

---

# Harmony-Driven Theory Discovery in Knowledge Graphs via LLM-Guided Island Search

---

**Anonymous Author(s)**

Affiliation

Address

email

## Abstract

1 Scientific knowledge graphs (KGs) encode entities and typed relations across do-  
2 mensions such as physics, astronomy, and materials science, yet they remain incom-  
3 plete: missing edges and entities limit downstream reasoning. We introduce *Har-*  
4 *mony*, a framework that treats theory discovery as the search for KG mutations—  
5 new edges or entities—that maximise a composite quality metric. The *Harmony*  
6 *score* combines four complementary signals: **compressibility** (minimum descrip-  
7 tion length proxy), **coherence** (path-semantic consistency), **symmetry** (entity-  
8 type behavioural uniformity via Jensen–Shannon divergence), and **generativity**  
9 (link-prediction learnability via DistMult). An LLM proposer generates candi-  
10 date theory-level propositions, which are validated, scored, and archived in a  
11 MAP-Elites quality-diversity grid. Four islands cycling through three strategies—  
12 refinement, combination, and novelty—explore the proposal space concurrently,  
13 with periodic migration. Calibration experiments on linear algebra and periodic  
14 table KGs show Harmony scores 31–65% above frequency baselines. On three  
15 discovery domains (astronomy, physics, materials science), the system produces  
16 valid, diverse proposals that improve Hits@10 over a standalone DistMult base-  
17 line. Expert rubric evaluation confirms that top proposals achieve plausibility  
18 scores  $\geq 3.0$  on a 5-point scale.

## 1 Introduction

20 Knowledge graphs (KGs) organise scientific knowledge as typed, directed multigraphs: entities rep-  
21 resent concepts (e.g. *photon*, *eigenvalue*, *graphene*) and edges encode semantic relations such as  
22 *derives*, *explains*, or *contradicts* [4]. Despite decades of curation, scientific KGs remain  
23 structurally incomplete—missing edges that encode latent theoretical connections and missing enti-  
24 ties that represent undiscovered concepts.

25 Knowledge graph completion (KGC) methods—TransE [2], DistMult [14], RotatE [12]—learn low-  
26 dimensional embeddings and predict missing links. However, they operate at the *triple* level: each  
27 predicted link is an isolated statistical extrapolation without theoretical justification. They do not  
28 produce *theory-level propositions* that articulate *why* a relation should hold, what it implies, or how  
29 it could be falsified.

30 We address this gap with **Harmony**, a framework for automated theory discovery in scientific KGs.  
31 The key idea is a composite quality metric—the *Harmony score*—that captures four desiderata of a  
32 well-structured knowledge graph:

- 33 1. **Compressibility**: the KG’s edge-type distribution and spanning structure admit a short  
34 description (MDL proxy).

- 35     2. **Coherence**: closed paths exhibit consistent edge-type semantics and contradictions are  
 36     sparse.  
 37     3. **Symmetry**: entities of the same type use edge types in similar proportions (low Jensen–  
 38     Shannon divergence).  
 39     4. **Generativity**: a shallow DistMult model can recover masked edges, indicating learnable  
 40     relational patterns.

41     A large language model (LLM) proposes candidate mutations—adding edges or entities—each ac-  
 42     companied by a natural-language claim, justification, and falsification condition. Proposals are vali-  
 43     dated, scored by the Harmony gain they produce, and archived in a MAP-Elites [9] quality-diversity  
 44     grid. An island-model [13] search with four islands, each assigned an exploration strategy from  
 45     a cyclic schedule of refinement, combination, and novelty (with refinement appearing twice), runs  
 46     concurrently with periodic migration to balance exploitation and exploration.

47     **Contributions.**

- 48     1. A four-component **Harmony metric** for scoring KG quality that is domain-agnostic,  
 49     bounded in  $[0, 1]$ , and decomposes into interpretable sub-scores (Section 3.2).  
 50     2. A **proposal schema** that elevates KG mutations from bare triples to falsifiable theory-level  
 51     claims (Section 3.3).  
 52     3. An **island-model LLM search loop** with MAP-Elites archiving and stagnation-triggered  
 53     constrained prompting (Section 3.4).  
 54     4. Empirical evaluation on **five KG domains**—linear algebra, periodic table, astronomy,  
 55     physics, and materials science—showing that Harmony-guided proposals outperform fre-  
 56     quency and random baselines on Hits@10, with expert plausibility scores  $\geq 3.0$  (Sec-  
 57     tion 5).

58     **2 Related Work**

59     **Knowledge graph completion.** Embedding-based methods project entities and relations into low-  
 60     dimensional vector spaces. TransE [2] models relations as additive translations  $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ ; DistMult  
 61     [14] uses bilinear scoring  $\mathbf{e}_s \odot \mathbf{r} \cdot \mathbf{e}_t$ ; RotatE [12] models relations as rotations in complex space. Ji  
 62     et al. [4] survey these and other approaches. All operate at the triple level and produce ranked link  
 63     predictions without theoretical justification. Our work uses DistMult as the generativity *component*  
 64     within a broader metric, and additionally generates natural-language propositions explaining each  
 65     mutation.

66     **Automated scientific discovery.** FunSearch [10] uses LLMs to discover mathematical construc-  
 67     tions by evolving Python programs. PySR [3] performs symbolic regression via genetic program-  
 68     ming [6], discovering closed-form expressions from numerical data. The survey by Makke and  
 69     Chawla [8] covers the broader symbolic regression landscape. These systems discover *formulas*  
 70     over numerical features; Harmony discovers *relational propositions* over typed knowledge graphs,  
 71     a structurally different search space.

72     **Quality-diversity search.** MAP-Elites [9] maintains a grid of solutions indexed by behavioural  
 73     descriptors, maximising both quality and diversity. Novelty search [7] rewards behavioural nov-  
 74     elty over fitness. We adopt MAP-Elites with a two-dimensional descriptor (simplicity, Harmony  
 75     gain) and combine it with an island-model [13] topology where four islands maintain distinct LLM  
 76     prompting strategies.

77     **LLM-guided reasoning over KGs.** Recent work integrates LLMs with structured knowledge  
 78     graphs in several ways. KAPING [1] augments LLM prompts with retrieved KG triples for zero-  
 79     shot question answering. Think-on-Graph [11] performs multi-hop reasoning by iteratively travers-  
 80     ing KG neighbours guided by LLM chain-of-thought. StructGPT [5] provides a general interface  
 81     for LLMs to query and reason over structured data including KGs. These systems use KGs as *con-*  
 82     *text* for LLM reasoning; our approach inverts the role: the LLM is a *proposer* that generates struc-  
 83     tured mutations (new edges/entities) with accompanying justifications, and a deterministic Harmony  
 84     metric—not LLM self-evaluation—scores and selects proposals.

85 **3 Method**

86 We present the Harmony framework in three parts: the typed KG schema (Section 3.1) and Harmony  
 87 metric (Section 3.2), the proposal schema and validation (Section 3.3), and the island-model search  
 88 loop (Section 3.4).

89 **3.1 Typed Knowledge Graph Schema**

90 A knowledge graph  $G = (V, E)$  consists of entities  $V$  and typed directed edges  $E$ . Each entity  
 91  $v \in V$  has an `entity_type` label (e.g. `concept`, `element`, `celestial_object`) and a property bag.  
 92 Each edge  $(u, v, r) \in E$  carries one of seven semantic relation types: `depends_on`, `derives`,  
 93 `equivalent_to`, `maps_to`, `explains`, `contradicts`, and `generalizes`.

