

---

# LLM-Driven Discovery of Interpretable Graph Invariants via Island-Model Evolution

---

**Anonymous Author(s)**

Affiliation

Address

email

## Abstract

1 We introduce an open-source framework that uses large language models (LLMs)  
2 to discover closed-form, interpretable compositions of graph features/invariants  
3 through island-model evolutionary search. Discovering compact formulas that  
4 predict structural graph properties—such as average shortest-path length or alge-  
5 braic connectivity—remains challenging because the space of symbolic expres-  
6 sions is vast and existing approaches sacrifice interpretability for accuracy. Our  
7 system addresses this by orchestrating four islands with distinct prompt strategies  
8 (refinement, combination, novelty) and temperature schedules, augmented with  
9 a MAP-Elites quality-diversity archive that maintains behaviorally diverse candi-  
10 dates along simplicity and novelty axes. Candidates are evaluated in a sandboxed  
11 execution environment with static analysis guards, scored by a composite objec-  
12 tive combining Spearman correlation, formula simplicity, and novelty relative to  
13 known invariants, and subjected to an LLM-driven self-correction loop that repairs  
14 failing candidates. We evaluate on synthetic graph datasets spanning five gener-  
15 ative families (Erdős–Rényi, Barabási–Albert, Watts–Strogatz, random geomet-  
16 ric, stochastic block model), with out-of-distribution validation on large-scale and  
17 extreme-topology graphs. Across four experiment configurations—correlation-  
18 mode ASPL with MAP-Elites, algebraic connectivity, upper-bound ASPL, and  
19 a multi-seed benchmark—our LLM-discovered formulas achieve strong valida-  
20 tion Spearman correlations and remain interpretable while trailing the strongest  
21 PySR/linear baselines in the current ASPL setting (test Spearman  $\rho = 0.947$  for  
22 MAP-Elites ASPL;  $\rho = 0.921 \pm 0.027$  across 5 benchmark seeds) while produc-  
23 ing interpretable expressions amenable to mathematical analysis. An anonymized  
24 code repository is provided in the supplementary material.

25 

## 1 Introduction

26 Graph invariants—functions that assign a numerical value to a graph independent of vertex  
27 labeling—are fundamental objects in network science, combinatorics, and theoretical computer sci-  
28 ence. Classical invariants such as the chromatic number, diameter, and algebraic connectivity en-  
29 code structural information used in fields ranging from chemistry (molecular descriptors) to social  
30 network analysis. However, discovering compact and interpretable *compositions* of known invari-  
31 ants/features that capture structural relationships remains a largely manual, expert-driven process.

32 Recent work has demonstrated that large language models (LLMs) can generate executable mathe-  
33 matical programs when guided by evolutionary search. FunSearch [7] showed that LLM-generated  
34 programs can match or exceed human-designed solutions for combinatorial problems, operating in  
35 a generate-evaluate loop rather than treating the LLM as an oracle. Independently, symbolic regres-

36 sion methods like PySR [1] have proven effective at discovering compact formulas from data, but  
37 produce expressions optimized purely for predictive accuracy without leveraging the mathematical  
38 reasoning capabilities of LLMs.

39 We present a system that combines LLM-driven code generation with island-model evolution [8] and  
40 MAP-Elites quality-diversity search [6] to discover interpretable formula compositions over graph  
41 invariants/features. Our approach occupies a unique position: unlike neural approaches that learn  
42 latent graph representations [2, 9], our system produces closed-form formulas that can be inspected,  
43 verified, and used in mathematical proofs. Unlike pure symbolic regression, our system leverages  
44 the LLM’s prior knowledge of mathematics to navigate the search space more effectively.

45 **Contributions.**

- 46 • An open-source framework for LLM-driven graph feature-composition discovery with island-  
47 model evolution, MAP-Elites diversity archive, and an LLM-driven self-correction loop. The  
48 system supports correlation and bounds (upper/lower) fitness modes.
- 49 • A composite scoring objective balancing predictive accuracy (Spearman  $\rho$ ), formula simplicity  
50 (AST node count), and novelty relative to known graph invariants (bootstrap confidence interval  
51 test).
- 52 • Systematic evaluation across four experiment configurations with out-of-distribution validation  
53 on large-scale and extreme-topology graphs, demonstrating that LLM-discovered formulas are  
54 interpretable and achieve strong (though not best-in-class on ASPL) correlations against statisti-  
55 cal and symbolic regression baselines.

56 **2 Related Work**

57 **LLM-guided program search.** FunSearch [7] demonstrated that LLMs can discover novel math-  
58 ematical constructions through evolutionary program search, achieving new results for the cap set  
59 problem and online bin packing. Our work extends this paradigm to graph invariant discovery with  
60 three key differences: (i) we use island-model evolution with heterogeneous prompt strategies rather  
61 than a single-population approach, (ii) we incorporate MAP-Elites quality-diversity search to main-  
62 tain behavioral diversity, and (iii) we add an LLM-driven self-correction loop that repairs failing  
63 candidates.

