growth depend on the observed gap between leading candidates.

2.4 OOD robustness and model selection across environments

The broader machine-learning literature offers several complementary views on robustness under multiple environments, including invariant-prediction formulations [28–30], extrapolation-based penalties [31], group distributional robustness [32, 34], and systematic evaluations of domain-generalization algorithms [35]. A common lesson across these works is that average empirical risk can hide severe subgroup or environment failures. In symbolic regression, analogous concerns arise when shortcut variables or nuisance structure differ across environments. Recent GP work on high-dimensional SR has also emphasized the role of generalization-aware feature selection [27]. ABTR-SGP draws from this broader robustness perspective but remains fully within the GP/SR paradigm: rather than enforcing exact invariance constraints, it uses a smooth worst-group surrogate as a robust selection objective and integrates that objective with batched semantic search.

3 Method

3.1 Problem setting and robust objective

We consider symbolic regression with E training environments. For each environment $e \in \{1, \dots, E\}$, let

$$\mathcal{D}_e^{\mathrm{tr}} = \{(x_{e,i}, y_{e,i})\}_{i=1}^{N_e}$$

denote the evolution split used during search, and let $\mathcal{D}_e^{\mathrm{val}}$ denote the corresponding validation split used only for final model selection. The search space is the common arithmetic–trigonometric grammar shared by all methods in the comparison.

To obtain bounded finite-sample guarantees, we use clipped squared loss

$$\ell_M(\hat{y}, y) = \min\left\{(\hat{y} - y)^2, M\right\},$$

where $M > 0$ is a fixed clipping constant. For a symbolic model f, the environment risk on a generic split $S_e \subseteq \mathcal{D}_e$ is

$$R_e^S(f) = \frac{1}{|S_e|} \sum_{(x,y) \in S_e} \ell_M(f(x), y).$$

When the split is clear from context, we simply write $R_e(f)$.

The robust score used throughout the paper is

$$J_\lambda(f) = \bar{R}(f) + \lambda\, \phi_\beta\big(R_1(f), \dots, R_E(f)\big) + \eta \log\big(1 + \Omega(f)\big), \tag{1}$$

where

$$\bar{R}(f) = \frac{1}{E} \sum_{e=1}^{E} R_e(f), \qquad \phi_\beta(\mathbf{u}) = \frac{1}{\beta} \log\left(\frac{1}{E} \sum_{e=1}^{E} e^{\beta u_e}\right),$$

$\Omega(f)$ is the node count of f, $\lambda \geq 0$ is the robustness weight, $\beta > 0$ controls the smoothness of the worst-group surrogate, and $\eta \geq 0$ regulates structural complexity. During evolution we use the training version $J_{\lambda_{\mathrm{tr}}}$; final model selection uses the same functional form on the validation splits, potentially with a stronger robustness weight $\lambda_{\mathrm{val}} \geq \lambda_{\mathrm{tr}}$. To avoid notational clutter, the theory below is written for a generic λ.

Remark 1 For every $\mathbf{u} \in \mathbb{R}^E$,

$$\max_{1 \leq e \leq E} u_e - \frac{\log E}{\beta} \leq \phi_\beta(\mathbf{u}) \leq \max_{1 \leq e \leq E} u_e.$$

Indeed,

$$\max_e e^{\beta u_e} \leq \sum_{e=1}^{E} e^{\beta u_e} \leq E \max_e e^{\beta u_e},$$

and taking $\frac{1}{\beta} \log(\cdot / E)$ yields the claim. Thus β directly controls the approximation error of the soft worst-group term.

3.2 Adaptive stratified batch evaluation

At generation t, ABTR-SGP does not evaluate the population on the full evolution data. Instead, it draws a stratified batch

$$B_t = \bigcup_{e=1}^{E} B_{t,e},$$

where each $B_{t,e} \subseteq \mathcal{D}_e^{\mathrm{tr}}$ is sampled uniformly without replacement and the per-environment quotas are balanced as evenly as possible:

$$m_{t,e} \in \left\{ \left\lfloor \frac{m_t}{E} \right\rfloor, \left\lceil \frac{m_t}{E} \right\rceil \right\}, \qquad \sum_{e=1}^{E} m_{t,e} = m_t.$$

The batched environment risk is

$$\hat{R}_{t,e}(f) = \frac{1}{m_{t,e}} \sum_{(x,y)\in B_{t,e}} \ell_M(f(x), y),$$

and the batched robust score is

$$\hat{J}_t(f) = \frac{1}{E} \sum_{e=1}^{E} \hat{R}_{t,e}(f) + \lambda\, \phi_\beta\big(\hat{R}_{t,1}(f), \ldots, \hat{R}_{t,E}(f)\big) + \eta \log\big(1 + \Omega(f)\big). \tag{2}$$

The key question is how to choose m_t. Let $f_{(1),t}$ and $f_{(2),t}$ denote the first- and second-ranked individuals under $\hat{J}_t$, and define the observed elite gap

$$\hat{\Delta}_t = \hat{J}_t\big(f_{(2),t}\big) - \hat{J}_t\big(f_{(1),t}\big) \geq 0.$$

Let

$$\hat{v}_{t,i} = \frac{1}{E} \sum_{e=1}^{E} \widehat{\mathrm{Var}}_{(x,y)\in B_{t,e}} \left[\ell_M\big(f_{(i),t}(x), y\big) \right], \qquad i \in \{1, 2\},$$

be the mean empirical environment-loss variance of the top two individuals. We then update the total batch size via

$$m_{t+1} = \mathrm{clip}\left(\left\lceil \frac{c\big(\hat{v}_{t,1} + \hat{v}_{t,2}\big)}{\big(\hat{\Delta}_t + \varepsilon\big)^2} \right\rceil, m_{\min}, m_{\max} \right), \tag{3}$$

where $c > 0$ is a tuning constant, $\varepsilon > 0$ prevents numerical blow-up when the observed gap is extremely small, and $\mathrm{clip}(a, \ell, u) = \min\{u, \max\{\ell, a\}\}$. Intuitively, generations with a clear elite ordering can be processed on smaller batches, while ambiguous generations force larger batches.

3.3 Trust-region semantic interpolation

Given a parent f and a raw syntax-generated proposal $\tilde{f}$ obtained by subtree crossover or mutation, ABTR-SGP measures the semantic displacement on the current batch by

$$d_t(f, \tilde{f}) = \max_{1 \le e \le E} \left(\frac{1}{|B_{t,e}|} \sum_{x \in B_{t,e}} \big(\tilde{f}(x) - f(x)\big)^2 \right)^{1/2}. \tag{4}$$

Instead of accepting $\tilde{f}$ directly, the method projects the semantic step into a trust region of radius τ_t:

$$\alpha_t = \min\left(1, \frac{\tau_t}{d_t(f, \tilde{f}) + \varepsilon}\right), \qquad f^+ = (1 - \alpha_t) f + \alpha_t \tilde{f}. \tag{5}$$

The trust-region radius is coupled to the batch size:

$$\tau_t = \max\left(\tau_{\min}, \tau_{\max} \sqrt{\frac{m_{\min}}{m_t}}\right). \tag{6}$$

Hence ABTR-SGP behaves aggressively when the ranking is still coarse and cheap to estimate, but becomes increasingly conservative as the population is evaluated on larger batches. This coupling is deliberate: the method uses exploratory semantic motion when score uncertainty is high and contracts the semantic radius as selection becomes sharper.