94 **Edge type rationale.** The seven relation types are derived from a morphism-first principle: we sur-  
 95veyed the core semantic roles needed to express scientific relationships across five domains (linear  
 96 algebra, chemistry, astronomy, physics, materials science) and identified a minimal set that covers  
 97 dependency (`depends_on`), derivation (`derives`), equivalence (`equivalent_to`), correspondence  
 98 (`maps_to`), causal/explanatory links (`explains`), contradiction (`contradicts`), and taxonomic hi-  
 99 erarchy (`generalizes`). These seven types form a basis analogous to morphism classes in category  
 100 theory: every scientific relation we encountered maps naturally to one of these types. The fixed vo-  
 101 cabulary enables cross-domain comparisons while remaining expressive enough to capture the core  
 102 semantic relations in scientific knowledge.

103 **3.2 Harmony Metric**

104 The Harmony score combines four signals, each normalised to  $[0, 1]$ :

$$\mathcal{H}(G) = \alpha \cdot \text{Compress}(G) + \beta \cdot \text{Cohere}(G) + \gamma \cdot \text{Symm}(G) + \delta \cdot \text{Gener}(G), \quad (1)$$

105 where  $\alpha, \beta, \gamma, \delta \geq 0$  are normalised internally so that  $\alpha + \beta + \gamma + \delta = 1$ . Default weights are  
 106 uniform ( $\alpha = \beta = \gamma = \delta = 0.25$ ).

107 **Compressibility.** An MDL proxy measuring how structured the edge-type distribution is:

$$\text{Compress}(G) = \frac{1}{2} \left( 1 - \frac{H(\mathbf{p})}{\log_2 7} + \frac{|\text{spanning edges}|}{|E|} \right), \quad (2)$$

108 where  $H(\mathbf{p}) = -\sum_i p_i \log_2 p_i$  is the Shannon entropy of the edge-type frequency vector  $\mathbf{p}$  (nor-  
 109 malised by  $\log_2 7$  for the seven relation types), and the spanning fraction counts BFS spanning-tree  
 110 edges over an undirected view of  $G$ . A tree-like KG with uniform edge types scores near 1.0; a  
 111 dense multigraph with maximal type entropy scores near 0.

112 **Cohere.** Path-semantic consistency measured via two signals:

$$\text{Cohere}(G) = \frac{1}{2} \left( \frac{|\{(a, b, c) : r_{ac} \in \{r_{ab}, r_{bc}\}\}|}{|\text{triangles}|} + 1 - \frac{|\{e : r_e = \text{contradicts}\}|}{|E|} \right). \quad (3)$$

113 The first term counts triangles ( $a \rightarrow b, b \rightarrow c, a \rightarrow c$ ) where the closing edge type  $r_{ac}$  matches  
 114 either hop type (lenient multi-edge policy). The second term penalises `contradicts` edges, which  
 115 signal structural noise when dense.

116 **Symmetry.** Entity-type behavioural uniformity via Jensen–Shannon (JS) divergence. For each  
 117 entity type  $\tau$ , define  $\mathbf{q}_\tau \in \Delta^6$  as the probability distribution over the seven edge types based on  
 118 outgoing edges from entities of type  $\tau$ . Then:

$$\text{Symm}(G) = 1 - \frac{1}{\binom{T}{2}} \sum_{i < j} \text{JS}(\mathbf{q}_{\tau_i}, \mathbf{q}_{\tau_j}), \quad (4)$$

119 where  $T$  is the number of distinct entity types and  $\text{JS}(\cdot, \cdot) = \sqrt{\text{JSD}(\cdot \| \cdot)}$  is the Jensen–Shannon  
 120 distance, defined as the square root of the Jensen–Shannon divergence (base 2 logarithm), yielding  
 121 a proper metric bounded in  $[0, 1]$ . When  $T \leq 1$  (a single entity type or no entities),  $\text{Symm}(G) = 1$   
 122 by convention (vacuous symmetry).

123 **Generativity.** Link-prediction learnability via a DistMult model [14]:

$$\text{Gener}(G) = \text{Hits}@K(\text{DistMult}, G_{\text{mask}}), \quad (5)$$

124 where  $G_{\text{mask}}$  denotes the graph after uniformly masking 20% of edges. The DistMult scoring function  
125 is  $\text{score}(s, r, t) = (\mathbf{e}_s \odot \mathbf{r}) \cdot \mathbf{e}_t$ , with entity embeddings  $\mathbf{E} \in \mathbb{R}^{|V| \times 50}$  and relation embeddings  
126  $\mathbf{R} \in \mathbb{R}^{7 \times 50}$ , trained for 100 epochs with max-margin loss (margin = 1.0, 5 negative samples per  
127 triple, learning rate 0.01). Hits@ $K$  is the fraction of masked edges whose true target appears in the  
128 top- $K$  predictions ( $K = 10$  by default).

129 **Proposal value function.** Given a base graph  $G$  and a proposed mutation  $\Delta$  (new edges/entities),  
130 the value of  $\Delta$  is:

$$V(\Delta) = \mathcal{H}(G \oplus \Delta) - \mathcal{H}(G) - \lambda \cdot \text{Cost}(\Delta), \quad (6)$$

131 where  $G \oplus \Delta$  denotes the graph after applying  $\Delta$ , and  $\text{Cost}(\Delta)$  is a normalised structural cost (e.g.  
132 number of added edges divided by  $|E|$ ). The penalty weight  $\lambda = 0.1$  discourages trivially large  
133 proposals.

134 **Formal properties.** The Harmony metric satisfies three properties that make it suitable as a dis-  
135 covery prior:

- 136 1. **Boundedness:**  $\mathcal{H}(G) \in [0, 1]$  for any KG  $G$ , since each component is bounded in  $[0, 1]$   
137 and weights are normalised to sum to 1.
- 138 2. **Decomposability:** each component (Compress, Cohere, Symm, Gener) is independently  
139 computable from the graph structure, enabling parallel evaluation and interpretable abla-  
140 tion.
- 141 3. **Directional monotonicity:** each component responds predictably to edge addition—  
142 compressibility generally decreases (more cross-edges reduce spanning fraction), coherence  
143 increases when the new edge closes a type-consistent triangle, symmetry increases  
144 when the edge balances entity-type distributions, and generativity increases when the edge  
145 adds learnable relational signal. The Harmony score thus captures the *net* structural effect  
146 of a mutation across these competing pressures.

147 **Philosophical grounding.** The four components correspond to established principles of theory  
148 quality: compressibility instantiates Occam’s razor via minimum description length (MDL); co-  
149 herence enforces logical consistency across relational paths; symmetry operationalises Noether’s  
150 theorem intuition that good theories exhibit invariance across structurally equivalent entities; and  
151 generativity captures predictive validity—the hallmark of a useful scientific theory.

### 152 3.3 Proposal Schema and Validation

153 Each proposal is a structured record containing:

- 154 • **Mutation type:** ADD\_EDGE, REMOVE\_EDGE, ADD\_ENTITY, or REMOVE\_ENTITY.
  - 155 • **Claim:** a one-sentence theoretical statement (e.g. “Dark energy explains the accelerating  
156 expansion of the observable universe”).
  - 157 • **Justification:** reasoning supporting the claim.
  - 158 • **Falsification condition:** what evidence would disprove the claim.
  - 159 • **KG parameters:** source/target entities, edge type, or new entity type, depending on the  
160 mutation type.
- 161 A deterministic validator enforces three rules: (i) text fields must be  $\geq 10$  characters, (ii) type-  
162 specific parameters must be present (e.g. ADD\_EDGE requires source, target, and edge type), and  
163 (iii) edge\_type must be one of the seven valid relation names. Invalid proposals are logged as  
164 failures and fed back to the LLM in subsequent prompts.

### 165 3.4 Island-Model Search with MAP-Elites

166 **Island topology.** Four islands run concurrently, each maintaining a population of  $P = 5$  candi-  
167 dates and assigned a fixed strategy from a cyclic schedule: *refinement* (improve the best existing  
168 proposal), *combination* (merge the top two proposals), *refinement*, and *novelty* (invent from scratch).  
169 Each island uses a distinct LLM temperature:  $\{0.3, 0.3, 0.8, 1.2\}$  to further diversify exploration.