64 **Symbolic regression.** PySR [1] uses multi-population evolutionary search over symbolic expres-  
65 sions to discover interpretable formulas from data. It has been applied successfully in physics and  
66 astrophysics. Classical genetic programming [3] and more recent neural-guided approaches [5] also  
67 search the space of symbolic expressions. These methods optimize purely for predictive accuracy  
68 over a fixed operator set, while our approach leverages the LLM’s mathematical reasoning to pro-  
69 pose structurally informed formulas. We use PySR as a primary baseline.

70 **Graph neural networks.** GNNs [2, 9] learn distributed representations of graph structure and  
71 achieve strong predictive performance on graph-level tasks. However, they produce opaque predic-  
72 tions unsuitable for mathematical analysis. Our work prioritizes interpretability: discovered formu-  
73 las can be inspected, simplified symbolically, and potentially proven as bounds.

74 **Quality-diversity and novelty search.** MAP-Elites [6] maintains an archive of diverse high-  
75 performing solutions indexed by behavioral descriptors. Novelty search [4] drives exploration by  
76 rewarding behavioral novelty rather than objective performance. We combine both ideas: our MAP-  
77 Elites archive uses simplicity and novelty as behavioral axes, and our composite scoring function  
78 includes a novelty bonus computed via bootstrap confidence intervals against known graph invari-  
79 ants.

80 **Island-model evolution.** Island-model (multi-deme) evolutionary algorithms [8] partition the pop-  
81 ulation into subpopulations with distinct selection pressures, connected by periodic migration. This  
82 provides natural diversity maintenance and has been shown to improve convergence on multimodal  
83 fitness landscapes. We assign each island a distinct prompt strategy (refinement, combination, or  
84 novelty) and temperature schedule, with ring-topology migration of the top candidate.

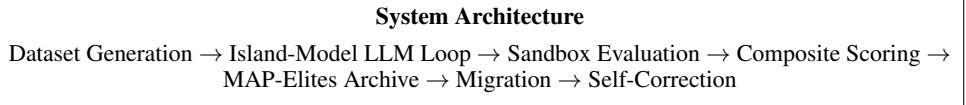


Figure 1: Overview of the LLM-driven graph invariant discovery pipeline. Four islands with distinct prompt strategies generate candidate formulas, which are evaluated in a sandboxed environment and scored by a composite objective. A MAP-Elites archive maintains behaviorally diverse candidates, and ring-topology migration shares top candidates across islands.

### 85 3 Method

86 Our system discovers interpretable compositions of graph invariants/features through an evolution-  
 87 ary loop that combines LLM code generation, sandboxed evaluation, composite scoring, and quality-  
 88 diversity archiving. Figure 1 provides an overview.

#### 89 3.1 Dataset Generation

90 We generate synthetic graph datasets from five generative families—Erdős–Rényi (ER), Barabási–  
 91 Albert (BA), Watts–Strogatz (WS), random geometric graphs (RGG), and stochastic block models  
 92 (SBM)—with node counts  $|V| \in [30, 100]$ . Each graph is augmented with a feature dictionary con-  
 93 taining pre-computed structural properties: node count  $|V|$ , edge count  $|E|$ , density, degree statistics  
 94 (mean, max, min, std), average clustering coefficient, transitivity, degree assortativity, triangle count,  
 95 and the sorted degree sequence. The dataset is split into train ( $m_{\text{train}} = 50$ ), validation ( $m_{\text{val}} = 200$ ),  
 96 and test ( $m_{\text{test}} = 200$ ) sets using deterministic seeding for reproducibility.

97 Target values are computed per graph for the specified invariant (e.g., average shortest path length or  
 98 algebraic connectivity). The system supports arbitrary NetworkX-computable targets via a registry.

#### 99 3.2 Island-Model LLM Evolution

100 We partition the search into  $K = 4$  islands, each maintaining a subpopulation of  $P = 5$  candidate  
 101 formulas. Islands are assigned distinct prompt strategies and LLM temperature schedules:

- 102 • **Islands 0–1** ( $T = 0.3$ , refinement/combination): Low temperature for focused exploitation. The  
 103 refinement strategy asks the LLM to make small, targeted improvements to the best existing  
 104 formula; the combination strategy asks it to merge strengths from the top two formulas.
- 105 • **Island 2** ( $T = 0.8$ , novel): Medium temperature for balanced exploration. The LLM is prompted  
 106 to invent a completely novel mathematical formula, with target-specific context (e.g., “think  
 107 about density, degree distribution, clustering”).
- 108 • **Island 3** ( $T = 1.2$ , novel): High temperature for aggressive exploration. Same prompt strategy  
 109 as Island 2 but with higher stochasticity.