3.4 Validation-time model selection

The algorithm explicitly separates *search-time scoring* from *final model choice*. Parent selection, offspring comparison, and survivor selection are all driven by the batched training objective (2). By contrast, the final model is chosen from the elite archive $\mathcal{H}$ using the full validation splits:

$$f^\star = \arg\min_{f \in \mathcal{H}} J^{\text{val}}_{\lambda_{\text{val}}}(f).$$

This design is important. Search-time batching reduces cost but introduces estimation noise; final selection therefore uses the full validation environments and a stronger robustness emphasis so that a candidate is not chosen solely because it looked favorable on a sequence of noisy mini-batches.

4 Theoretical analysis

4.1 Standing assumptions and scope of the probabilistic claims

All results in this section are stated at a *fixed generation* t. To separate the within-generation sampling question from the long-range adaptivity of the evolutionary

process, let $\mathcal{F}_t$ denote the σ-field generated by all randomness revealed *before* the evaluation batch at generation t is sampled. Thus the current population, the elite archive accumulated so far, and any candidate family determined prior to evaluation are $\mathcal{F}_t$-measurable. Every probability statement below is conditional on $\mathcal{F}_t$; for readability we suppress this conditioning from the notation.

We use the following standing assumptions.

(A1)

For each environment e, the batch $B_{t,e}$ is sampled uniformly without replacement from $\mathcal{D}_e^{\mathrm{tr}}$.

(A2)

The clipped loss satisfies

$$0 \le \ell_M(\hat{y}, y) \le M \qquad \text{for all } (\hat{y}, y).$$

(A3)

Whenever a high-probability statement is made for a model f or for a finite family $\mathcal{C}_t$, that model or family is $\mathcal{F}_t$-measurable. Equivalently, the compared candidates are fixed before the evaluation batch is drawn.

(A4)

Whenever a finite candidate family is invoked, it has cardinality

$$K_t = |\mathcal{C}_t| < \infty.$$

We also define

$$m_{\min,t} = \min_{1 \le e \le E} m_{t,e}.$$

Assumption (A3) is essential. Without it, concentration for a *data-dependent* candidate chosen from the same batch used to evaluate it generally fails because selection on the batch can introduce optimistic bias. Deterministic statements such as the semantic trust-region bound below do *not* require (A3); only the concentration-based transfer from batch quantities to population quantities does.

The next lemma records the two elementary properties of the clipped squared loss used repeatedly in the sequel.

Lemma 1 (Basic properties of the clipped squared loss) *For fixed $y \in \mathbb{R}$, the map*

$$\hat{y} \mapsto \ell_M(\hat{y}, y) = \min\{(\hat{y} - y)^2, M\}$$

takes values in $[0, M]$ and is globally $2\sqrt{M}$-Lipschitz in $\hat{y}$. That is,

$$\left|\ell_M(\hat{y}_1, y) - \ell_M(\hat{y}_2, y)\right| \le 2\sqrt{M}\,|\hat{y}_1 - \hat{y}_2| \qquad \textit{for all } \hat{y}_1, \hat{y}_2 \in \mathbb{R}.$$

Proof Boundedness is immediate from the definition. For the Lipschitz claim, define

$$g(z) = \min\{z^2, M\}, \qquad z \in \mathbb{R}.$$

Then $\ell_M(\hat{y}, y) = g(\hat{y} - y)$. The function g is differentiable on $\mathbb{R} \setminus \{\pm\sqrt{M}\}$, with

$$g'(z) = \begin{cases} 2z, & |z| < \sqrt{M}, \\ 0, & |z| > \sqrt{M}. \end{cases}$$

Hence $|g'(z)| \le 2\sqrt{M}$ wherever the derivative exists. Since g is continuous and piecewise C^1, the mean-value theorem applied on each interval between the breakpoints implies

$$|g(z_1) - g(z_2)| \le 2\sqrt{M}\,|z_1 - z_2| \qquad \text{for all } z_1, z_2 \in \mathbb{R}.$$

Substituting $z_i = \hat{y}_i - y$ yields the claim. □

4.2 Lipschitz transfer from environment risks to the robust score

The first step is to show that the robust objective is stable under coordinatewise perturbations of the environment-risk vector.

Proposition 1 (Lipschitz transfer) *Let*

$$G_\lambda(\mathbf{u}) = \frac{1}{E}\sum_{e=1}^{E} u_e + \lambda\,\phi_\beta(\mathbf{u}), \qquad \mathbf{u} \in \mathbb{R}^E,$$

where

$$\phi_\beta(\mathbf{u}) = \frac{1}{\beta}\log\left(\frac{1}{E}\sum_{e=1}^{E} e^{\beta u_e}\right).$$

Then, for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^E$,

$$|G_\lambda(\mathbf{u}) - G_\lambda(\mathbf{v})| \le (1+\lambda)\|\mathbf{u} - \mathbf{v}\|_\infty.$$

Consequently, if

$$J_\lambda(f) = G_\lambda(\mathbf{R}(f)) + \eta\log(1 + \Omega(f)),$$

then for all symbolic models f, g,

$$|J_\lambda(f) - J_\lambda(g)| \le (1+\lambda)\|\mathbf{R}(f) - \mathbf{R}(g)\|_\infty + \eta\,|\log(1+\Omega(f)) - \log(1+\Omega(g))|\,.$$

Proof For the average-risk term,

$$\left|\frac{1}{E}\sum_{e=1}^{E} u_e - \frac{1}{E}\sum_{e=1}^{E} v_e\right| \le \frac{1}{E}\sum_{e=1}^{E} |u_e - v_e| \le \|\mathbf{u} - \mathbf{v}\|_\infty.$$

For the smooth worst-group term, write

$$S(\mathbf{u}) = \frac{1}{E}\sum_{e=1}^{E} e^{\beta u_e}, \qquad \phi_\beta(\mathbf{u}) = \frac{1}{\beta}\log S(\mathbf{u}).$$

A direct differentiation gives

$$\frac{\partial \phi_\beta}{\partial u_e}(\mathbf{u}) = \frac{e^{\beta u_e}}{\sum_{j=1}^{E} e^{\beta u_j}} =: p_e(\mathbf{u}),$$

where $p_e(\mathbf{u}) \geq 0$ and $\sum_{e=1}^{E} p_e(\mathbf{u}) = 1$. Therefore

$$\|\nabla \phi_\beta(\mathbf{u})\|_1 = \sum_{e=1}^{E} p_e(\mathbf{u}) = 1.$$

Now define the line segment

$$\gamma(s) = \mathbf{v} + s(\mathbf{u} - \mathbf{v}), \qquad s \in [0, 1].$$

By the fundamental theorem of calculus,

$$\phi_\beta(\mathbf{u}) - \phi_\beta(\mathbf{v}) = \int_0^1 \nabla \phi_\beta(\gamma(s))^\top (\mathbf{u} - \mathbf{v})\, ds.$$