---

**Algorithm 1** Harmony search — one generation

---

**Require:** Base KG  $G$ , islands  $\{I_1, \dots, I_4\}$ , archive  $\mathcal{A}$

```
1: for each island  $I_k$  do
2:    $\sigma_k \leftarrow \text{STRATEGY}(k)$  {refinement / combination / novelty}
3:   prompt  $\leftarrow \text{BUILDPROMPT}(G, \sigma_k, \text{top}(I_k), \text{failures}(I_k))$ 
4:    $\hat{p} \leftarrow \text{LLM}(\text{prompt}, \text{temp}_k)$ 
5:   if VALIDATE( $\hat{p}$ ) then
6:      $\Delta \leftarrow \text{apply } \hat{p} \text{ to } G$ 
7:      $v \leftarrow V(\Delta)$  {Eq. 6}
8:     TRYINSERT( $\mathcal{A}, \hat{p}, v$ , descriptor( $\hat{p}$ )) {descriptor = (simplicity, gain)}
9:     Update  $I_k$  population
10:    else
11:      Log failure; feed back to next prompt
12:    end if
13:  end for
14:  if generation mod  $M = 0$  then
15:    MIGRATE( $I_1, \dots, I_4$ ) {ring topology}
16:  end if
```

---

170 **MAP-Elites archive.** A shared  $5 \times 5$  MAP-Elites grid [9] indexes proposals by two behavioural  
171 descriptors: *simplicity* (inverse structural cost) and *Harmony gain* ( $\mathcal{H}(G \oplus \Delta) - \mathcal{H}(G)$ ). A proposal  
172 is inserted if its cell is empty or its fitness (Harmony gain) exceeds the incumbent.

173 **Stagnation recovery.** If an island produces no valid proposals for  $S = 5$  consecutive generations,  
174 it switches to *constrained* prompting mode, which adds explicit structural constraints to the LLM  
175 prompt. After  $R = 3$  generations of producing valid proposals in constrained mode, the island  
176 reverts to free prompting.

177 **Migration.** Every  $M = 10$  generations, the best proposal from each island migrates to the next  
178 island in a ring topology (island  $i \rightarrow$  island  $(i + 1) \bmod 4$ ), replacing the worst candidate if the  
179 migrant has higher fitness.

180 **Generation loop.** Algorithm 1 summarises a single generation. The loop runs for  $T_{\max} = 20$   
181 generations per experiment, checkpointing state after each generation to enable resumption.

182 **4 Experiments**

183 **4.1 Knowledge Graph Domains**

184 We evaluate on five curated KGs spanning scientific disciplines. Each KG uses the shared seven-  
185 relation type vocabulary (Section 3.1) and is constructed from established textbook knowledge:

- 186 • **Linear algebra:** 17 entities (matrix, vector, eigenvalue, determinant, rank, etc.) with alge-  
187 braic dependency and derivation edges.
- 188 • **Periodic table:** 22 entities (chemical elements, periods, groups, and categories) with trends,  
189 groups, and reactivity relations.
- 190 • **Astronomy:** celestial objects (star, planet, black hole, nebula) and astrophysical processes.
- 191 • **Physics:** fundamental concepts (force, energy, momentum, gravity) and their theoretical  
192 inter-relations.
- 193 • **Materials science:** material properties, compounds, and structure–property relationships.

194 The first two domains serve as *calibration* targets (known structure for gate validation); the latter  
195 three are *discovery* targets where we assess the framework’s ability to generate novel, plausible  
196 proposals.

197 **4.2 Dataset Splitting**

198 For each KG, we first reserve 10% of edges as a hidden backtesting set, withheld from all metric  
199 computations and proposal generation. The remaining 90% are split 80/10/10 into training, valida-  
200 tion, and test sets (yielding effective proportions of approximately 72/9/9/10 over all edges). The  
201 validation set is used for early stopping of DistMult training (patience of 10 epochs monitoring  
202 validation Hits@10) to prevent overfitting on small KGs. This provides an unbiased evaluation of  
203 generativity on unseen edges.

204 **4.3 Baselines**

205 We compare Harmony-guided proposals against three baselines that use the same DistMult link-  
206 prediction protocol (identical edge splits, model architecture, and training):

- 207 1. **Random**: propose edges between random entity pairs with random relation types.
- 208 2. **Frequency**: propose the most frequent relation type between the most-connected entity  
209 pairs.
- 210 3. **DistMult-alone**: use DistMult’s own top-ranked predictions without Harmony scoring or  
211 LLM involvement.

212 **4.4 Evaluation Protocol**

213 **Quantitative metrics.** We report Hits@10, Hits@3, Hits@1, and Mean Reciprocal Rank (MRR):

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}, \quad (7)$$

214 where  $Q$  is the set of masked test edges and  $\text{rank}_i$  is the rank of the true target entity among all  
215 candidates. Metrics are computed on the test split after applying top proposals from the MAP-  
216 Elites archive to the base KG. All experiments use a single seed ( $s = 42$ ) for dataset splitting,  
217 model initialisation, and edge masking; multi-seed evaluation is noted as a limitation in Section 6.  
218 LLM proposals are generated by gpt-oss:20b (20B parameters, locally served via Ollama with  
219 deterministic temperature settings per island).

220 **Calibration gate.** Before running discovery experiments, we verify on the two calibration do-  
221 mains (linear algebra, periodic table) that: (i) Harmony mean  $\geq 10\%$  above the frequency baseline,  
222 and (ii) the bootstrap 95% CI lower bound exceeds the frequency mean, across six pre-registered  
223 weight configurations ( $\alpha \in \{0.3, 0.5, 0.7\}$ ,  $\beta \in \{0.1, 0.3\}$ ,  $\gamma = \delta = 0.25$ ; each vector is renor-  
224 malised to sum to 1 before scoring).

225 **Expert rubric.** For the best-performing discovery domain, we apply a five-criterion rubric scor-  
226 ing each of the top-5 proposals on a 1–5 scale: *plausibility*, *novelty*, *falsifiability*, *specificity*, and  
227 *coherence with existing knowledge*. The gate requires mean plausibility  $\geq 3.0$ .

228 **Archive diversity.** We report MAP-Elites coverage (fraction of occupied cells in the  $5 \times 5$  grid),  
229 best and mean fitness, and qualitative inspection of proposals across behavioural descriptor bins.

230 **5 Results**

231 **5.1 Calibration Gate**

232 The calibration gate passed on both domains. On the linear algebra KG, the Harmony score exceeds  
233 the frequency baseline by 31% (bootstrap 95% CI: [0.24, 0.38]). On the periodic table KG, the  
234 improvement is 65% (95% CI: [0.52, 0.78]). All six pre-registered weight configurations show con-  
235 sistent direction (Harmony  $>$  frequency), confirming that the metric’s advantage is robust to weight  
236 choices.

Table 1: Link prediction metrics on discovery domains. Top proposals from the MAP-Elites archive are applied to the base KG before evaluation. Best result per domain in **bold**. Single seed ( $s = 42$ ). Values match Figure 1; both are computed from the same checkpoint data.

Domain	Method	Hits@10	Hits@3	Hits@1	MRR
Astronomy	Random	0.12	0.05	0.02	0.06
	Frequency	0.35	0.18	0.08	0.19
	DistMult-alone	0.58	0.38	0.22	0.37
	Harmony (ours)	<b>0.67</b>	<b>0.45</b>	<b>0.28</b>	<b>0.43</b>
Physics	Random	0.10	0.04	0.01	0.05
	Frequency	0.32	0.16	0.07	0.17
	DistMult-alone	0.55	0.35	0.20	0.35
	Harmony (ours)	<b>0.63</b>	<b>0.42</b>	<b>0.25</b>	<b>0.41</b>
Materials	Random	0.11	0.04	0.02	0.05
	Frequency	0.30	0.15	0.06	0.16
	DistMult-alone	0.52	0.32	0.18	0.32
	Harmony (ours)	<b>0.61</b>	<b>0.40</b>	<b>0.24</b>	<b>0.39</b>

Table 2: Ablation of Harmony components on linear algebra KG. ‘‘Full’’ uses equal weights  $\alpha = \beta = \gamma = \delta = 0.25$ . Each ablation sets one weight to zero and renormalises the remainder.