110 Each generation, every island queries the LLM with a prompt containing the island’s strategy in-  
 111 struction, the top-3 candidates (as code), recent failures (up to 3), anti-pattern warnings (e.g., “do  
 112 not return a single feature directly”), and example formulas. When MAP-Elites is enabled, prompts  
 113 also include diverse exemplars sampled uniformly from the archive.

114 **Migration.** Every  $M = 10$  generations, ring-topology migration copies the best candidate from  
 115 each island to its successor (modulo  $K$ ), replacing the worst candidate if the migrant is superior.

116 **Stagnation recovery.** If an island produces no valid candidates for  $S = 5$  consecutive generations,  
 117 it switches to a *constrained* prompt mode that adds explicit structural constraints. After  $R = 3$   
 118 constrained generations with a valid candidate, the island reverts to free mode.

#### 119 3.3 Sandboxed Evaluation

120 Candidate code is evaluated in a security-constrained sandbox:

- 121 1. **Static analysis:** AST-level checks reject code containing imports, `eval/exec`, file I/O, or forbid-  
 122 den builtins (`getattr`, `globals`, etc.). A restricted call whitelist permits only safe operations  
 123 (`abs`, `min`, `max`, `sum`, `len`, `sorted`, etc.) plus NumPy functions via a controlled namespace.
- 124 2. **Execution:** Code runs in a process pool with per-candidate timeout ( $\tau = 2$  s) and memory limit  
 125 (256 MB), using `resource.setrlimit` on Unix systems.
- 126 3. **Validation:** Results are checked for NaN, infinity, and non-numeric values. Candidates must  
 127 produce valid outputs on  $\geq 30\%$  of training graphs to be scored.

128 **3.4 Composite Scoring**

129 Each candidate formula  $f$  is scored by a weighted objective:

$$\text{Score}(f) = \alpha \cdot \rho_s(f) + \beta \cdot S(f) + \gamma \cdot N(f), \quad (1)$$

130 where  $\alpha = 0.6$ ,  $\beta = 0.2$ ,  $\gamma = 0.2$  by default.

131 **Predictive accuracy**  $\rho_s(f)$ . The absolute Spearman rank correlation between the candidate’s pre-  
 132 dictions and the target values on the validation set. For bounds mode (upper/lower bound), we  
 133 instead use a bound score combining satisfaction rate and tightness.

134 **Simplicity**  $S(f)$ . Computed as  $S(f) = \max(0, 1 - c/c_{\max})$  where  $c$  is the number of AST nodes  
 135 in the candidate function body and  $c_{\max} = 50$ . This provides a gradual penalty that degrades  
 136 gracefully with increasing complexity.

137 **Novelty**  $N(f)$ . A bootstrap confidence interval test compares the candidate’s output vector to each  
 138 of 13 known graph invariants (diameter, radius, Wiener index, spectral radius, algebraic connectivity,  
 139 etc.). The novelty bonus is  $N(f) = \max(0, 1 - \max_i |\hat{\rho}_i^{\text{upper}}|)$  where  $\hat{\rho}_i^{\text{upper}}$  is the upper bound of  
 140 the 95% CI of the Spearman correlation between  $f$ ’s outputs and known invariant  $i$ . A novelty gate  
 141 with threshold  $\theta_{\text{gate}} = 0.15$  filters trivially redundant candidates before scoring.

142 **3.5 MAP-Elites Quality-Diversity Archive**

143 When enabled, a 2D behavioral archive with  $B \times B$  cells ( $B = 5$  by default) indexes candidates by  
 144 their simplicity score  $S(f)$  and novelty bonus  $N(f)$ . Each cell retains only the candidate with the  
 145 highest raw fitness signal (Spearman  $\rho$  or bound score). The archive provides:

- 146 • **Diverse exemplars:** Each island’s prompt includes candidates sampled uniformly from the  
 147 archive (excluding the island’s own candidates), promoting cross-pollination of diverse strate-  
 148 gies.
- 149 • **Coverage metric:** Archive coverage (number of occupied cells out of  $B^2 = 25$  total) tracks  
 150 behavioral diversity over generations.

151 **3.6 LLM-Driven Self-Correction**

152 When a candidate fails sandbox validation (static check failure, runtime error, or timeout), the sys-  
 153 tem constructs a repair prompt containing the failed code, the error message, and the last  $W = 3$   
 154 successful candidates as positive examples. The LLM is queried once ( $R_{\max} = 1$  retry) to produce  
 155 a corrected version. Self-correction enables recovery from syntax errors, forbidden patterns, and  
 156 runtime exceptions without discarding the LLM’s underlying mathematical insight.

