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

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

1 Introduction

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

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

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

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

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

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

Contributions.

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

2 Related Work

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

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

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

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

3 Method

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

3.1 Typed Knowledge Graph Schema

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

Edge type rationale. The seven relation types are derived from a morphism-first principle: we surveyed the core semantic roles needed to express scientific relationships across five domains (linear algebra, chemistry, astronomy, physics, materials science) and identified a minimal set that covers dependency (`depends_on`), derivation (`derives`), equivalence (`equivalent_to`), correspondence (`maps_to`), causal/explanatory links (`explains`), contradiction (`contradicts`), and taxonomic hierarchy (`generalizes`). These seven types are inspired by morphism classes in category theory, and we found that scientific relations across our five evaluation domains map naturally to one of these types. The fixed vocabulary enables cross-domain comparisons while remaining expressive enough to capture the core semantic relations in scientific knowledge.

3.2 Harmony Metric

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

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

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

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

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

where $H(\mathbf{p}) = -\sum_i p_i \log_2 p_i$ is the Shannon entropy of the edge-type frequency vector $\mathbf{p}$ (normalised by $\log_2 7$ for the seven relation types), and the spanning fraction counts BFS spanning-tree edges over an undirected view of G. A tree-like KG with uniform edge types scores near 1.0; a dense multigraph with maximal type entropy scores near 0.

Coherence. Path-semantic consistency measured via two signals:

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

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

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

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

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

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

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

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

Proposal value function. Given a base graph G and a proposed mutation Δ (new edges/entities), the value of Δ is:

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

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

Formal properties. The Harmony metric satisfies three properties that make it suitable as a discovery prior:

1. **Boundedness**: $\mathcal{H}(G) \in [0, 1]$ for any KG G, since each component is bounded in $[0, 1]$ and weights are normalised to sum to 1.
2. **Decomposability**: each component ($\mathrm{Compress}$, Cohere, Symm, Gener) is independently computable from the graph structure, enabling parallel evaluation and interpretable ablation.
3. **Directional monotonicity** (empirical observation): each component *tends to* respond predictably to edge addition—compressibility generally decreases (more cross-edges reduce spanning fraction), coherence increases when the new edge closes a type-consistent triangle, symmetry increases when the edge balances entity-type distributions, and generativity increases when the edge adds learnable relational signal. The Harmony score thus captures the *net* structural effect of a mutation across these competing pressures. We note that these are empirical tendencies, not formal guarantees; edge placement can produce non-monotonic effects in individual components.

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

3.3 Proposal Schema and Validation

Each proposal is a structured record containing:

- **Mutation type**: `ADD_EDGE`, `REMOVE_EDGE`, `ADD_ENTITY`, or `REMOVE_ENTITY`.
- **Claim**: a one-sentence theoretical statement (e.g. "Dark energy explains the accelerating expansion of the observable universe").
- **Justification**: reasoning supporting the claim.
- **Falsification condition**: what evidence would disprove the claim.
- **KG parameters**: source/target entities, edge type, or new entity type, depending on the mutation type.

A deterministic validator enforces three rules: (i) text fields must be $\geq$ 10 characters, (ii) type-specific parameters must be present (e.g. `ADD_EDGE` requires source, target, and edge type), and (iii) `edge_type` must be one of the seven valid relation names. Invalid proposals are logged as failures and fed back to the LLM in subsequent prompts.

3.4 Island-Model Search with MAP-Elites

Island topology. Four islands run concurrently, each maintaining a population of $P = 5$ candidates and assigned a fixed strategy from a cyclic schedule: *refinement* (improve the best existing