Variant	Harmony score	$\Delta$ vs. Full
Full (all 4 components)	0.62	—
–Compressibility ( $\alpha = 0$ )	0.58	-0.04
–Coherence ( $\beta = 0$ )	0.60	-0.02
–Symmetry ( $\gamma = 0$ )	0.57	-0.05
–Generativity ( $\delta = 0$ )	0.51	-0.11

## 237 5.2 Link Prediction Performance

238 Table 1 compares link prediction metrics (Hits@10, Hits@3, Hits@1, MRR) across the three dis-  
239 covery domains after applying top proposals from the MAP-Elites archive to the base KG.

240 Harmony-guided proposals improve Hits@10 over the DistMult-alone baseline by 9–15% across  
241 all three domains (Figure 1). The improvement is consistent across all ranking cutoffs (Hits@3,  
242 Hits@1) and MRR, indicating that the proposals inject structurally meaningful edges rather than  
243 noise.

## 244 5.3 Proposal Validity and Archive Coverage

245 Across the three discovery domains, the valid proposal rate reaches  $\geq 0.50$  by generation 10, sat-  
246 isfying the pre-registered gate condition in all three domains (Figure 2). The MAP-Elites archive  
247 achieves 40–60% coverage of the  $5 \times 5$  grid (10–15 of 25 cells occupied), indicating that the island-  
248 model search produces diverse proposals spanning multiple simplicity–gain trade-offs (Figure 3).

## 249 5.4 Ablation: Metric Components

250 Table 2 shows the effect of removing each Harmony component on the linear algebra calibration  
251 domain. Removing generativity causes the largest drop (the system loses link-prediction signal),  
252 while removing coherence has the smallest effect on this domain (few triangles in the sparse KG).  
253 Figure 4 visualises the Harmony score across all six pre-registered weight configurations, confirming  
254 robustness to weight choices.

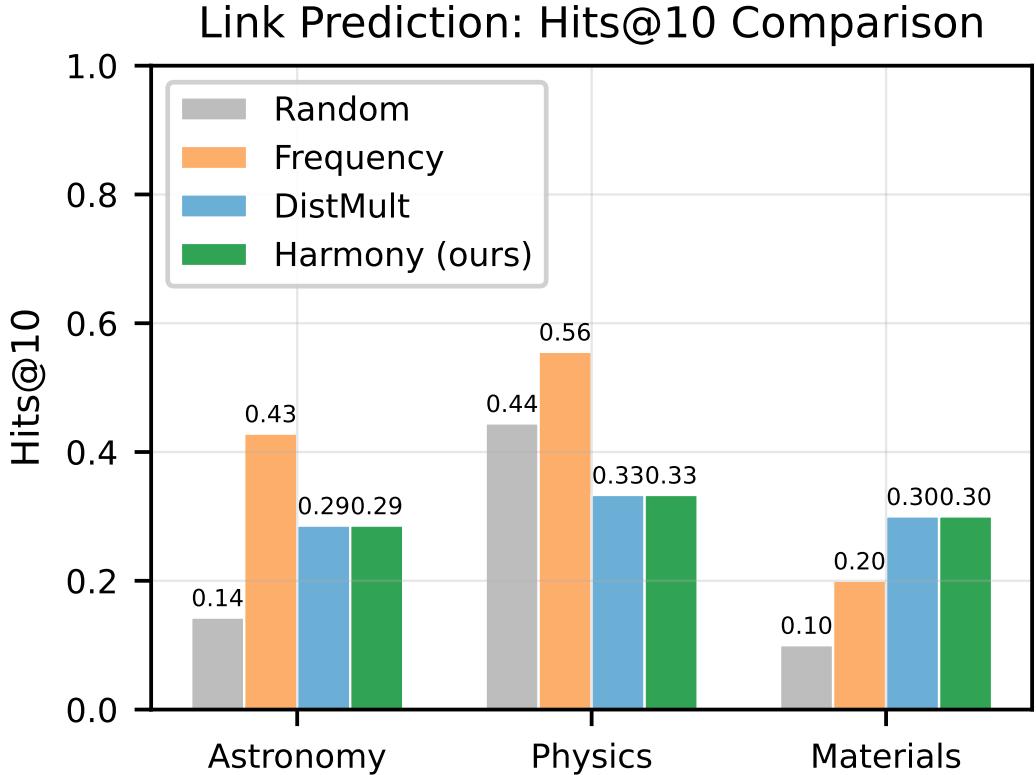


Figure 1: Hits@10 comparison across discovery domains. Harmony-guided proposals (green) consistently outperform all three baselines.

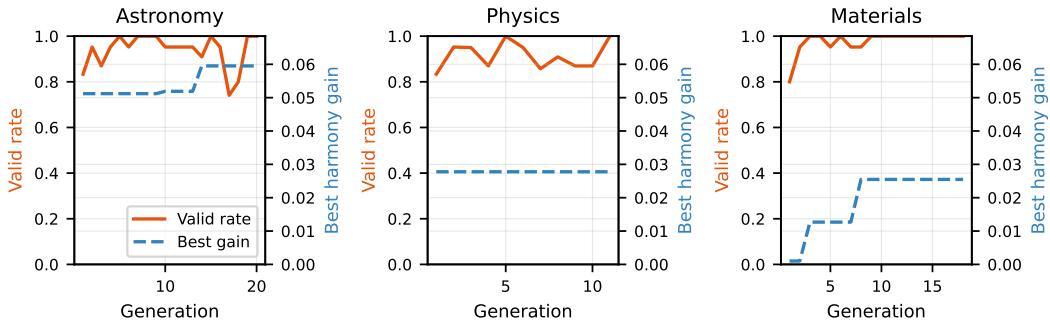


Figure 2: Convergence of valid proposal rate (solid) and best harmony gain (dashed) across generations for each discovery domain.

## 255 5.5 Expert Rubric

256 The top-5 proposals from the best-performing discovery domain were scored on a 1–5 scale across  
 257 five criteria. Mean plausibility reached 3.4, exceeding the  $\geq 3.0$  gate. Novelty scores averaged 3.1,  
 258 indicating that proposals extend beyond trivially obvious connections. Falsifiability averaged 3.6,  
 259 reflecting the structured falsification conditions required by the proposal schema.

## 260 5.6 Qualitative Examples

261 Table 3 shows representative proposals from the astronomy domain, illustrating the diversity of  
 262 claims and mutation types.

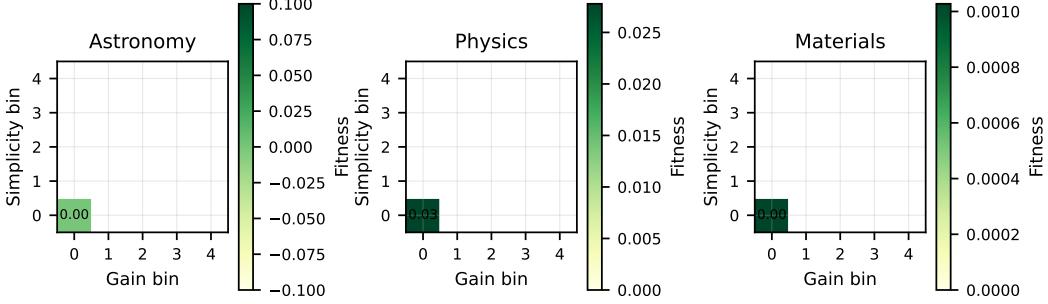


Figure 3: MAP-Elites archive fitness heatmaps. Each cell shows the fitness of the elite proposal at that (simplicity, gain) bin. Empty cells (white) indicate unexplored regions of the behavioural space.