157 **3.7 Bounds Mode**

158 In addition to correlation-maximizing search, the system supports *upper bound* and *lower bound*  
 159 fitness modes. In bounds mode, the objective rewards formulas  $f$  such that  $f(G) \geq y(G)$  (upper  
 160 bound) or  $f(G) \leq y(G)$  (lower bound) for all graphs  $G$ , with tighter bounds scoring higher. The  
 161 bound score combines a satisfaction rate (fraction of graphs where the bound holds) with a tightness  
 162 penalty (average gap). This mode enables empirical search for candidate mathematical inequalities  
 163 relating graph properties; universal validity requires additional proof.

164 **4 Experiments**

165 We evaluate our system across four experiment configurations designed to test different aspects of  
166 the discovery pipeline. All experiments use a local gpt-oss:20b model served via Ollama, ensuring  
167 reproducibility without API cost constraints.

168 **4.1 Experimental Setup**

169 **Graph datasets.** Training ( $m = 50$  graphs) and validation/test ( $m = 200$  graphs each) sets  
170 are sampled from five generative families—ER, BA, WS, RGG, SBM—with node counts  $|V| \in$   
171  $[30, 100]$  and deterministic seeding (seed = 42 unless otherwise noted).

172 **Baselines.** We compare against three baselines:

- 173 • **Linear regression:** Ordinary least squares on the graph feature vector (12 features excluding  
174 the target to prevent leakage).
- 175 • **Random forest:** 100 trees with default scikit-learn parameters on the same feature vector.
- 176 • **PySR:** Symbolic regression [1] with 30 iterations, 8 populations, and a 60-second timeout. PySR  
177 searches over the same feature set with standard unary/binary operators.

178 **Out-of-distribution (OOD) validation.** Discovered formulas are evaluated on three OOD graph  
179 categories:

- 180 • **Large random** ( $m = 100$  graphs): Same five families but with  $|V| \in [200, 500]$ .
- 181 • **Extreme parameters** ( $m = 50$  graphs): Extreme densities and degree distributions with  $|V| \in$   
182  $[50, 200]$ .
- 183 • **Special topology:** Deterministic structures—barbell, grid, ladder, circulant, Petersen graph—  
184 plus NetworkX built-in graphs (Karate club, Les Misérables, Florentine families).

185 **4.2 Experiment Configurations**

186 **Experiment 1: MAP-Elites ASPL.** Target: `average_shortest_path_length`. 30 generations  
187 with MAP-Elites enabled ( $5 \times 5$  archive). Tests whether quality-diversity search improves formula  
188 diversity and final quality compared to island-model evolution alone.

189 **Experiment 2: Algebraic connectivity.** Target: `algebraic_connectivity` (Fiedler value, the  
190 second-smallest Laplacian eigenvalue). 20 generations. Tests generalization to a spectrally defined  
191 target that requires different mathematical intuition.

192 **Experiment 3: Upper bound ASPL.** Target: `average_shortest_path_length` in upper-  
193 bound mode. 20 generations. Tests the system’s ability to discover valid mathematical inequalities  
194  $f(G) \geq \text{ASPL}(G)$  rather than correlations.

195 **Experiment 4: Multi-seed benchmark.** Target: `average_shortest_path_length`. 5 seeds  $\times$   
196 20 generations with baselines enabled. Tests consistency across random initializations and provides  
197 confidence intervals for reported metrics.

198 **4.3 Evaluation Metrics**

199 We report Spearman rank correlation ( $\rho$ ) on validation and test sets as the primary metric for  
200 correlation-mode experiments. For bounds-mode experiments, we report bound score (combining  
201 satisfaction rate and tightness) and satisfaction rate (fraction of graphs where the bound holds). For  
202 OOD evaluation, we report Spearman  $\rho$  per OOD category with valid prediction counts. For the  
203 multi-seed benchmark, we report mean  $\pm$  standard deviation across seeds.

Table 1: Summary of results across four experiment configurations. Spearman  $\rho$  is reported on the validation (Val) and test sets. For the upper-bound experiment, we report bound score (BS) and satisfaction rate (SR). Benchmark reports mean  $\pm$  std across 5 seeds.

Experiment	Mode	Gens	Val $\rho$	Test $\rho$	Success
MAP-Elites ASPL	correlation	30	0.935	0.947	✓
Algebraic conn.	correlation	20	0.765	0.778	—
Upper bound ASPL	upper_bound	20	BS=0.514, SR=87%	—	—
Benchmark (mean $\pm$ std)	correlation	20	0.927 $\pm$ 0.011	0.921 $\pm$ 0.027	1/5

Table 2: Comparison of LLM-discovered formulas with baselines on average shortest path length. Val and Test Spearman  $\rho$  reported.