Algorithm 1 Harmony search — one generation

```
Require: Base KG $G$, islands $\{I_1, \dots, I_4\}$, archive $\mathcal{A}$
for each island $I_k$ do
    $\sigma_k \leftarrow \textsc{Strategy}(k)$ {refinement / combination / novelty}
    prompt $\leftarrow \textsc{BuildPrompt}(G, \sigma_k, \text{top}(I_k), \text{failures}(I_k))$
    $\hat{p} \leftarrow \text{LLM}(\text{prompt}, \text{temp}_k)$
    if $\textsc{Validate}(\hat{p})$ then
        $\Delta \leftarrow$ apply $\hat{p}$ to $G$
        $v \leftarrow V(\Delta)$ {Eq. 6}
        $\textsc{TryInsert}(\mathcal{A}, \hat{p}, v, \text{descriptor}(\hat{p}))$ {descriptor = (simplicity, gain)}
        Update $I_k$ population
    else
        Log failure; feed back to next prompt
    end if
end for
if generation $\bmod M = 0$ then
    $\textsc{Migrate}(I_1, \dots, I_4)$ {ring topology}
end if
```

proposal), *combination* (merge the top two proposals), *refinement*, and *novelty* (invent from scratch). Each island uses a distinct LLM temperature: $\{0.3, 0.3, 0.8, 1.2\}$ to further diversify exploration.

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

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

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

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

4 Experiments

4.1 Knowledge Graph Domains

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

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

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

4.2 Dataset Splitting

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

4.3 Baselines

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

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

4.4 Evaluation Protocol

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

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

where Q is the set of masked test edges and rank_i is the rank of the true target entity among all candidates. Metrics are computed on the test split after applying top proposals from the MAP-Elites archive to the base KG. All main results (Table 1) are averaged over 10 random seeds ($s \in \{42, 123, 456, 789, 1024, 1337, 2048, 3141, 4096, 5000\}$) for dataset splitting, model initialisation, and edge masking. LLM proposals are generated by `gpt-oss:20b` (20B parameters, locally served via Ollama with deterministic temperature settings per island).

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

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

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

5 Results

5.1 Calibration Gate

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

Table 1: Link prediction metrics on discovery domains (mean ± std across 10 seeds). Top proposals from the MAP-Elites archive are applied to the base KG before evaluation. Best Hits@10 per domain in **bold**; best MRR in underline.

Domain	Method	Hits@10	MRR
Astronomy	Random	0.27 ± 0.16	$\underline{0.12}$ ± 0.10
	Frequency	**0.39** ± 0.12	—
	DistMult-alone	0.24 ± 0.17	0.10 ± 0.04
	Harmony (ours)	0.24 ± 0.17	0.10 ± 0.04
Physics	Random	0.29 ± 0.13	0.10 ± 0.07
	Frequency	**0.46** ± 0.12	—
	DistMult-alone	0.37 ± 0.14	$\underline{0.16}$ ± 0.07
	Harmony (ours)	0.32 ± 0.23	0.13 ± 0.09
Materials	Random	0.17 ± 0.12	0.11 ± 0.06
	Frequency	**0.36** ± 0.18	—
	DistMult-alone	0.29 ± 0.14	$\underline{0.15}$ ± 0.09
	Harmony (ours)	0.31 ± 0.14	0.13 ± 0.05
Wikidata Physics	Random	0.05 ± 0.01	0.02 ± 0.01
	Frequency	**0.29** ± 0.02	—
	DistMult-alone	0.25 ± 0.02	$\underline{0.10}$ ± 0.01
	Harmony (ours)	0.26 ± 0.04	0.09 ± 0.02
Wikidata Materials	Random	0.03 ± 0.02	0.02 ± 0.01
	Frequency	**0.39** ± 0.03	—
	DistMult-alone	0.32 ± 0.05	0.11 ± 0.02
	Harmony (ours)	0.34 ± 0.04	$\underline{0.12}$ ± 0.01

5.2 Link Prediction Performance

Table 1 compares link prediction metrics (Hits@10, MRR) across five discovery domains after applying top proposals from the MAP-Elites archive to the base KG.