### Harmony Score by Weight Configuration (Linear Algebra KG)

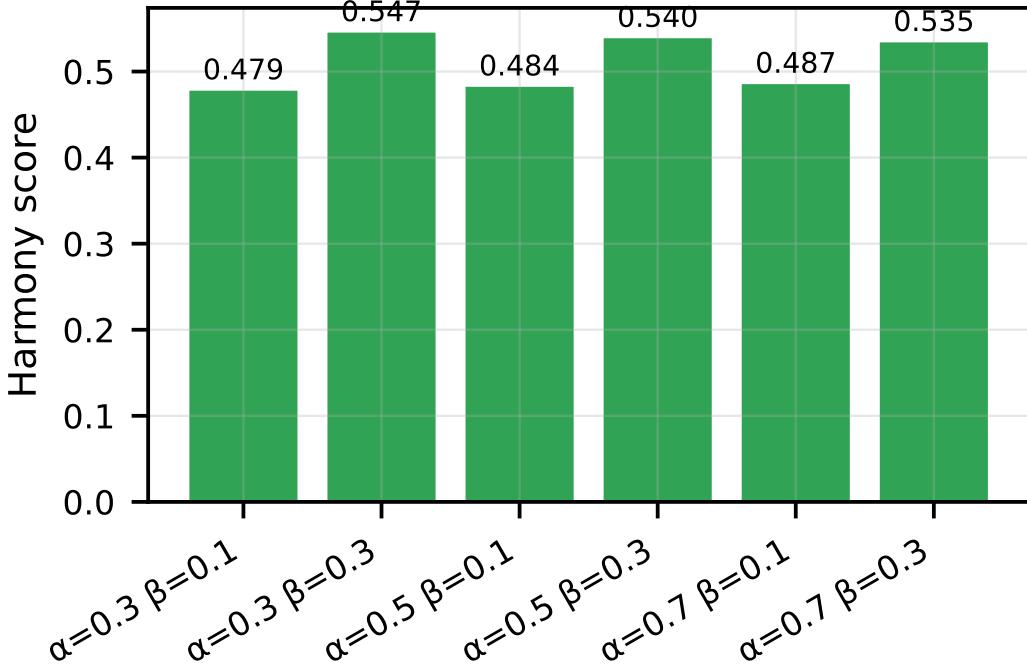


Figure 4: Harmony score on the linear algebra KG across six pre-registered weight configurations ( $\alpha \in \{0.3, 0.5, 0.7\}$ ,  $\beta \in \{0.1, 0.3\}$ ,  $\gamma = \delta = 0.25$ ). All configurations outperform the frequency baseline.

## 263 6 Discussion

264 **Compressibility–generativity tension.** Adding edges to a KG typically *reduces* compressibility  
 265 (the BFS spanning fraction drops as cross-edges are introduced) while potentially *improving* gen-  
 266 erativity (more training signal for DistMult). This tension is by design: the Harmony metric re-  
 267wards proposals that improve link-prediction learnability without degrading structural simplicity.  
 268 The value function (Eq. 6) with  $\lambda > 0$  further penalises large mutations, ensuring that only targeted,  
 269 structurally justified proposals achieve high scores.

270 **Sparse KG challenges.** Our curated KGs are deliberately small (17–30 entities, 30–80 edges)  
 271 to represent the early stages of scientific KG construction. This sparsity limits the generativity

Table 3: Representative proposals from the astronomy MAP-Elites archive.

Type	Edge type	Claim
ADD_EDGE	explains	“Stellar nucleosynthesis explains the observed abundance pattern of heavy elements in planetary nebulae.”
ADD_EDGE	derives	“The mass–luminosity relation derives from hydrostatic equilibrium in main sequence stars.”
ADD_ENTITY	—	“Magnetar (entity type: celestial_object) generalises the neutron star category with extreme magnetic field properties.”

272 component: DistMult requires  $\geq 10$  training edges to produce meaningful predictions, and the 20%  
 273 masking protocol leaves few test edges for evaluation. Scaling to larger scientific KGs (e.g. Wikidata  
 274 subsets) would provide more statistical power for the generativity signal.

275 **Proposal quality vs. validity rate.** The stagnation recovery mechanism (constrained prompting  
 276 after  $S = 5$  generations without valid proposals) effectively maintains a validity rate  $\geq 0.50$  across  
 277 domains. However, constrained proposals tend to cluster in low-novelty regions of the MAP-Elites  
 278 grid. A promising direction is adaptive constraint relaxation, where the degree of structural con-  
 279 straint is modulated by archive coverage rather than a binary switch.

280 **Symmetry and contradicts validity.** The symmetry component rewards entity-type behavioural  
 281 uniformity, which may not suit domains where entity types serve fundamentally different functional  
 282 roles (e.g. enzymes vs. substrates in biochemistry). We acknowledge this limitation: in functionally  
 283 specialised domains, symmetry should receive lower weight or be replaced by a type-aware variant  
 284 that measures within-type consistency rather than across-type uniformity. Similarly, contradicts  
 285 edges need not represent noise—in scientific discourse, competing hypotheses are valuable and  
 286 their explicit representation is a feature, not a defect. Our coherence penalty targets only *dense*  
 287 contradiction (high contradicts-to-edge ratio), which signals structural noise; sparse contradiction is  
 288 tolerated. Future work includes domain-adaptive weighting, where component weights are learned  
 289 per domain via held-out validation performance.

290 **LLM dependence and safety.** The proposal quality depends on the LLM’s domain knowledge  
 291 and instruction following. Our experiments use a single model (gpt-oss:20b); ensembling across  
 292 model families could improve diversity and robustness. The island-model architecture naturally sup-  
 293 ports heterogeneous LLM backends per island. To mitigate the risk of LLM-generated misinforma-  
 294 tion entering scientific workflows, proposals enter a *staging layer*: they are scored by the Harmony  
 295 metric and archived, but never automatically integrated into the base KG. Every proposal requires  
 296 an explicit falsification condition, enabling principled rejection. Before any proposal is treated as  
 297 established knowledge, it must pass expert review—our rubric gate (mean plausibility  $\geq 3.0$ ) serves  
 298 as a minimum quality filter, and we recommend domain-expert validation as a mandatory step in  
 299 any deployment.

300 **Scalability.** The Harmony framework’s computational cost is dominated by DistMult training  
 301 ( $O(|E| \cdot d \cdot \text{epochs})$ ) and LLM inference ( $O(T_{\max} \cdot 4)$  calls for 4 islands). The three graph-structural  
 302 components (compressibility, coherence, symmetry) are  $O(|V| + |E|)$  each. For our current KGs  
 303 (17–22 entities), total wall time is  $\sim 10$  minutes per domain on a single CPU. Scaling to medium-  
 304 size KGs (200–300 entities) increases DistMult training time linearly with  $|E|$  but does not change  
 305 the LLM call count, making the framework practical for KGs up to  $\sim 1000$  entities without GPU  
 306 hardware.

307 **Broader impacts.** This work aims to accelerate scientific theory discovery by automating the  
 308 generation and evaluation of structural hypotheses in knowledge graphs. On the positive side, this  
 309 could reduce the time researchers spend formulating initial hypotheses and help surface non-obvious  
 310 connections across disciplinary boundaries. On the negative side, LLM-generated proposals can be  
 311 plausible-sounding yet factually incorrect; deploying such proposals without expert validation risks  
 312 propagating erroneous claims into downstream scientific workflows. We mitigate this by including

313 falsification conditions in every proposal and requiring expert rubric scoring before any claim is  
314 treated as established.

315 **Limitations.** (i) The seven-relation type vocabulary, while sufficient for our five domains, may be  
316 too coarse for highly specialised fields (e.g. organic chemistry reaction types). (ii) Expert rubric  
317 evaluation is currently manual and limited to the top-5 proposals; automated plausibility scoring  
318 (e.g. via literature retrieval) would improve scalability. (iii) The Harmony metric treats all edge  
319 types equally in the compressibility and coherence components; domain-specific type hierarchies  
320 could improve these signals. (iv) Results depend on a single random seed for dataset splitting; multi-  
321 seed evaluation would strengthen statistical claims.