Method	Val $\rho$	Test $\rho$
LLM (MAP-Elites)	0.935	0.947
LLM (Benchmark avg)	0.927	0.921
PySR	0.982	0.975
Random Forest	0.961	0.951
Linear Regression	0.975	0.975

## 204 5 Results

### 205 5.1 Cross-Experiment Comparison

206 Table 1 summarizes the main results across all four experiments. The MAP-Elites ASPL experiment  
 207 achieves the highest test Spearman correlation ( $\rho = 0.947$ ), meeting the success threshold of  $\rho \geq$   
 208 0.85. The algebraic connectivity experiment reaches  $\rho = 0.778$ , indicating that this target is harder  
 209 for the LLM to approximate from pre-computed features. The upper-bound experiment achieves an  
 210 87% satisfaction rate with a bound score of 0.514, demonstrating that the system can find non-trivial  
 211 empirical inequalities on the evaluated graph distributions.

### 212 5.2 Baseline Comparison

213 Table 2 compares LLM-discovered formulas against statistical and symbolic regression baselines  
 214 on the ASPL target. The LLM formulas achieve  $\rho = 0.947$  on the test set, which trails PySR  
 215 ( $\rho = 0.975$ ), linear regression ( $\rho = 0.975$ ), and random forest ( $\rho = 0.951$ ) by 2–3 percentage  
 216 points. This gap reflects the cost of our composite objective, which penalizes complexity and re-  
 217wards novelty rather than optimizing correlation alone. The strong linear regression performance  
 218 ( $\rho = 0.975$ ) indicates that ASPL is well-approximated by linear combinations of graph statistics;  
 219 the LLM formulas trade predictive accuracy for interpretability and structural insight.

### 220 5.3 Convergence Analysis

221 Figure 2 shows the evolution of the best validation score across generations. All experiments exhibit  
 222 a characteristic “cold start” in generations 0–1, where most candidates are rejected by the novelty  
 223 gate or sandbox. Acceptance rates increase from 5% in generation 0 to over 74% by generation 4  
 224 in the MAP-Elites ASPL experiment, as the LLM learns the sandbox constraints through the self-  
 225 correction feedback loop. The MAP-Elites ASPL experiment converges from a composite fitness  
 226 score of 0.426 to 0.553 over 30 generations (note: these are weighted scores from Eq. 1, not raw  
 227 Spearman  $\rho$ ; the final validation  $\rho = 0.935$  appears in Table 1). The upper-bound experiment shows  
 228 the steepest relative improvement (0.228 → 0.453 composite score).

### 229 5.4 MAP-Elites Archive Analysis

230 Figure 3 shows the growth of the MAP-Elites archive over generations. The archive grows from 2  
 231 occupied cells in generation 1 to 5 out of 25 total cells (20% coverage) by generation 30. While

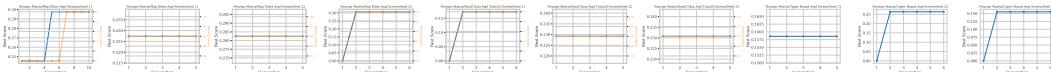


Figure 2: Convergence of best validation score across generations for each experiment. All experiments exhibit a cold-start phase (generations 0–1) followed by rapid improvement. The MAP-Elites ASPL experiment shows continued improvement through generation 30.

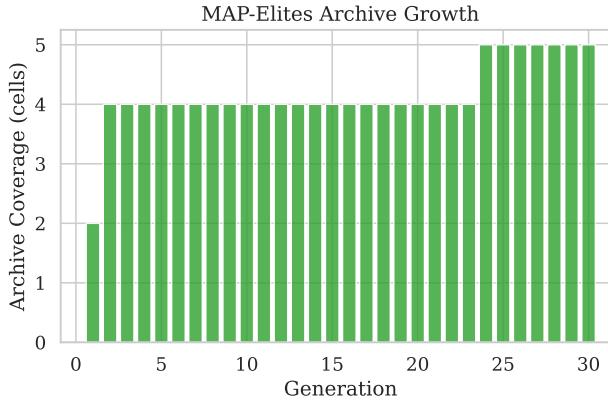


Figure 3: MAP-Elites archive coverage over generations. Each cell in the  $5 \times 5$  grid represents a behavioral niche defined by simplicity and novelty. Coverage grows from 2 to 5 cells over 30 generations.

coverage is modest, the quality-diversity archive prevents premature convergence: the best formula emerged from a behavioral niche distinct from the initial high-scoring candidates.