Hence

$$\begin{aligned} |\phi_\beta(\mathbf{u}) - \phi_\beta(\mathbf{v})| &\leq \int_0^1 \|\nabla \phi_\beta(\gamma(s))\|_1 \|\mathbf{u} - \mathbf{v}\|_\infty \, ds \\ &\leq \int_0^1 1 \cdot \|\mathbf{u} - \mathbf{v}\|_\infty \, ds = \|\mathbf{u} - \mathbf{v}\|_\infty. \end{aligned}$$

Combining the average-risk and smooth-max bounds yields

$$|G_\lambda(\mathbf{u}) - G_\lambda(\mathbf{v})| \leq (1 + \lambda)\|\mathbf{u} - \mathbf{v}\|_\infty.$$

Finally, substituting $\mathbf{u} = \mathbf{R}(f)$ and $\mathbf{v} = \mathbf{R}(g)$ gives

$$|J_\lambda(f) - J_\lambda(g)| \leq |G_\lambda(\mathbf{R}(f)) - G_\lambda(\mathbf{R}(g))| + \eta \left|\log(1 + \Omega(f)) - \log(1 + \Omega(g))\right|,$$

and the claimed inequality follows from the first part. □

4.3 Concentration of the batched robust score

We next control the discrepancy between the batched robust score and the full evolution-split score.

Theorem 1 (Concentration of the batched robust score) *Fix a generation t and let f be an $\mathcal{F}_t$-measurable symbolic model. Under* (A1)–(A3)*, for any $\delta \in (0, 1)$,*

$$\Pr\left(\left| \hat{J}_t(f) - J_\lambda(f) \right| > (1 + \lambda) M \sqrt{\frac{\log(2E/\delta)}{2m_{\min,t}}} \right) \leq \delta.$$

Proof Fix an environment $e \in \{1, \ldots, E\}$. Conditional on $\mathcal{F}_t$, the model f is fixed, so the finite population

$$\left\{ \ell_M(f(x_{e,1}), y_{e,1}), \ldots, \ell_M(f(x_{e,N_e}), y_{e,N_e}) \right\}$$

is deterministic and takes values in $[0, M]$ by Lemma 1. The batch mean $\hat{R}_{t,e}(f)$ is the average of $m_{t,e}$ elements sampled uniformly without replacement from this population, while $R_e(f)$ is the full-population mean. Hoeffding's inequality for sampling without replacement (equivalently, the basic Serfling bound) therefore gives, for every $\epsilon > 0$,

$$\Pr\left(\left| \hat{R}_{t,e}(f) - R_e(f) \right| > \epsilon \right) \leq 2 \exp\left(-\frac{2m_{t,e}\epsilon^2}{M^2} \right) \leq 2 \exp\left(-\frac{2m_{\min,t}\epsilon^2}{M^2} \right).$$

Applying the union bound over $e = 1, \dots, E$ yields

$$\Pr\left(\max_{1\le e\le E}\left|\hat{R}_{t,e}(f) - R_e(f)\right| > \epsilon\right) \le 2E\exp\left(-\frac{2m_{\min,t}\epsilon^2}{M^2}\right).$$

Choose

$$\epsilon = M\sqrt{\frac{\log(2E/\delta)}{2m_{\min,t}}}.$$

Then, with probability at least $1-\delta$,

$$\left\|\hat{\mathbf{R}}_t(f) - \mathbf{R}(f)\right\|_\infty \le M\sqrt{\frac{\log(2E/\delta)}{2m_{\min,t}}},$$

where

$$\hat{\mathbf{R}}_t(f) = \big(\hat{R}_{t,1}(f), \dots, \hat{R}_{t,E}(f)\big), \qquad \mathbf{R}(f) = \big(R_1(f), \dots, R_E(f)\big).$$

Now observe that

$$\hat{J}_t(f) = G_\lambda\big(\hat{\mathbf{R}}_t(f)\big) + \eta\log(1+\Omega(f)), \qquad J_\lambda(f) = G_\lambda\big(\mathbf{R}(f)\big) + \eta\log(1+\Omega(f)).$$

The complexity term is identical in both quantities, so Proposition 1 implies

$$\left|\hat{J}_t(f) - J_\lambda(f)\right| = \left|G_\lambda\big(\hat{\mathbf{R}}_t(f)\big) - G_\lambda\big(\mathbf{R}(f)\big)\right| \le (1+\lambda)\left\|\hat{\mathbf{R}}_t(f) - \mathbf{R}(f)\right\|_\infty.$$

Substituting the previous deviation bound proves the claim. □

Remark 2 A sharper version of Theorem 1 can incorporate the finite-population correction from Serfling's inequality. We display the simpler conservative form because it isolates the dominant $m_{\min,t}^{-1/2}$ dependence that drives the batch-growth logic.

Corollary 1 (Uniform deviation over a finite candidate set) *Let $\mathcal{C}_t$ be a finite $\mathcal{F}_t$-measurable candidate set with $|\mathcal{C}_t| = K_t$. Under the assumptions of Theorem 1, for any $\delta \in (0,1)$,*

$$\Pr\left(\sup_{g\in\mathcal{C}_t}\left|\hat{J}_t(g) - J_\lambda(g)\right| > (1+\lambda)M\sqrt{\frac{\log(2EK_t/\delta)}{2m_{\min,t}}}\right) \le \delta.$$

Proof For each fixed $g \in \mathcal{C}_t$, Theorem 1 gives

$$\Pr\left(\left|\hat{J}_t(g) - J_\lambda(g)\right| > (1+\lambda)M\sqrt{\frac{\log(2EK_t/\delta)}{2m_{\min,t}}}\right) \le \frac{\delta}{K_t}.$$

Since $\mathcal{C}_t$ is $\mathcal{F}_t$-measurable and finite, a union bound over the K_t candidates yields

$$\Pr\left(\exists g \in \mathcal{C}_t : \left|\hat{J}_t(g) - J_\lambda(g)\right| > (1+\lambda)M\sqrt{\frac{\log(2EK_t/\delta)}{2m_{\min,t}}}\right) \le K_t \cdot \frac{\delta}{K_t} = \delta,$$

which is exactly the stated bound. □

Corollary 2 (Pairwise ranking preservation and gap-squared batch scaling) *Let $\mathcal{G}_t$ be any finite $\mathcal{F}_t$-measurable comparison family, and let $g^\star_{1,t}, g^\star_{2,t} \in \mathcal{G}_t$ denote the first- and second-ranked elements under the full score J_λ:*

$$J_\lambda(g^\star_{1,t}) \le J_\lambda(g^\star_{2,t}) \le J_\lambda(g) \qquad \textit{for all } g \in \mathcal{G}_t.$$