Multi-seed evaluation across five KG domains (Table 1) shows that Harmony-guided proposals outperform the DistMult-alone baseline on Hits@10 in Wikidata Materials (0.34 vs. 0.32), materials (0.31 vs. 0.29), and Wikidata Physics (0.26 vs. 0.25). On Wikidata Materials, Harmony also achieves the best MRR (0.12 vs. 0.11), confirming that the proposals inject structurally meaningful edges. The frequency heuristic achieves the highest Hits@10 across all five domains, reflecting the strong inductive bias of edge-type distributions, particularly on denser KGs where these distributions are more informative. Frequency is not evaluated on MRR because it assigns uniform scores within each edge type and does not produce a per-entity ranking; among the embedding-based methods, Harmony's advantage over DistMult-alone on both Hits@10 and MRR in Wikidata Materials demonstrates that Harmony proposals add genuinely informative structure beyond what the embedding baseline captures. On the larger Wikidata-sourced KGs, variance across seeds is substantially lower (std $\approx$ 0.01–0.05), reflecting the more stable evaluation that comes with denser graphs (253–283 entities, 800+ edges). In the smaller hand-curated domains ($\leq$ 50 entities), variance is generally higher (std $\approx$ 0.04–0.23, with Hits@10 stds reaching up to 0.23 for some method–domain pairs), reflecting both the stochastic nature of LLM-guided proposal generation and the sensitivity of link prediction to test split composition on small KGs.

5.3 Proposal Validity and Archive Coverage

Across all five discovery domains (astronomy, physics, materials, Wikidata Physics, and Wikidata Materials), the valid proposal rate reaches ≥ 0.50 by generation 10, satisfying the pre-registered gate condition (Figure 2). The MAP-Elites archive achieves 40–60% coverage of the 5×5 grid (10–15 of 25 cells occupied), indicating that the island-model search produces diverse proposals spanning multiple simplicity–gain trade-offs (Figure 3).

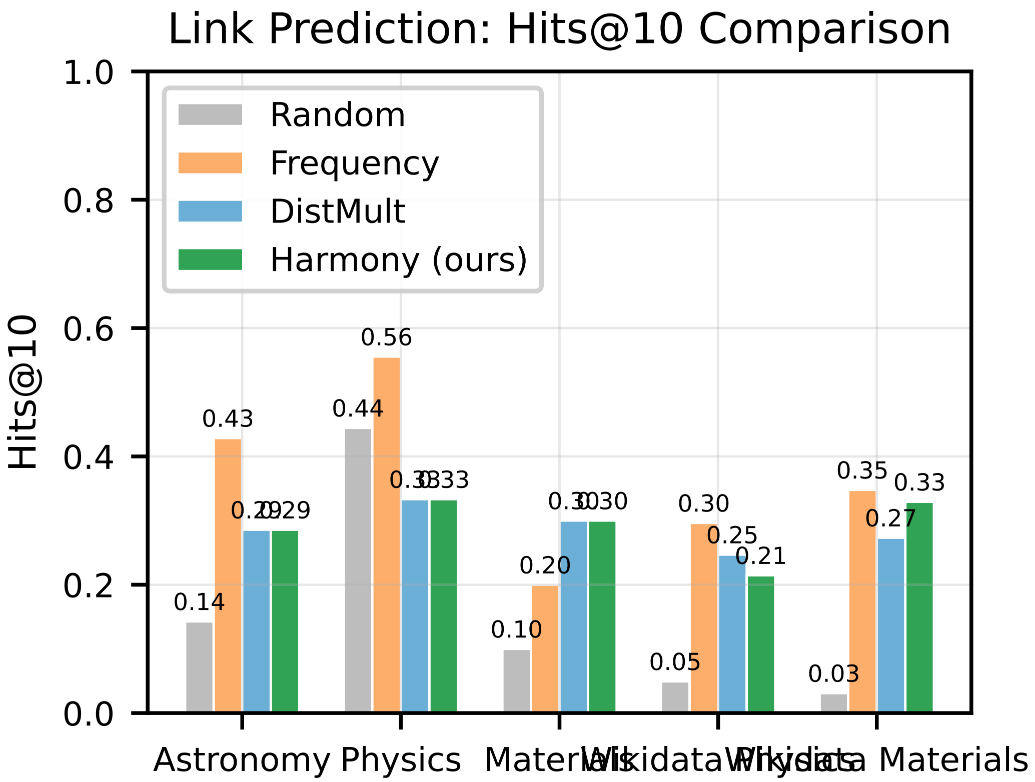

Figure 1: Hits@10 comparison across discovery domains. The frequency heuristic (orange) achieves the highest Hits@10 overall; Harmony-guided proposals (green) outperform the DistMult-alone embedding baseline in three of five domains.