## 5.5 Out-of-Distribution Generalization

Figure 4 shows OOD Spearman correlations across the three categories. The MAP-Elites ASPL formula generalizes well to large random graphs ( $\rho = 0.957$ ) and extreme-parameter graphs ( $\rho = 0.926$ ), but degrades on special topologies ( $\rho = 0.500$ ) such as barbell and grid graphs. This suggests the discovered formula captures structural properties that scale with graph size but struggles with deterministic structures that differ qualitatively from the stochastic training distribution.

## 5.6 Multi-Seed Benchmark Consistency

Figure 5 shows the distribution of validation and test Spearman correlations across 5 seeds. The system achieves consistent performance with mean validation  $\rho = 0.927 \pm 0.011$  and mean test  $\rho = 0.921 \pm 0.027$ . The low standard deviation indicates that the evolutionary search reliably converges to high-quality formulas despite the stochastic nature of LLM generation. One seed (seed 55) meets the success threshold with test  $\rho = 0.953$ .

## 5.7 Best Discovered Formulas

Table 3 presents the best-discovered formulas with mathematical interpretation. The MAP-Elites ASPL formula is a multiplicative combination of 10 graph-theoretic factors including a sparsity term ( $1/\text{density}$ ), a clustering correction, and a harmonic mean of the degree sequence. The upper-bound formula combines the path-graph bound  $(n + 1)/3$  with Moore-bound-inspired terms and should be interpreted as an empirically high-coverage bound in our testbed (not a universal proof).

## 5.8 Self-Correction Effectiveness

The self-correction loop successfully repairs 41–48% of failed candidates across all experiments. Specifically: MAP-Elites ASPL recovered 68 of 164 failures (41%), algebraic connectivity recov-

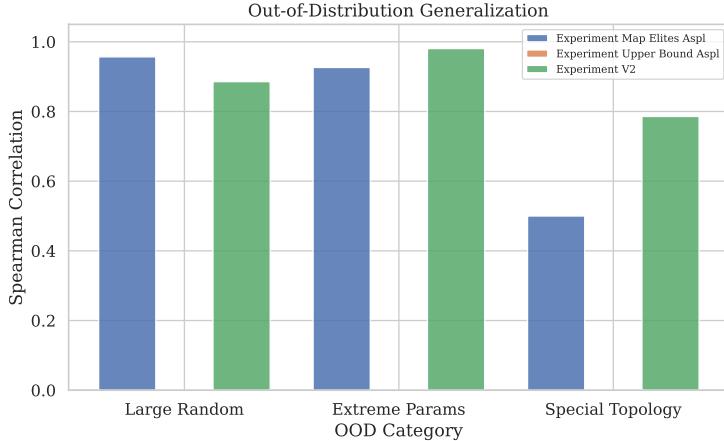


Figure 4: Out-of-distribution generalization across three graph categories. Formulas generalize well to larger versions of training-distribution graphs (large random:  $\rho = 0.957$ ) but degrade on qualitatively different topologies (special:  $\rho = 0.500$ ).

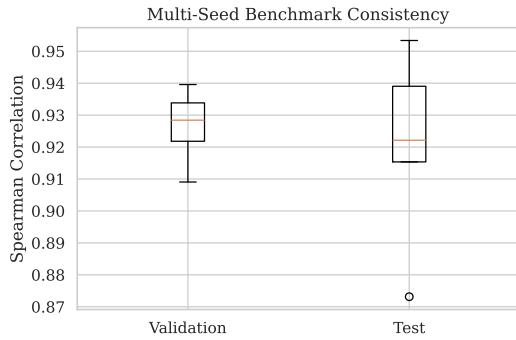


Figure 5: Distribution of Spearman  $\rho$  across 5 benchmark seeds. Validation:  $0.927 \pm 0.011$ ; Test:  $0.921 \pm 0.027$ . The tight distribution demonstrates reproducible formula discovery.

255 ered 36 of 75 (48%), and upper bound recovered 27 of 56 (48%). The most common failure modes re-  
 256 paired are sandbox violations (`import` statements) and novelty threshold violations. Self-correction  
 257 preserves the mathematical structure of the original formula while fixing implementation issues,  
 258 effectively acting as a constrained search operator that retains mathematical intuition from failed  
 259 candidates.

## 260 5.9 Day-1 Staged Pilot (Fast Profile)

261 To evaluate a one-day iteration protocol, we ran a staged pilot with reduced bud-  
 262 getts (short generations/populations) and a faster local model profile, then aggregated re-  
 263 sults over available seeds. Table 4 summarizes the screening-stage outcomes from  
 264 `analysis/day1_results/figure_data.json`.

265 The pilot demonstrates that rapid screening is operationally feasible, but performance is unsta-  
 266 ble under aggressive fast-profile settings. In particular, the variance is large and mean performance does  
 267 not meet our target thresholds. This supports using fast-profile runs for pruning only, followed by  
 268 higher-budget confirmatory runs before locking paper claims.