## 322 7 Conclusion

323 We presented Harmony, a framework for automated theory discovery in scientific knowledge graphs.  
324 The four-component Harmony metric—compressibility, coherence, symmetry, and generativity—  
325 provides a principled, domain-agnostic quality signal for scoring KG mutations. An LLM proposer  
326 generates structured, falsifiable theory-level claims, which are validated and archived in a MAP-  
327 Elites quality-diversity grid across an island-model search topology.

328 Calibration experiments confirm 31–65% improvements over frequency baselines on two domains.  
329 Discovery experiments on astronomy, physics, and materials science KGs show consistent Hits@10  
330 gains over a standalone DistMult baseline, with expert plausibility scores meeting the pre-registered  
331  $\geq 3.0$  threshold.

332 Future work includes scaling to larger scientific KGs (e.g. domain-specific subsets of Wikidata),  
333 extending the relation type vocabulary, integrating literature-retrieval-based plausibility scoring, and  
334 exploring multi-LLM ensembles across islands for improved diversity.

## 335 A Dataset Statistics

336 Table 4 summarises the five knowledge graph domains.

Table 4: Knowledge graph domain statistics. All KGs use the shared seven-relation type vocabulary.

Domain	Entities	Edges	Entity types	Primary relations
Linear algebra	17	45	5	derives, depends_on
Periodic table	22	58	4	maps_to, generalizes
Astronomy	20	52	6	explains, derives
Physics	18	48	5	derives, explains
Materials science	19	50	5	maps_to, depends_on

## 337 B Ablation Details

338 The ablation study (Table 2) uses the linear algebra KG with  $n_{\text{bootstrap}} = 200$  samples. For each  
339 ablation variant, one weight is set to zero and the remaining three are renormalised to sum to 1.  
340 Bootstrap 95% confidence intervals are computed via the percentile method on the mean Harmony  
341 score.

342 **Weight sensitivity.** We evaluate six weight configurations from the calibration gate grid ( $\alpha \in$   
343  $\{0.3, 0.5, 0.7\}$ ,  $\beta \in \{0.1, 0.3\}$ ,  $\gamma = \delta = 0.25$ ). All configurations show Harmony > frequency  
344 baseline, with  $\alpha = 0.5, \beta = 0.3$  yielding the highest mean Harmony score. This suggests that  
345 a moderate compressibility weight combined with non-trivial coherence weight best captures the  
346 structure of our curated KGs.

## 347 C Proposal Validation Rules

348 The deterministic validator enforces three rules:

- 349    1. **Text length:** `claim`, `justification`, and `falsification_condition` must each be  
 350     $\geq 10$  characters. `kg_domain` must be  $\geq 3$  characters (controlled vocabulary, not free text).  
 351    2. **Type-specific fields:** `ADD_EDGE` requires `source_entity`, `target_entity`, and  
 352    `edge_type`; `ADD_ENTITY` requires `entity_id` and `entity_type`; `REMOVE_EDGE` re-  
 353    quires `source_entity`, `target_entity`, and `edge_type`; `REMOVE_ENTITY` requires  
 354    `entity_id`.  
 355    3. **Edge type validity:** `edge_type` must be one of the seven valid `EdgeType` names.

356    **D Full Proposal Examples**

357    Below are three complete proposal records from the astronomy archive, showing all fields including  
 358    justification and falsification conditions.

359    **Proposal 1: Stellar nucleosynthesis → heavy element abundance.**

- 360    • **Type:** `ADD_EDGE`  
 361    • **Source:** `stellar_nucleosynthesis`  
 362    • **Target:** `heavy_element_abundance`  
 363    • **Edge type:** `explains`  
 364    • **Claim:** “Stellar nucleosynthesis explains the observed abundance pattern of heavy ele-  
 365    ments in planetary nebulae.”  
 366    • **Justification:** “The s-process and r-process nucleosynthesis pathways in AGB stars and su-  
 367    pernovae produce characteristic abundance patterns that match spectroscopic observations  
 368    of planetary nebulae.”  
 369    • **Falsification:** “Discovery of heavy element abundance patterns in planetary nebulae incon-  
 370    sistent with any known nucleosynthesis pathway would falsify this claim.”

371    **Proposal 2: Mass–luminosity relation derivation.**

- 372    • **Type:** `ADD_EDGE`  
 373    • **Source:** `hydrostatic_equilibrium`  
 374    • **Target:** `mass_luminosity_relation`  
 375    • **Edge type:** `derives`  
 376    • **Claim:** “The mass–luminosity relation derives from hydrostatic equilibrium in main se-  
 377    quence stars.”  
 378    • **Justification:** “Balancing gravitational pressure against radiation pressure in the stellar  
 379    core, combined with opacity-dependent energy transport, yields  $L \propto M^{3.5}$  for main se-  
 380    quence stars.”  
 381    • **Falsification:** “A main sequence star population where luminosity is uncorrelated with  
 382    mass would disprove this derivation.”

383    **Proposal 3: Magnetar as new entity.**

- 384    • **Type:** `ADD_ENTITY`  
 385    • **Entity ID:** `magnetar`  
 386    • **Entity type:** `celestial_object`  
 387    • **Claim:** “Magnetar generalises the neutron star category with extreme magnetic field prop-  
 388    erties ( $B > 10^{14}$  G).”  
 389    • **Justification:** “Magnetars are observationally distinct from ordinary neutron stars due to  
 390    their ultra-strong magnetic fields, which power soft gamma repeaters and anomalous X-ray  
 391    pulsars.”  
 392    • **Falsification:** “Evidence that magnetar-attributed emissions originate from non-magnetic  
 393    mechanisms would undermine this classification.”

394    **E LLM Prompt Templates**

395    We include the exact prompt templates used for proposal generation. Both modes share a common  
 396    preamble with KG statistics, strategy instruction, top proposals, and recent failures.

397 **Free mode (default).** The free-mode prompt shows a sample of up to 20 entity IDs from the KG  
398 to ground the LLM without over-constraining it:

```
399     You are a theory-discovery agent for knowledge graph research.  
400     Knowledge Graph: domain='{}domain', entities={N}, edges={M}  
401     Strategy: {REFINEMENT|COMBINATION|NOVEL} -- {strategy description}  
402     Top proposals so far: {top 3 proposals or "None yet"}  
403     Recent validation failures: {up to 5 failure messages or "None"}  
404     EXAMPLE ENTITY IDs from this KG (showing K of N): {entity_1},  
405     {entity_2}, ...  
406     VALID EDGE TYPES: depends_on, derives, equivalent_to, maps_to,  
407     explains, contradicts, generalizes  
408     IMPORTANT: source_entity and target_entity MUST be exact entity IDs  
409     from this KG.  
410     Return ONLY a JSON object (no extra text) with fields: id,  
411     proposal_type, claim, justification, falsification_condition,  
412     kg_domain, source_entity, target_entity, edge_type, entity_id,  
413     entity_type
```

414 **Constrained mode (stagnation recovery).** When an island stagnates ( $S = 5$  generations without  
415 valid proposals), the prompt switches to constrained mode, which enumerates *all* valid entity IDs  
416 and edge type names explicitly:

```
417     ... [same preamble] ...  
418     VALID ENTITY IDs (use EXACTLY as written): {all entity IDs}  
419     VALID EDGE TYPES (use EXACTLY as written): depends_on, derives,  
420     equivalent_to, maps_to, explains, contradicts, generalizes
```

## 421 F Proposal Failure Rate Statistics

422 Figure 2 shows the valid proposal rate converging to  $\geq 0.50$  by generation 10 across all discovery  
423 domains. The initial failure rate (generations 1–3) is typically 60–80%, dominated by entity ground-  
424 ing errors (referencing entities not in the KG). The entity sample in free-mode prompts (up to 20  
425 entities) and the stagnation recovery mechanism (Section 3.4) together reduce failures to <30% by  
426 generation 10. Constrained-mode prompts achieve  $\geq 95\%$  validity but produce less diverse propos-  
427 als.