Define the true top-two gap

$$\Delta_t^\star = J_\lambda(g_{2,t}^\star) - J_\lambda(g_{1,t}^\star) > 0.$$

If

$$\Delta_t^\star > 2(1+\lambda)M\sqrt{\frac{\log(4E/\delta)}{2m_{\min,t}}},$$

then

$$\hat{J}_t(g_{1,t}^\star) < \hat{J}_t(g_{2,t}^\star)$$

with probability at least $1-\delta$*. Equivalently, it is sufficient that*

$$m_{\min,t} \geq \frac{2(1+\lambda)^2 M^2 \log(4E/\delta)}{(\Delta_t^\star)^2}.$$

Proof Apply Theorem 1 to $g_{1,t}^\star$ and $g_{2,t}^\star$ separately with confidence level $\delta/2$. Then with probability at least $1-\delta$, both inequalities

$$\left|\hat{J}_t(g_{1,t}^\star) - J_\lambda(g_{1,t}^\star)\right| \leq \epsilon, \qquad \left|\hat{J}_t(g_{2,t}^\star) - J_\lambda(g_{2,t}^\star)\right| \leq \epsilon$$

hold simultaneously, where

$$\epsilon = (1+\lambda)M\sqrt{\frac{\log(4E/\delta)}{2m_{\min,t}}}.$$

On this event,

$$\begin{aligned}\hat{J}_t(g_{2,t}^\star) - \hat{J}_t(g_{1,t}^\star) &\geq \big(J_\lambda(g_{2,t}^\star) - \epsilon\big) - \big(J_\lambda(g_{1,t}^\star) + \epsilon\big) \\ &= \Delta_t^\star - 2\epsilon.\end{aligned}$$

Therefore, if $\Delta_t^\star > 2\epsilon$, then

$$\hat{J}_t(g_{2,t}^\star) - \hat{J}_t(g_{1,t}^\star) > 0,$$

which is equivalent to

$$\hat{J}_t(g_{1,t}^\star) < \hat{J}_t(g_{2,t}^\star).$$

Solving the inequality $2\epsilon < \Delta_t^\star$ for $m_{\min,t}$ yields the stated sufficient condition. □

Remark 3 (Why the scheduler scales like variance divided by gap squared) Corollary 2 already shows the inverse-gap-squared law. The empirical scheduler in (3) adds variance information, and this is also mathematically natural. Indeed, for two fixed models f and g, define the environment-wise loss difference

$$Z_e(x,y) = \ell_M(g(x),y) - \ell_M(f(x),y).$$

Then

$$\mathrm{Var}(Z_e) = \mathrm{Var}\big((\ell_M(g) - \mathbb{E}\ell_M(g)) - (\ell_M(f) - \mathbb{E}\ell_M(f))\big) \leq 2\,\mathrm{Var}(\ell_M(g)) + 2\,\mathrm{Var}(\ell_M(f)),$$

because $(a-b)^2 \leq 2a^2 + 2b^2$. Hence any variance-sensitive concentration inequality for comparing the score gap between f and g necessarily involves a numerator proportional to the sum of their loss variances, and a denominator proportional to the square of the score gap. This is the formal reason that (3) uses the structure

$$m_{t+1} \propto \frac{\hat{v}_{t,1} + \hat{v}_{t,2}}{(\hat{\Delta}_t + \varepsilon)^2}.$$

4.4 Trust-region semantic stability

We now show that the trust-region operator directly controls the score drift induced by semantic interpolation.

Proposition 2 (Stability on the monitored batch) *Let* $f^+ = (1-\alpha_t)f + \alpha_t \tilde{f}$ *be produced by* (5). *Then*

$$d_t(f, f^+) \le \tau_t,$$

and

$$\left|\hat{J}_t(f^+) - \hat{J}_t(f)\right| \le (1+\lambda) L_\ell \tau_t + \eta\, \Delta_\Omega(f^+, f),$$

where

$$L_\ell := 2\sqrt{M}, \qquad \Delta_\Omega(f^+, f) = \left|\log(1+\Omega(f^+)) - \log(1+\Omega(f))\right|.$$

Proof By definition,

$$f^+(x) - f(x) = \alpha_t\big(\tilde{f}(x) - f(x)\big).$$

Therefore, for each environment e,

$$\left(\frac{1}{|B_{t,e}|}\sum_{x\in B_{t,e}} \big(f^+(x) - f(x)\big)^2\right)^{1/2} = \alpha_t \left(\frac{1}{|B_{t,e}|}\sum_{x\in B_{t,e}} \big(\tilde{f}(x) - f(x)\big)^2\right)^{1/2}.$$

Taking the maximum over e gives

$$d_t(f, f^+) = \alpha_t\, d_t(f, \tilde{f}).$$

Since

$$\alpha_t = \min\left(1, \frac{\tau_t}{d_t(f,\tilde{f}) + \varepsilon}\right),$$

we have $\alpha_t\, d_t(f, \tilde{f}) \le \tau_t$, hence

$$d_t(f, f^+) \le \tau_t.$$

Next fix an environment e. Using Lemma 1 with $L_\ell = 2\sqrt{M}$, for every $(x, y) \in B_{t,e}$,

$$\left|\ell_M(f^+(x), y) - \ell_M(f(x), y)\right| \le L_\ell\, |f^+(x) - f(x)|.$$

Averaging over the batch and applying the triangle inequality,

$$\begin{aligned}
\left|\hat{R}_{t,e}(f^+) - \hat{R}_{t,e}(f)\right| &= \left|\frac{1}{|B_{t,e}|}\sum_{(x,y)\in B_{t,e}} \Big(\ell_M(f^+(x), y) - \ell_M(f(x), y)\Big)\right| \\
&\le \frac{1}{|B_{t,e}|}\sum_{(x,y)\in B_{t,e}} \left|\ell_M(f^+(x), y) - \ell_M(f(x), y)\right| \\
&\le \frac{L_\ell}{|B_{t,e}|}\sum_{(x,y)\in B_{t,e}} |f^+(x) - f(x)|.
\end{aligned}$$

Now apply Cauchy–Schwarz:

$$\frac{1}{|B_{t,e}|}\sum_{(x,y)\in B_{t,e}} |f^+(x) - f(x)| \le \left(\frac{1}{|B_{t,e}|}\sum_{x\in B_{t,e}} \big(f^+(x) - f(x)\big)^2\right)^{1/2} \le d_t(f, f^+) \le \tau_t.$$

Therefore

$$\left|\hat{R}_{t,e}(f^+) - \hat{R}_{t,e}(f)\right| \le L_\ell \tau_t \qquad \text{for all } e.$$

Equivalently,

$$\left\|\hat{\mathbf{R}}_t(f^+) - \hat{\mathbf{R}}_t(f)\right\|_\infty \le L_\ell \tau_t.$$

Applying Proposition 1 to the batched environment-risk vectors yields

$$\left|G_\lambda\big(\hat{\mathbf{R}}_t(f^+)\big) - G_\lambda\big(\hat{\mathbf{R}}_t(f)\big)\right| \le (1+\lambda) L_\ell \tau_t.$$

Finally,

$$\hat{J}_t(f^+) - \hat{J}_t(f) = \Big(G_\lambda(\hat{\mathbf{R}}_t(f^+)) - G_\lambda(\hat{\mathbf{R}}_t(f))\Big) + \eta\Big(\log(1+\Omega(f^+)) - \log(1+\Omega(f))\Big),$$

so another application of the triangle inequality gives

$$\left|\hat{J}_t(f^+) - \hat{J}_t(f)\right| \le (1+\lambda) L_\ell \tau_t + \eta\, \Delta_\Omega(f^+, f).$$

□

Corollary 3 (Population robust-score stability under trust-region updates) *Suppose f and f^+ are $\mathcal{F}_t$-measurable symbolic models. Under the assumptions of Theorem 1 and Proposition 2, for any $\delta \in (0,1)$,*

$$\left|J_\lambda(f^+) - J_\lambda(f)\right| \le (1+\lambda) L_\ell \tau_t + 2(1+\lambda) M \sqrt{\frac{\log(4E/\delta)}{2m_{\min,t}}} + \eta\, \Delta_\Omega(f^+, f)$$

with probability at least $1-\delta$.