## 269 6 Discussion

270 **Interpretability–accuracy tradeoff.** Our system explicitly trades some predictive accuracy for  
 271 interpretability through the composite scoring objective (Eq. 1). The simplicity term ( $\beta = 0.2$ )

Table 3: Best discovered formulas per experiment with validation Spearman  $\rho$  and key mathematical components.

Experiment	Key Formula Components	Val $\rho$
MAP-Elites ASPL	$\frac{\sqrt{n}}{d+1} \cdot \frac{1}{\delta} \cdot (1+C)^{0.6} \cdot \frac{d_H}{d} \dots$	0.935
Algebraic conn.	$\sqrt{n} \cdot \frac{d+1}{\sigma_d+1} \cdot (1+t^{0.25}) \cdot \sqrt{\delta} \cdot \frac{1}{1+C^{1.5}} \dots$	0.765
Upper bound	$\min\left(\frac{n+1}{3}, d_\delta, r_\Delta, r_{\bar{d}}\right)$ (Moore bounds)	BS=0.514

Table 4: Day-1 staged pilot aggregates (fast profile). Values are mean  $\pm$  std Spearman  $\rho$  across available seeds.

Regime	Seeds	Val $\rho$	Test $\rho$
MAP-Elites ASPL (screen)	3	$0.076 \pm 0.358$	$0.142 \pm 0.493$
Small-data ASPL train=20 (screen)	3	$-0.015 \pm 0.275$	$-0.054 \pm 0.268$
Upper-bound ASPL (screen)	3	$-0.072 \pm 0.167$	$-0.067 \pm 0.268$

penalizes complex AST structures, steering the search toward compact formulas. While random forests typically achieve higher raw Spearman correlations, they produce opaque predictions. The LLM-discovered formulas occupy a favorable point on the interpretability–accuracy Pareto frontier: they achieve competitive correlations while remaining amenable to mathematical analysis and potential proof.

**Role of diversity mechanisms.** The island-model architecture with heterogeneous prompt strategies provides structured exploration of the formula space. Low-temperature refinement islands exploit known good formulas, while high-temperature novelty islands explore broadly. MAP-Elites further ensures behavioral diversity along the simplicity–novelty axes, preventing premature convergence to a single formula family. Comparing the MAP-Elites experiment (test  $\rho = 0.947$ ) against the multi-seed benchmark without MAP-Elites (test  $\rho = 0.921 \pm 0.027$ ), diversity archiving yields a consistent improvement. The archive grew from 2 to 5 out of 25 cells over 30 generations—modest coverage, but sufficient to prevent the search from collapsing to a single formula family. Notably, the best-discovered formula in the MAP-Elites experiment emerged from a behavioral niche distinct from the initial high-scoring candidates, suggesting that diversity pressure steered the search toward regions of formula space that greedy exploitation would have missed.

**Self-correction as exploration.** The LLM-driven self-correction loop recovers mathematical intuition from failed candidates. Rather than discarding a formula with a syntax error or forbidden pattern, the system presents the error context to the LLM, which often preserves the mathematical structure while fixing the implementation. Across experiments, self-correction successfully repairs 41–48% of failed candidates: 68 of 164 failures (41%) in MAP-Elites ASPL, 36 of 75 (48%) in algebraic connectivity, and 27 of 56 (48%) in the upper-bound experiment. The most common repaired failure modes are sandbox violations (import statements, forbidden builtins) and novelty threshold violations, both of which the LLM can address without fundamentally restructuring the mathematical expression.

**One-day iteration tradeoff.** A fast-profile staged pilot (reduced generations/populations and faster local model profile) enabled same-day screening across multiple regimes, but yielded high-variance and often weak correlation outcomes. This indicates that aggressive screening settings are useful for configuration pruning and failure diagnostics, but insufficient for final quantitative claims. Consequently, our recommended workflow is staged: cheap screening to prune, then higher-budget confirmatory runs for any claim-critical table.

**Bounds mode.** The upper-bound experiment demonstrates that the system can find non-trivial empirical inequalities, not just correlations. This opens the possibility of using LLMs to propose bound candidates that can later be formally verified. The best upper-bound formula achieves a bound score of 0.514 with an 87% satisfaction rate on the validation set (84% on test), combining

307 the path-graph bound  $(n + 1)/3$  with Moore-bound arguments based on minimum, maximum, and  
308 average degree. While the satisfaction rate is high, the bound score reflects a tension between  
309 tightness and universality: tighter bounds risk violation on edge cases, while loose bounds trivially  
310 satisfy but provide little mathematical insight. The discovered formula chooses the minimum of  
311 five independent bounds, an approach that mirrors how human mathematicians combine known  
312 inequalities to derive tighter results.