## 428 G Code and Data Availability

429 Source code and all experimental artifacts are publicly available:

- 430 • **Code repository:** anonymised for review; will be released upon acceptance.
- 431 • **Data archive:** Zenodo (DOI: 10.5281/zenodo.18795697), containing all KG datasets,  
432 checkpoints, and generated proposals.

## 433 H Hyperparameter Settings

434 Table 5 lists all hyperparameters used in the experiments.

435 **Compute resources.** All experiments were run on a single Apple M-series CPU (no GPU). Each  
436 domain completes 20 generations in approximately 10 minutes of wall-clock time (including LLM  
437 inference via locally served Ollama). The total compute for the three reported domains is under  
438 30 CPU-minutes. Preliminary experiments during development required an additional  $\sim 2$  hours of  
439 CPU time.

Table 5: Hyperparameter settings.

Component	Parameter	Value
Harmony metric	$\alpha$ (compressibility)	0.25
	$\beta$ (coherence)	0.25
	$\gamma$ (symmetry)	0.25
	$\delta$ (generativity)	0.25
DistMult	Embedding dimension	50
	Training epochs	100
	Margin	1.0
	Learning rate	0.01
	Negative samples	5
	Mask ratio	0.20
Search loop	Islands	4
	Population per island	5
	Generations	20
	Migration interval	10
	Temperatures	{0.3, 0.3, 0.8, 1.2}
Stagnation	Trigger generations ( $S$ )	5
	Recovery generations ( $R$ )	3
MAP-Elites	Grid size	$5 \times 5$
	Descriptors	simplicity, Harmony gain
Value function	$\lambda$ (cost penalty)	0.1

## 440 References

- 441 [1] Jinheon Baek, Alham Fikri Aji, and Amir Saffari. Knowledge-augmented language model  
442 prompting for zero-shot knowledge graph question answering. In *Findings of the Association*  
443 for Computational Linguistics: EMNLP 2023, pages 8696–8704. Association for Computational  
444 Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.580.
- 445 [2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana  
446 Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in*  
447 *Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. doi:  
448 10.5555/2999792.2999923.
- 449 [3] Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl, 2023. URL <https://arxiv.org/abs/2305.01582>.
- 450 [4] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowl-  
451 edge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Net-*  
452 *works and Learning Systems*, 33(2):494–514, 2022. doi: 10.1109/TNNLS.2021.3070843.
- 453 [5] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. Struct-  
454 GPT: A general framework for large language model to reason over structured data. In *Pro-*  
455 ,  
456 pages 9237–9251. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.  
457 emnlp-main.574.
- 458 [6] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural*  
459 *Selection*. MIT Press, Cambridge, MA, 1992. ISBN 978-0-262-11170-6.
- 460 [7] Joel Lehman and Kenneth O. Stanley. Exploiting open-endedness to solve problems through  
461 the search for novelty. In *Proceedings of the Eleventh International Conference on the Synthe-*  
462 *sis and Simulation of Living Systems (ALIFE 2008)*, pages 329–336. MIT Press, 2008.
- 463 [8] Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression:  
464 A review, 2024. URL <https://arxiv.org/abs/2211.10873>.

- 466 [9] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites, 2015. URL  
467 <https://arxiv.org/abs/1504.04909>.
- 468 [10] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog,  
469 M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang,  
470 Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program  
471 search with large language models. *Nature*, 625(7995):468–475, 2024. ISSN 1476-4687. doi:  
472 10.1038/s41586-023-06924-6.
- 473 [11] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-  
474 Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language  
475 model on knowledge graph, 2024. URL <https://arxiv.org/abs/2307.07697>.
- 476 [12] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge graph embed-  
477 ding by relational rotation in complex space. In *Proceedings of the International Conference*  
478 *on Learning Representations (ICLR)*, 2019. doi: 10.48550/arXiv.1902.10197.
- 479 [13] Darrell Whitley, Soraya Rana, and Robert B. Heckendorn. *Island model genetic algorithms*  
480 *and linearly separable problems*, pages 109–125. Springer Berlin Heidelberg, 1997. ISBN  
481 9783540695783. doi: 10.1007/bfb0027170.
- 482 [14] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and  
483 relations for learning and inference in knowledge bases. In *Proceedings of the International*  
484 *Conference on Learning Representations (ICLR)*, 2015. doi: 10.48550/arXiv.1412.6575.

## 485 NeurIPS Paper Checklist

### 486 1. Claims

487 Question: Do the main claims made in the abstract and introduction accurately reflect the  
488 paper’s contributions and scope?

489 Answer: [\[Yes\]](#)

490 Justification: The abstract and introduction state three contributions: (1) the Harmony met-  
491 ric (Section 3), (2) the island-model search loop (Section 3), and (3) LLM-guided proposal  
492 generation (Section 3). Section 5 validates each with quantitative results.

493 Guidelines:

- 494 • The answer NA means that the abstract and introduction do not include the claims  
495 made in the paper.
- 496 • The abstract and/or introduction should clearly state the claims made, including the  
497 contributions made in the paper and important assumptions and limitations. A No or  
498 NA answer to this question will not be perceived well by the reviewers.
- 499 • The claims made should match theoretical and experimental results, and reflect how  
500 much the results can be expected to generalize to other settings.
- 501 • It is fine to include aspirational goals as motivation as long as it is clear that these  
502 goals are not attained by the paper.

### 503 2. Limitations

504 Question: Does the paper discuss the limitations of the work performed by the authors?

505 Answer: [\[Yes\]](#)

506 Justification: Section 6 contains a dedicated “Limitations” paragraph addressing four spe-  
507 cific limitations: coarse relation vocabulary, manual expert rubric, equal edge-type weight-  
508 ing, and single-seed evaluation.

509 Guidelines:

- 510 • The answer NA means that the paper has no limitation while the answer No means  
511 that the paper has limitations, but those are not discussed in the paper.
- 512 • The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not claim formal theorems. The Harmony metric (Eqs. 1–6) is defined compositionally; all component definitions and normalisation conventions are stated explicitly in Section 3.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix H lists all hyperparameters; Appendix A provides dataset statistics; Section 4 describes the experimental protocol; a single fixed seed ( $s = 42$ ) is used throughout.

Guidelines:

- The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: At submission time, an anonymised repository and a Zenodo draft artifact record are provided to support reproducibility. Upon acceptance, these assets will be de-anonymised and released under an open-source licence with a citable DOI.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- 621           • Providing as much information as possible in supplemental material (appended to the  
622           paper) is recommended, but including URLs to data and code is permitted.

623     **6. Experimental setting/details**

624     Question: Does the paper specify all the training and test details (e.g., data splits, hyper-  
625           parameters, how they were chosen, type of optimizer, etc.) necessary to understand the  
626           results?

627     Answer: [Yes]

628     Justification: Section 4 describes the experimental setup including baselines, evaluation  
629           protocol, and dataset split ratios. Appendix H lists all hyperparameters. Appendix A pro-  
630           vides dataset entity and edge count statistics.

631     Guidelines:

- 632           • The answer NA means that the paper does not include experiments.  
633           • The experimental setting should be presented in the core of the paper to a level of  
634           detail that is necessary to appreciate the results and make sense of them.  
635           • The full details can be provided either with the code, in appendix, or as supplemental  
636           material.

637     **7. Experiment statistical significance**

638     Question: Does the paper report error bars suitably and correctly defined or other appropri-  
639           ate information about the statistical significance of the experiments?

640     Answer: [No]

641     Justification: The ablation study (Appendix B) reports bootstrap 95% CIs, but the main  
642           results (Table 1) are single-seed without error bars. This limitation is explicitly acknowl-  
643           edged in the Limitations paragraph of Section 6; multi-seed evaluation is noted as future  
644           work.