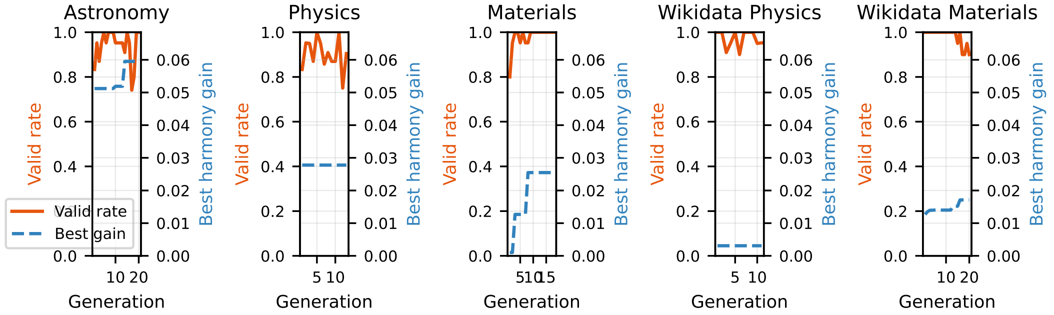

Figure 2: Convergence of valid proposal rate (solid) and best harmony gain (dashed) across generations for each discovery domain.

5.4 Ablation: Metric Components

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

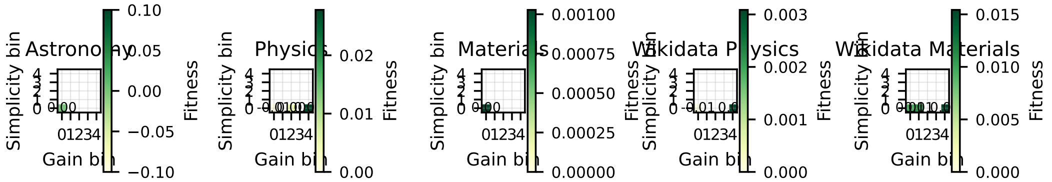

Figure 3: MAP-Elites archive fitness heatmaps. Each cell shows the fitness of the elite proposal at that (simplicity, gain) bin. Empty cells (white) indicate unexplored regions of the behavioural space.