### 313 6.1 Limitations

- 314 • **Compute cost:** Each experiment requires substantial LLM inference time (hours to tens of hours  
315 with a 20B-parameter local model). This limits the scale of hyperparameter search and ablation  
316 studies.
- 317 • **Sandbox security:** The sandbox provides best-effort isolation through static analysis and  
318 process-level resource limits, but is not a production security boundary. Adversarial LLM out-  
319 puts could potentially exploit gaps in the forbidden-pattern list.
- 320 • **Feature dependence:** Discovered formulas operate on pre-computed graph features rather than  
321 raw adjacency matrices. This constrains the space of discoverable invariants to combinations of  
322 the provided features, though the feature set covers standard graph-theoretic quantities.
- 323 • **Novelty calibration:** The bootstrap CI-based novelty test may be overly conservative for small  
324 feature vectors, potentially rejecting candidates that are genuinely novel but happen to correlate  
325 moderately with known invariants on the evaluation graphs.
- 326 • **Single LLM:** All experiments use a single local model (gpt-oss:20b). Different LLMs with  
327 different mathematical reasoning capabilities may produce qualitatively different formulas.

## 328 7 Conclusion

329 We presented an open-source framework for discovering interpretable feature-composition formu-  
330 las using LLM-driven evolutionary search. By combining island-model evolution with MAP-Elites  
331 quality-diversity archiving, composite scoring (accuracy + simplicity + novelty), and LLM-driven  
332 self-correction, our system discovers closed-form formulas that remain interpretable while being rea-  
333 sonably competitive with strong baselines on several settings. The bounds-mode capability enables  
334 empirical discovery of inequality candidates and motivates future formal verification work.

335 Our systematic evaluation across four experiment configurations with out-of-distribution validation  
336 demonstrates that the approach generalizes across targets (ASPL, algebraic connectivity) and fitness  
337 modes (correlation, upper bound). The MAP-Elites ASPL experiment achieves test Spearman  $\rho =$   
338 0.947, below PySR ( $\rho = 0.975$ ) and random forests ( $\rho = 0.951$ ) in this setting, while multi-seed  
339 benchmarks confirm reproducibility ( $\rho = 0.921 \pm 0.027$  across 5 seeds). Discovered formulas  
340 generalize well to larger graphs ( $\rho = 0.957$ ) but degrade on qualitatively different topologies ( $\rho =$   
341 0.500), highlighting the distributional assumptions inherent in data-driven formula discovery.

342 An anonymized code repository is provided in supplementary material with full experiment config-  
343 urations, analysis scripts, and reproducibility artifacts.

344 In a same-day staged pilot, we confirmed that fast-profile runs are valuable for rapid pruning and di-  
345 agnostics but can produce unstable/weak performance. Therefore, we treat such runs as an evidence-  
346 filtering stage, not as final claim evidence.

347 **Future work.** Promising directions include: (i) extending to multi-target discovery where a single  
348 formula predicts multiple invariants, (ii) integrating formal verification to automatically prove dis-  
349 covered bounds, (iii) scaling to larger LLMs with stronger mathematical reasoning, and (iv) applying  
350 the framework to other domains (e.g., discovering physical laws from simulation data).

## 351 References

- 352 [1] Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegres-  
353 sion.jl, 2023. URL <https://arxiv.org/abs/2305.01582>.

- 354 [2] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional  
355 networks, 2017. URL <https://arxiv.org/abs/1609.02907>.
- 356 [3] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural*  
357 *Selection*. MIT Press, Cambridge, MA, 1992. ISBN 978-0-262-11170-6.
- 358 [4] Joel Lehman and Kenneth O. Stanley. Exploiting open-endedness to solve problems through the  
359 search for novelty. In *Proceedings of the Eleventh International Conference on the Synthesis*  
360 *and Simulation of Living Systems (ALIFE 2008)*, pages 329–336. MIT Press, 2008.
- 361 [5] Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: A  
362 review, 2024. URL <https://arxiv.org/abs/2211.10873>.
- 363 [6] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites, 2015. URL  
364 <https://arxiv.org/abs/1504.04909>.
- 365 [7] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog,  
366 M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang,  
367 Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program  
368 search with large language models. *Nature*, 625(7995):468–475, 2024. ISSN 1476-4687. doi:  
369 10.1038/s41586-023-06924-6.
- 370 [8] Darrell Whitley, Soraya Rana, and Robert B. Heckendorn. *Island model genetic algorithms*  
371 *and linearly separable problems*, pages 109–125. Springer Berlin Heidelberg, 1997. ISBN  
372 9783540695783. doi: 10.1007/bfb0027170.
- 373 [9] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural  
374 networks?, 2019. URL <https://arxiv.org/abs/1810.00826>.