Proof Apply Theorem 1 to f and f^+ separately with confidence level $\delta/2$. With probability at least $1-\delta$,

$$|J_\lambda(f) - \hat{J}_t(f)| \le \epsilon, \qquad |J_\lambda(f^+) - \hat{J}_t(f^+)| \le \epsilon,$$

where

$$\epsilon = (1+\lambda) M \sqrt{\frac{\log(4E/\delta)}{2m_{\min,t}}}.$$

On this event,

$$\begin{aligned} |J_\lambda(f^+) - J_\lambda(f)| &\le |J_\lambda(f^+) - \hat{J}_t(f^+)| + |\hat{J}_t(f^+) - \hat{J}_t(f)| + |\hat{J}_t(f) - J_\lambda(f)| \\ &\le \epsilon + \Big((1+\lambda) L_\ell \tau_t + \eta\, \Delta_\Omega(f^+, f)\Big) + \epsilon \\ &= (1+\lambda) L_\ell \tau_t + 2\epsilon + \eta\, \Delta_\Omega(f^+, f). \end{aligned}$$

Substituting the value of ϵ proves the claim. □

Remark 4 Proposition 2 is deterministic and remains valid even if f^+ is constructed using the monitored batch itself. Corollary 3, however, additionally requires the fixed-candidate condition in (A3). If the same batch is reused both to construct f^+ and to score it, extra uniform-control arguments over the induced candidate class are needed; the corollary above is the clean version valid for prespecified pairs or for decoupled implementations that evaluate candidates on a fresh batch.

Corollary 3 is the key bridge between batching and semantics. The first term is the explicit trust-region drift, the second is the residual sampling uncertainty from batch evaluation, and the third is the exact structural-complexity change. Hence robust-score drift is controlled jointly by semantic step size and batch fidelity.

4.5 Local suboptimality of batched selection

The final result formalizes the accuracy of batched minimization over a fixed finite candidate family.

Theorem 2 (Local suboptimality of batched selection) *Fix generation t, and let $\mathcal{C}_t$ be a finite $\mathcal{F}_t$-measurable candidate set. Define*

$$\hat{g}_t \in \arg\min_{g \in \mathcal{C}_t} \hat{J}_t(g), \qquad g_t^\star \in \arg\min_{g \in \mathcal{C}_t} J_\lambda(g).$$

If

$$\sup_{g \in \mathcal{C}_t} \left| \hat{J}_t(g) - J_\lambda(g) \right| \le \varepsilon_t,$$

then

$$J_\lambda(\hat{g}_t) \le J_\lambda(g_t^\star) + 2\varepsilon_t.$$

In particular, under Corollary 1, with probability at least $1-\delta$,

$$J_\lambda(\hat{g}_t) \le \min_{g \in \mathcal{C}_t} J_\lambda(g) + 2(1+\lambda) M \sqrt{\frac{\log(2EK_t/\delta)}{2m_{\min,t}}}.$$

Proof Assume

$$\sup_{g \in \mathcal{C}_t} \left| \hat{J}_t(g) - J_\lambda(g) \right| \le \varepsilon_t.$$

Then, in particular,

$$J_\lambda(\hat{g}_t) \le \hat{J}_t(\hat{g}_t) + \varepsilon_t$$

and

$$\hat{J}_t(g_t^\star) \le J_\lambda(g_t^\star) + \varepsilon_t.$$

Since $\hat{g}_t$ minimizes $\hat{J}_t$ over $\mathcal{C}_t$,

$$\hat{J}_t(\hat{g}_t) \le \hat{J}_t(g_t^\star).$$

Combining the last three inequalities gives

$$J_\lambda(\hat{g}_t) \le \hat{J}_t(\hat{g}_t) + \varepsilon_t \le \hat{J}_t(g_t^\star) + \varepsilon_t \le J_\lambda(g_t^\star) + 2\varepsilon_t.$$

This proves the deterministic $2\varepsilon_t$ bound.

For the explicit high-probability statement, Corollary 1 shows that with probability at least $1-\delta$,

$$\sup_{g \in \mathcal{C}_t} \left| \hat{J}_t(g) - J_\lambda(g) \right| \le (1+\lambda) M \sqrt{\frac{\log(2EK_t/\delta)}{2m_{\min,t}}}.$$

Substituting this value for ε_t into the deterministic part yields

$$J_\lambda(\hat{g}_t) \le J_\lambda(g_t^\star) + 2(1+\lambda) M \sqrt{\frac{\log(2EK_t/\delta)}{2m_{\min,t}}}.$$

Since $g_t^\star$ is a minimizer of J_λ over $\mathcal{C}_t$,

$$J_\lambda(g_t^\star) = \min_{g \in \mathcal{C}_t} J_\lambda(g),$$

and the result follows. □

Theorem 2 is intentionally local: it does not claim global convergence of the entire GP process. Its meaning is narrower and precise. At a fixed generation, once a finite comparison family has been specified independently of the evaluation batch, minimizing the batched robust score is nearly as good as minimizing the exact robust score over that same family, with an additive penalty that decays at the standard $m_{\min,t}^{-1/2}$ rate.

5 Experimental protocol

5.1 Benchmark design

The empirical study uses a controlled multi-environment SR benchmark with four analytic target functions:

$$\text{Poly2}: \; x_0 x_1 + x_1^2, \qquad \text{Rational2}: \; \frac{x_0}{1+x_1^2} + x_1,$$

$$\text{Trig2}: \; \sin(x_0) + x_1^2, \qquad \text{Mix3}: \; x_0 x_1 + \frac{x_2}{1+x_0^2}.$$

Each task is augmented by appending one *spurious* feature as the final input variable. Across the three training environments, this feature is correlated with the target with coefficients 1.5, 0.9, and -0.8, respectively, plus Gaussian noise. Thus the appended variable is genuinely predictive on the training mixture, but in an environment-dependent and ultimately misleading way.