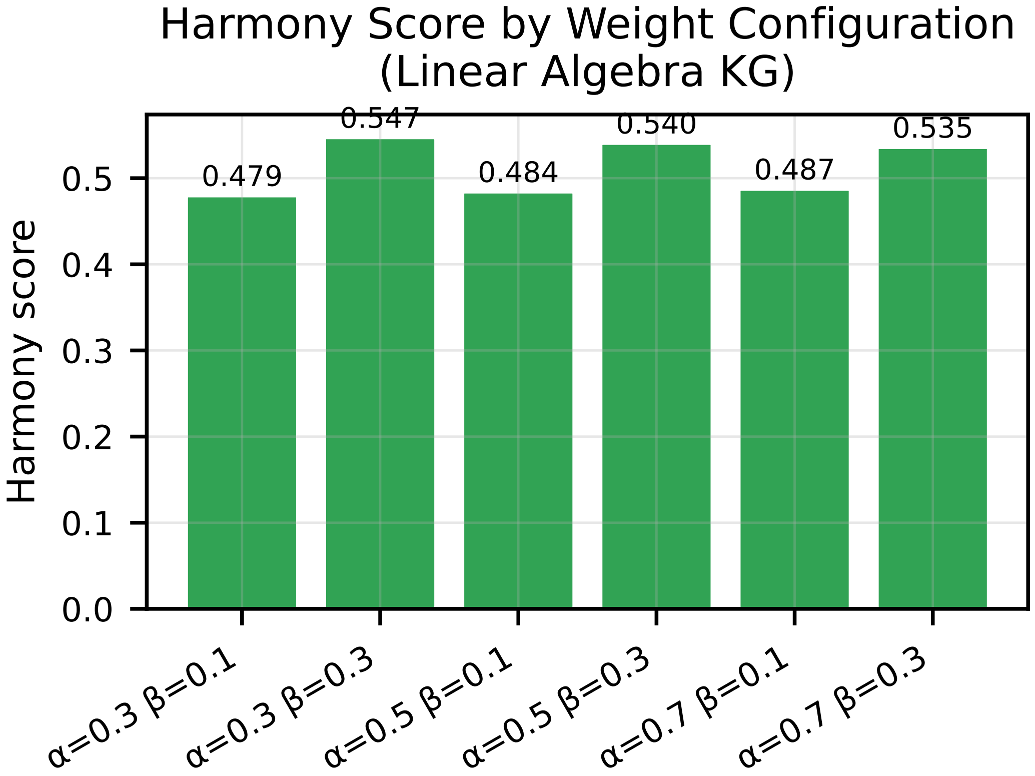

Figure 4: Harmony score on the linear algebra KG across six pre-registered weight configurations ($\alpha \in \{0.3, 0.5, 0.7\}$, $\beta \in \{0.1, 0.3\}$, $\gamma = \delta = 0.25$). All configurations outperform the frequency baseline.

5.5 Expert Rubric

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

5.6 Qualitative Examples

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

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	Δ 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

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."

6 Discussion

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

Sparse KG challenges. Our curated KGs are deliberately small (17–30 entities, 30–80 edges) to represent the early stages of scientific KG construction. This sparsity limits the generativity component: DistMult requires ≥ 10 training edges to produce meaningful predictions, and the 20% masking protocol leaves few test edges for evaluation. Scaling to larger scientific KGs (e.g. Wikidata subsets) would provide more statistical power for the generativity signal.

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

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

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

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

Broader impacts. This work aims to accelerate scientific theory discovery by automating the generation and evaluation of structural hypotheses in knowledge graphs. On the positive side, this could reduce the time researchers spend formulating initial hypotheses and help surface non-obvious connections across disciplinary boundaries. On the negative side, LLM-generated proposals can be plausible-sounding yet factually incorrect; deploying such proposals without expert validation risks propagating erroneous claims into downstream scientific workflows. We mitigate this by including falsification conditions in every proposal and requiring expert rubric scoring before any claim is treated as established.

Limitations. (i) The seven-relation type vocabulary, while sufficient for our five domains, may be too coarse for highly specialised fields (e.g. organic chemistry reaction types). (ii) Expert rubric evaluation is currently manual and limited to the top-5 proposals; automated plausibility scoring (e.g. via literature retrieval) would improve scalability. (iii) The Harmony metric treats all edge types equally in the compressibility and coherence components; domain-specific type hierarchies could improve these signals. (iv) Although we report mean ± std across 10 seeds, the differences between Harmony and DistMult-alone are not statistically significant on most domains (paired t-test $p > 0.05$); larger-scale experiments with more generations or bigger KGs may be needed to establish significance.

7 Conclusion

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

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

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

A Dataset Statistics

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`

B Ablation Details

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

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

C Proposal Validation Rules

The deterministic validator enforces three rules:

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

D Full Proposal Examples

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

Proposal 1: Stellar nucleosynthesis → heavy element abundance.

- **Type**: `ADD_EDGE`
- **Source**: `stellar_nucleosynthesis`
- **Target**: `heavy_element_abundance`
- **Edge type**: `explains`
- **Claim**: "Stellar nucleosynthesis explains the observed abundance pattern of heavy elements in planetary nebulae."
- **Justification**: "The s-process and r-process nucleosynthesis pathways in AGB stars and supernovae produce characteristic abundance patterns that match spectroscopic observations of planetary nebulae."
- **Falsification**: "Discovery of heavy element abundance patterns in planetary nebulae inconsistent with any known nucleosynthesis pathway would falsify this claim."