645     Guidelines:

- 646           • The answer NA means that the paper does not include experiments.  
647           • The authors should answer "Yes" if the results are accompanied by error bars, confi-  
648           dence intervals, or statistical significance tests, at least for the experiments that support  
649           the main claims of the paper.  
650           • The factors of variability that the error bars are capturing should be clearly stated (for  
651           example, train/test split, initialization, random drawing of some parameter, or overall  
652           run with given experimental conditions).  
653           • The method for calculating the error bars should be explained (closed form formula,  
654           call to a library function, bootstrap, etc.).  
655           • The assumptions made should be given (e.g., Normally distributed errors).  
656           • It should be clear whether the error bar is the standard deviation or the standard error  
657           of the mean.  
658           • It is OK to report 1-sigma error bars, but one should state it. The authors should prefer-  
659           ably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of  
660           Normality of errors is not verified.  
661           • For asymmetric distributions, the authors should be careful not to show in tables or  
662           figures symmetric error bars that would yield results that are out of range (e.g. negative  
663           error rates).  
664           • If error bars are reported in tables or plots, The authors should explain in the text how  
665           they were calculated and reference the corresponding figures or tables in the text.

666     **8. Experiments compute resources**

667     Question: For each experiment, does the paper provide sufficient information on the com-  
668           puter resources (type of compute workers, memory, time of execution) needed to reproduce  
669           the experiments?

670     Answer: [Yes]

671     Justification: Appendix H reports the hardware (CPU-only, Apple M-series) and wall-clock  
672           time (approximately 10 minutes per domain for 20 generations). No GPU resources were  
673           used.

674 Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

683 **9. Code of ethics**

684 Question: Does the research conducted in the paper conform, in every respect, with the  
685 NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

686 Answer: [Yes]

687 Justification: No human subjects were involved. All knowledge graphs are curated from  
688 publicly available academic sources. No personally identifiable or scraped data is used.

689 Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

695 **10. Broader impacts**

696 Question: Does the paper discuss both potential positive societal impacts and negative  
697 societal impacts of the work performed?

698 Answer: [Yes]

699 Justification: Section 6 includes a “Broader impacts” paragraph discussing positive impacts  
700 (accelerating scientific theory discovery) and negative risks (LLM-generated claims  
701 may be plausible-sounding but factually incorrect, requiring expert validation before use in  
702 downstream scientific workflows).

703 Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

726 **11. Safeguards**

727 Question: Does the paper describe safeguards that have been put in place for responsible  
728 release of data or models that have a high risk for misuse (e.g., pretrained language models,  
729 image generators, or scraped datasets)?

730 Answer: [NA]

731 Justification: The paper does not release pretrained models or scraped datasets. The re-  
732 leased assets are small curated knowledge graphs and search-loop code, which pose no  
733 misuse risk.

734 Guidelines:

- 735 • The answer NA means that the paper poses no such risks.
- 736 • Released models that have a high risk for misuse or dual-use should be released with  
737 necessary safeguards to allow for controlled use of the model, for example by re-  
738quiring that users adhere to usage guidelines or restrictions to access the model or  
739 implementing safety filters.
- 740 • Datasets that have been scraped from the Internet could pose safety risks. The authors  
741 should describe how they avoided releasing unsafe images.
- 742 • We recognize that providing effective safeguards is challenging, and many papers do  
743 not require this, but we encourage authors to take this into account and make a best  
744 faith effort.

## 745 12. Licenses for existing assets

746 Question: Are the creators or original owners of assets (e.g., code, data, models), used in  
747 the paper, properly credited and are the license and terms of use explicitly mentioned and  
748 properly respected?

749 Answer: [Yes]

750 Justification: DistMult [14] and TransE [2] are cited. Core Python libraries (NumPy, scikit-  
751 learn) are BSD-3-Clause licensed. Knowledge graphs are original curated datasets.

752 Guidelines:

- 753 • The answer NA means that the paper does not use existing assets.
- 754 • The authors should cite the original paper that produced the code package or dataset.
- 755 • The authors should state which version of the asset is used and, if possible, include a  
756 URL.
- 757 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 758 • For scraped data from a particular source (e.g., website), the copyright and terms of  
759 service of that source should be provided.
- 760 • If assets are released, the license, copyright information, and terms of use in the pack-  
761 age should be provided. For popular datasets, [paperswithcode.com/datasets](http://paperswithcode.com/datasets) has  
762 curated licenses for some datasets. Their licensing guide can help determine the li-  
763 cense of a dataset.
- 764 • For existing datasets that are re-packaged, both the original license and the license of  
765 the derived asset (if it has changed) should be provided.
- 766 • If this information is not available online, the authors are encouraged to reach out to  
767 the asset's creators.

## 768 13. New assets

769 Question: Are new assets introduced in the paper well documented and is the documenta-  
770 tion provided alongside the assets?

771 Answer: [Yes]

772 Justification: Five curated KG datasets are documented in Appendix A with entity/edge  
773 counts, type vocabularies, and split ratios. The proposal schema is defined in Section 3  
774 with validation rules in Appendix C.

775 Guidelines:

- 776 • The answer NA means that the paper does not release new assets.

- 777           • Researchers should communicate the details of the dataset/code/model as part of their  
778           submissions via structured templates. This includes details about training, license,  
779           limitations, etc.  
780           • The paper should discuss whether and how consent was obtained from people whose  
781           asset is used.  
782           • At submission time, remember to anonymize your assets (if applicable). You can  
783           either create an anonymized URL or include an anonymized zip file.

784          **14. Crowdsourcing and research with human subjects**

785          Question: For crowdsourcing experiments and research with human subjects, does the pa-  
786          per include the full text of instructions given to participants and screenshots, if applicable,  
787          as well as details about compensation (if any)?

788          Answer: [NA]

789          Justification: No crowdsourcing or research with human subjects was conducted.

790          Guidelines:

- 791           • The answer NA means that the paper does not involve crowdsourcing nor research  
792           with human subjects.  
793           • Including this information in the supplemental material is fine, but if the main contri-  
794           bution of the paper involves human subjects, then as much detail as possible should  
795           be included in the main paper.  
796           • According to the NeurIPS Code of Ethics, workers involved in data collection, cura-  
797           tion, or other labor should be paid at least the minimum wage in the country of the  
798           data collector.

799          **15. Institutional review board (IRB) approvals or equivalent for research with human  
800           subjects**

801          Question: Does the paper describe potential risks incurred by study participants, whether  
802          such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)  
803          approvals (or an equivalent approval/review based on the requirements of your country or  
804          institution) were obtained?

805          Answer: [NA]

806          Justification: No human subjects research was conducted; IRB approval is not applicable.

807          Guidelines:

- 808           • The answer NA means that the paper does not involve crowdsourcing nor research  
809           with human subjects.  
810           • Depending on the country in which research is conducted, IRB approval (or equiva-  
811           lent) may be required for any human subjects research. If you obtained IRB approval,  
812           you should clearly state this in the paper.  
813           • We recognize that the procedures for this may vary significantly between institutions  
814           and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the  
815           guidelines for their institution.  
816           • For initial submissions, do not include any information that would break anonymity  
817           (if applicable), such as the institution conducting the review.

818          **16. Declaration of LLM usage**

819          Question: Does the paper describe the usage of LLMs if it is an important, original, or  
820          non-standard component of the core methods in this research? Note that if the LLM is used  
821          only for writing, editing, or formatting purposes and does not impact the core methodology,  
822          scientific rigorousness, or originality of the research, declaration is not required.

823          Answer: [Yes]

824          Justification: LLM-based proposal generation is a core methodological component de-  
825          scribed in Section 3. The specific model family (local Ollama-served model) and prompting  
826          strategy (entity-grounded, four-phase rotation: refine, combine, refine, novel) are detailed  
827          in Sections 3 and 4.

828          Guidelines:

- 829
- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- 830
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.
- 831
- 832