The OOD evaluation uses three protocols:

1. **Shift:** the sign pattern of the spurious correlation is flipped relative to the training mixture;
2. **Range:** the support of the genuine covariates is expanded while the spurious feature becomes independent noise;
3. **Heavy:** the covariates become heavy-tailed and the spurious feature exhibits strong inverse correlation.

The benchmark is intentionally controlled rather than encyclopedic: the grammar remains matched to the analytic tasks so that the experiment isolates robustness mechanisms rather than conflating them with search-space mismatch.

5.2 Compared methods

We compare the following five algorithms:

- **StdGP-Full:** syntax-only GP with full-batch evaluation;
- **StdGP-Adaptive:** syntax-only GP with adaptive batching;
- **SemGP-Full:** semantic interpolation GP with full-batch evaluation;
- **SemGP-Adaptive:** semantic interpolation GP with adaptive batching but without trust-region projection and without robust validation selection;
- **ABTR-SGP:** the full proposed method.

Table 1: Overall results across 4 tasks and 4 seeds (mean ± standard deviation). Lower is better for all metrics.

Method	ID RMSE	Robust OOD	Worst OOD	Size	Evals (k)
StdGP-Full	0.397 ± 0.161	0.946 ± 0.267	1.353 ± 0.408	13.9 ± 8.9	192.0 ± 0.0
StdGP-Adaptive	0.434 ± 0.158	0.952 ± 0.313	1.349 ± 0.506	10.7 ± 7.5	165.7 ± 17.7
SemGP-Full	0.354 ± 0.135	0.798 ± 0.329	1.128 ± 0.506	59.5 ± 21.4	192.0 ± 0.0
SemGP-Adaptive	0.363 ± 0.128	0.838 ± 0.300	1.201 ± 0.480	64.2 ± 18.2	163.2 ± 15.1
ABTR-SGP	0.316 ± 0.137	0.764 ± 0.369	1.098 ± 0.567	28.3 ± 18.0	145.2 ± 30.3

All methods use the same arithmetic–trigonometric grammar, population size (40), generation budget (20), and four random seeds, yielding 16 paired runs per method over the task–seed combinations.

5.3 Metrics and statistical testing

We report the following quantities: in-distribution (ID) RMSE, mean robust OOD RMSE averaged across the three OOD protocols, worst-shift RMSE, model size measured by node count, and total fitness-case evaluations. The last metric captures the actual computational benefit of batching because it aggregates the number of fitness cases consumed during evolution.

Planned pairwise comparisons use one-sided Wilcoxon signed-rank tests [37]. Overall ranking on robust OOD RMSE is assessed by a Friedman test with mean-rank reporting, following common recommendations for multi-algorithm comparisons [38–40].

6 Results

6.1 Aggregate performance

Table 1 summarizes the aggregate results across all tasks and seeds. ABTR-SGP achieves the best mean ID RMSE (0.316) and the best mean robust OOD RMSE (0.764). It also obtains the best Friedman mean rank (2.250), and the omnibus Friedman test rejects equality among methods ($\chi^2 = 9.650$, $p = 0.0468$). These results indicate that the proposed coupling of adaptive batching, trust-region semantics, and robust validation is beneficial at the level of the full benchmark rather than on a single isolated task.

Relative to the two syntax-only baselines, the gains are statistically clear. Compared with StdGP-Full and StdGP-Adaptive, ABTR-SGP reduces mean robust OOD RMSE by 19.2% and 19.7%, with one-sided Wilcoxon $p = 0.0038$ and $p = 0.0370$, respectively. Relative to the stronger semantic baselines, the mean robust OOD RMSE improvements are smaller—4.2% versus SemGP-Full and 8.8% versus SemGP-Adaptive on average—and do not reach significance at $\alpha = 0.05$. This is an important nuance: ABTR-SGP is not presented as a method that overwhelmingly dominates all semantic baselines in raw predictive accuracy, but rather as one that improves the robustness–efficiency–complexity trade-off.

Table 2: Pairwise Wilcoxon signed-rank tests used in the main discussion. Negative mean difference favors ABTR-SGP.

Metric	Comparison	Mean diff.	Wilcoxon p
Robust RMSE	ABTR-SGP vs StdGP-Full	-0.181	0.0038
Robust RMSE	ABTR-SGP vs StdGP-Adaptive	-0.188	0.0370
Robust RMSE	ABTR-SGP vs SemGP-Full	-0.034	0.3529
Robust RMSE	ABTR-SGP vs SemGP-Adaptive	-0.073	0.1613
Size	ABTR-SGP vs SemGP-Full	-31.188	0.0001
Fitness cases	ABTR-SGP vs SemGP-Full	-46830.000	0.0000
Size	ABTR-SGP vs SemGP-Adaptive	-35.875	0.0002
Fitness cases	ABTR-SGP vs SemGP-Adaptive	-17985.000	0.0062

Table 3: Per-task mean robust OOD RMSE.

	StdGP Full	StdGP Adaptive	SemGP Full	SemGP Adaptive	ABTR SGP
Poly2	1.219000	1.278000	0.985000	1.084000	0.896000
Rational2	0.675000	0.635000	0.330000	0.361000	0.271000
Trig2	1.052000	1.061000	0.971000	0.992000	1.015000
Mix3	0.836000	0.834000	0.905000	0.914000	0.876000

The pairwise tests underlying these statements are listed in Table 2. Taken together, Tables 1 and 2 support the central empirical claim of the paper: uncertainty-aware batching is valuable, but it is substantially more effective when paired with bounded semantic motion and a robustness-aware final selection rule.

6.2 Task-wise behavior under shift

Table 3 reports the task-level robust OOD RMSE. The results are heterogeneous in a useful way. ABTR-SGP is clearly strongest on Poly2 and Rational2, which are the two tasks in which environment disagreement induced by the appended spurious feature is most consequential. By contrast, SemGP-Full remains strongest on Trig2, and the syntax-only variants are comparatively competitive on Mix3.

This pattern is consistent with the intended use case of the method. When the search landscape is relatively smooth and environment disagreement is limited, unconstrained semantic interpolation can already be highly effective. ABTR-SGP is most beneficial when two risks coexist: the risk of elite misranking caused by batched evaluation, and the risk of semantic overshoot caused by large behavior-space steps taken on partially observed data.

Figure 1 visualizes the distribution of robust OOD RMSE over all task–seed runs. The figure confirms that ABTR-SGP improves not only the mean but also the overall concentration of robust performance on the tasks where environment mismatch is most damaging.

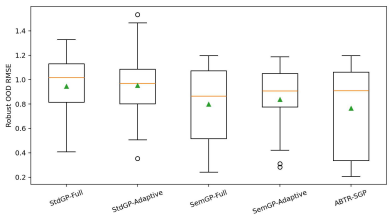
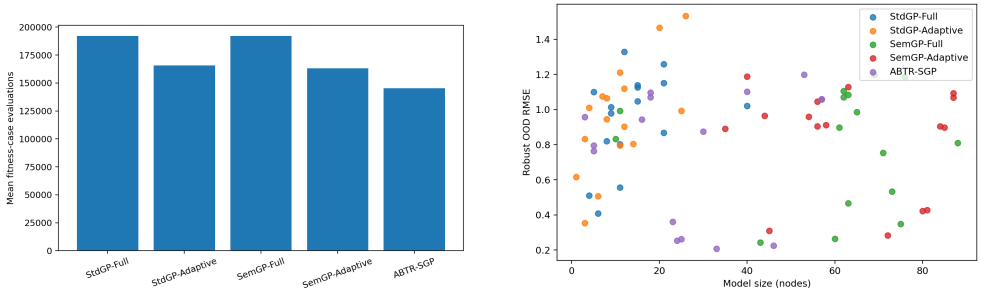


Fig. 1: Distribution of robust OOD RMSE over all task-seed runs.



(a) Mean fitness-case evaluations by method. (b) Per-run size versus robust OOD RMSE.

Fig. 2: Evaluation cost and complexity–robustness trade-off across methods.

6.3 Model size and evaluation budget

A particularly important result is that ABTR-SGP improves the practical efficiency profile of semantic GP. Relative to SemGP-Full, ABTR-SGP uses 24.4% fewer fitness-case evaluations and produces models that are 52.4% smaller on average. Relative to SemGP-Adaptive, it uses 11.0% fewer fitness-case evaluations and produces models that are 55.9% smaller. These differences are not marginal: the associated p-values are 0.0001 and 0.0002 for model size, and 0.0000 and 0.0062 for evaluation cost.

Figure 2a shows the evaluation-budget advantage. Full-batch semantic GP remains a strong accuracy baseline, but it pays substantially more in raw fitness-case evaluations. ABTR-SGP recovers a large fraction of that accuracy while using a meaningfully smaller budget.

Figure 2b shows the complexity–robustness trade-off at the run level. ABTR-SGP occupies a favorable region relative to the semantic baselines, achieving comparable or better robust OOD error with markedly smaller trees. This supports the interpretation that trust-region projection is doing more than dampening noise: it is also preventing the semantic operator from accumulating unnecessary structural baggage.

6.4 Adaptive batch dynamics

Figure 3 illustrates the batch-size trajectory for a representative Poly2 run. The qualitative behavior aligns with Corollary 2. When the elite gap is clear, small batches are

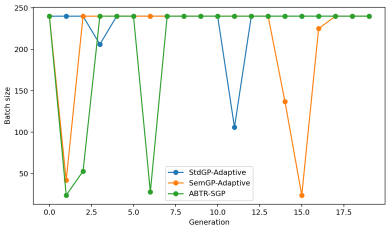


Fig. 3: Batch-size trajectories for a representative Poly2 run.

sufficient and the algorithm spends little computation. When the leading candidates become difficult to distinguish, the batch size increases sharply. ABTR-SGP couples these increases with a reduced semantic radius through (6); therefore, the generations with the highest ranking uncertainty are also those in which semantic movement is most tightly controlled.

This coupling is central to the method. Adaptive batching alone controls the fidelity of the score estimate, but not the magnitude of the semantic move proposed under that estimate. Conversely, trust-region semantics alone controls step size, but not whether the batch is large enough to support a reliable ranking decision. The empirical dynamics suggest that the two mechanisms reinforce each other in the intended way.

Representative models.

For reference and qualitative inspection, representative Rational2 expressions are listed in `results/example_models.csv` in the accompanying artifact.

7 Discussion

The empirical and theoretical evidence together supports three main conclusions.

First, *batching in GP should be uncertainty-aware rather than static.* The challenge is not merely to reduce computation, but to do so without corrupting the ranking signal that drives evolution. The concentration bound in Theorem 1 and the gap-based condition in Corollary 2 make this point explicit: the required batch size scales with the difficulty of the ranking problem, not with a fixed notion of iteration count.

Second, *semantic movement benefits from explicit trust regions.* Semantic GP is powerful precisely because it searches in behavior space, but that same property can become a liability when search-time scores are noisy or environment-specific. Proposition 2 and Corollary 3 show that bounding semantic displacement translates directly into bounded robust-score drift, up to the sampling error induced by batching and the structural change induced by interpolation. Empirically, this control appears to be one of the main reasons that ABTR-SGP attains far smaller trees than the stronger semantic baselines.

Third, *robustness is fundamentally a model-selection issue as well as a search issue.* In the present benchmark, the final choice of model cannot safely be delegated to average validation error because average error underweights catastrophic behavior

on adverse environments. The smooth worst-group surrogate in (1) gives a differentiable and selection-friendly approximation to worst-group validation, allowing the final criterion to express the stability preference that the benchmark is designed to test.

The study also has clear limitations. The theoretical results are local and conditional on the current generation; they justify batched scoring and trust-region selection within a generation, but they do not establish global convergence of the full evolutionary process. The trust-region guarantee is exact on the monitored batch and becomes a population guarantee only after invoking concentration. In addition, the benchmark is deliberately controlled and grammar-matched. This is a strength for mechanism isolation, but broader validation on larger community suites and harder scientific-discovery corpora remains necessary [4, 9, 10]. Finally, the current formulation assumes that environment labels are available. Extending the approach to latent or dynamically inferred environments is an important direction for future work, as is integrating high-dimensional feature selection ideas more directly into the robust semantic search process [27].

8 Conclusion

This paper introduced ABTR-SGP, a symbolic-regression algorithm that jointly addresses three sources of brittleness in GP under distribution shift: noisy batch-based fitness estimation, unconstrained semantic overshoot, and average-only model selection across heterogeneous environments. The method combines stratified adaptive batching, trust-region semantic interpolation, and a distributionally robust validation objective with explicit complexity regularization. The main-text theory shows that the batched robust score concentrates, that trust-region semantic updates induce bounded robust-score drift, and that batched minimization over a finite candidate set is locally near-optimal. In controlled multi-environment experiments with spurious features and multiple OOD protocols, ABTR-SGP achieves the best mean robust OOD accuracy among the tested methods while also reducing model size and evaluation cost relative to stronger semantic baselines. The broader methodological implication is that reliable symbolic discovery under shift requires the search operator, the evaluation budget, and the model-selection rule to be designed as a coupled system.

Code and data availability.

The accompanying package contains the benchmark generator, raw experimental outputs, tables, figures, and representative evolved expressions used to produce this manuscript.

References

[1] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992.

[2] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science*, vol. 324, no. 5923, pp. 81–85, 2009.

[3] S.-M. Udrescu and M. Tegmark, "AI Feynman: A physics-inspired method for symbolic regression," *Science Advances*, vol. 6, no. 16, Art. eaay2631, 2020.

[4] W. La Cava, P. Orzechowski, B. Burlacu, F. O. de Franca, M. Virgolin, Y. Jin, M. Kommenda, and J. H. Moore, "Contemporary symbolic regression methods and their relative performance," in *Proc. NeurIPS Datasets and Benchmarks Track*, 2021.

[5] L. Biggio, T. Bendinelli, A. Neitz, A. Lucchi, and G. Parascandolo, "Neural symbolic regression that scales," in *Proc. 38th Int. Conf. Machine Learning (ICML)*, *Proc. Mach. Learn. Res.*, vol. 139, pp. 936–945, 2021.

[6] P.-A. Kamienny, G. Lample, S. Lamprier, and M. Virgolin, "Deep generative symbolic regression with Monte-Carlo-Tree-Search," in *Proc. 40th Int. Conf. Machine Learning (ICML)*, *Proc. Mach. Learn. Res.*, vol. 202, pp. 15655–15668, 2023.

[7] T. Bendinelli, L. Biggio, and P.-A. Kamienny, "Controllable neural symbolic regression," in *Proc. 40th Int. Conf. Machine Learning (ICML)*, *Proc. Mach. Learn. Res.*, vol. 202, pp. 2063–2077, 2023.

[8] W. Li, W. Li, L. Yu, M. Wu, L. Sun, J. Liu, Y. Li, S. Wei, D. Yusong, and M. Hao, "A neural-guided dynamic symbolic network for exploring mathematical expressions from data," in *Proc. 41st Int. Conf. Machine Learning (ICML)*, *Proc. Mach. Learn. Res.*, vol. 235, pp. 28222–28242, 2024.

[9] K. S. Fong and M. Motani, "Pareto-optimal fronts for benchmarking symbolic regression algorithms," in *Proc. 42nd Int. Conf. Machine Learning (ICML)*, *Proc. Mach. Learn. Res.*, vol. 267, pp. 17392–17410, 2025.

[10] P. Shojaee, N.-H. Nguyen, K. Meidani, A. Barati Farimani, K. D. Doan, and C. K. Reddy, "LLM-SRBench: A new benchmark for scientific equation discovery with large language models," in *Proc. 42nd Int. Conf. Machine Learning (ICML)*, *Proc. Mach. Learn. Res.*, vol. 267, pp. 55325–55359, 2025.

[11] A. Moraglio, K. Krawiec, and C. G. Johnson, "Geometric semantic genetic programming," in *Proc. 12th Int. Conf. Parallel Problem Solving from Nature (PPSN XII)*, pp. 21–31, 2012.

[12] M. Castelli, S. Silva, and L. Vanneschi, "A C++ framework for geometric semantic genetic programming," *Genetic Programming and Evolvable Machines*, vol. 16, no. 1, pp. 73–81, 2015.

[13] M. Castelli, L. Trujillo, L. Vanneschi, and E. Z-Flores, "Geometric semantic genetic programming with local search," in *Proc. Genetic and Evolutionary Computation Conf. (GECCO)*, pp. 999–1006, 2015.

[14] I. Bakurov, J. M. Muñoz Contreras, M. Castelli, N. Rodrigues, S. Silva, L. Trujillo, and L. Vanneschi, "Geometric semantic genetic programming with normalized and standardized random programs," *Genetic Programming and Evolvable Machines*, vol. 25, Art. 6, 2024.

[15] L. Bonin, L. Rovito, A. De Lorenzo, and L. Manzoni, "Cellular geometric semantic genetic programming," *Genetic Programming and Evolvable Machines*, vol. 25, Art. 8, 2024.

[16] G. Dick, "An ensemble learning interpretation of geometric semantic genetic programming," *Genetic Programming and Evolvable Machines*, vol. 25, Art. 9, 2024.

[17] G. Nadizar, B. Sakallioglu, F. Garrow, S. Silva, and L. Vanneschi, "Geometric semantic GP with linear scaling: Darwinian versus Lamarckian evolution," *Genetic Programming and Evolvable Machines*, vol. 25, Art. 17, 2024.

[18] H. Zhang, Q. Chen, B. Xue, W. Banzhaf, and M. Zhang, "A geometric semantic macro-crossover operator for evolutionary feature construction in regression," *Genetic Programming and Evolvable Machines*, vol. 25, Art. 2, 2024.

[19] A. Leite and M. Schoenauer, "Memetic semantic boosting for symbolic regression," *Genetic Programming and Evolvable Machines*, vol. 26, Art. 11, 2025.

[20] M. Reissmann, Y. Fang, A. S. H. Ooi, and R. D. Sandberg, "Constraining genetic symbolic regression via semantic backpropagation," *Genetic Programming and Evolvable Machines*, vol. 26, Art. 12, 2025.

[21] L. Rovito, L. Bonin, D. Farinati, L. Vanneschi, L. Manzoni, A. De Lorenzo, and G. Pietropolli, "Semantic-based recombination and mutation in cellular-inspired genetic programming," *Genetic Programming and Evolvable Machines*, vol. 26, Art. 27, 2025.

[22] A. Moraglio, K. Krawiec, and C. Johnson, "Editorial introduction to the special issue for the tenth anniversary of geometric semantic genetic programming," *Genetic Programming and Evolvable Machines*, vol. 26, Art. 16, 2025.

[23] L. Rosenfeld and L. Vanneschi, "A survey on batch training in genetic programming," *Genetic Programming and Evolvable Machines*, vol. 26, Art. 2, 2025.

[24] E. Hemberg, S. Moskal, and U.-M. O'Reilly, "Evolving code with a large language model," *Genetic Programming and Evolvable Machines*, vol. 25, Art. 21, 2024.

[25] S. de Vries, S. Keemink, and M. van Gerven, "Discovering continuous-time memory-based symbolic policies using genetic programming," *Genetic Programming and Evolvable Machines*, vol. 27, Art. 5, 2026.

[26] K. Sen Fong and M. Motani, "Multi-level symbolic regression: Function structure learning for multi-level data," in *Proc. 27th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, *Proc. Mach. Learn. Res.*, vol. 238, pp. 2890–2898, 2024.

[27] C. Wang, Q. Chen, B. Xue, and M. Zhang, "Improving generalization of genetic programming for high-dimensional symbolic regression with Shapley value based feature selection," *Data Science and Engineering*, vol. 10, pp. 196–211, 2025.

[28] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, "Invariant models for causal transfer learning," *Journal of Machine Learning Research*, vol. 19, no. 36, pp. 1–34, 2018.

[29] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[30] K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar, "Invariant risk minimization games," in *Proc. 37th Int. Conf. Machine Learning (ICML)*, *Proc. Mach. Learn. Res.*, vol. 119, pp. 145–155, 2020.

[31] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (REx)," in *Proc. 38th Int. Conf. Machine Learning (ICML)*, *Proc. Mach. Learn. Res.*, vol. 139, pp. 5815–5826, 2021.

[32] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2020.

[33] J. Duchi and H. Namkoong, "Variance-based regularization with convex objectives," *Journal of Machine Learning Research*, vol. 20, no. 68, pp. 1–55, 2019.

[34] J. C. Duchi and H. Namkoong, "Learning models with uniform performance via distributionally robust optimization," *The Annals of Statistics*, vol. 49, no. 3, pp. 1378–1406, 2021.

[35] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[36] R. J. Serfling, "Probability inequalities for the sum in sampling without replacement," *The Annals of Statistics*, vol. 2, no. 1, pp. 39–48, 1974.

[37] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[38] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[39] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[40] S. García and F. Herrera, "An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons," *Journal of Machine Learning Research*, vol. 9, pp. 2677–2694, 2008.