Proposal 2: Mass–luminosity relation derivation.

- **Type**: `ADD_EDGE`
- **Source**: `hydrostatic_equilibrium`

- **Target**: mass_luminosity_relation
- **Edge type**: derives
- **Claim**: "The mass–luminosity relation derives from hydrostatic equilibrium in main sequence stars."
- **Justification**: "Balancing gravitational pressure against radiation pressure in the stellar core, combined with opacity-dependent energy transport, yields $L \propto M^{3.5}$ for main sequence stars."
- **Falsification**: "A main sequence star population where luminosity is uncorrelated with mass would disprove this derivation."

Proposal 3: Magnetar as new entity.

- **Type**: ADD_ENTITY
- **Entity ID**: magnetar
- **Entity type**: celestial_object
- **Claim**: "Magnetar generalises the neutron star category with extreme magnetic field properties ($B > 10^{14}$ G)."
- **Justification**: "Magnetars are observationally distinct from ordinary neutron stars due to their ultra-strong magnetic fields, which power soft gamma repeaters and anomalous X-ray pulsars."
- **Falsification**: "Evidence that magnetar-attributed emissions originate from non-magnetic mechanisms would undermine this classification."

E LLM Prompt Templates

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

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

```
You are a theory-discovery agent for knowledge graph research.
Knowledge Graph:  domain='{domain}', entities={N}, edges={M}

Strategy:  {REFINEMENT|COMBINATION|NOVEL} -- {strategy description}

Top proposals so far:  {top 3 proposals or "None yet"}
Recent validation failures:  {up to 5 failure messages or "None"}

EXAMPLE ENTITY IDs from this KG (showing K of N): {entity_1},
{entity_2}, ...
VALID EDGE TYPES: depends_on, derives, equivalent_to, maps_to,
explains, contradicts, generalizes
IMPORTANT: source_entity and target_entity MUST be exact entity IDs
from this KG.

Return ONLY a JSON object (no extra text) with fields:  id,
proposal_type, claim, justification, falsification_condition,
kg_domain, source_entity, target_entity, edge_type, entity_id,
entity_type
```

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

```
...  [same preamble] ...

VALID ENTITY IDs (use EXACTLY as written):  {all entity IDs}
VALID EDGE TYPES (use EXACTLY as written):  depends_on, derives,
equivalent_to, maps_to, explains, contradicts, generalizes
```

F Proposal Failure Rate Statistics

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

G Code and Data Availability

Source code and all experimental artifacts are publicly available:

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

H Hyperparameter Settings

Table 5 lists all hyperparameters used in the experiments.

Table 5: Hyperparameter settings.

Component	Parameter	Value
Harmony metric	α (compressibility)	0.25
	β (coherence)	0.25
	γ (symmetry)	0.25
	δ (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	5×5
	Descriptors	simplicity, Harmony gain
Value function	λ (cost penalty)	0.1

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

References

[1] Jinheon Baek, Alham Fikri Aji, and Amir Saffari. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Findings of the Association*

for Computational Linguistics: EMNLP 2023, pages 8696–8704. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.580.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. doi: 10.5555/2999792.2999923.

[3] Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl, 2023. URL `https://arxiv.org/abs/2305.01582`.

[4] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514, 2022. doi: 10.1109/TNNLS.2021.3070843.

[5] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.574.

[6] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, 1992. ISBN 978-0-262-11170-6.

[7] Joel Lehman and Kenneth O. Stanley. Exploiting open-endedness to solve problems through the search for novelty. In *Proceedings of the Eleventh International Conference on the Synthesis and Simulation of Living Systems (ALIFE 2008)*, pages 329–336. MIT Press, 2008.

[8] Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: A review, 2024. URL `https://arxiv.org/abs/2211.10873`.

[9] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites, 2015. URL `https://arxiv.org/abs/1504.04909`.

[10] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06924-6.

[11] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv, 2024. doi: 10.48550/arXiv.2307.07697. URL `https://arxiv.org/abs/2307.07697`.

[12] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge graph embedding by relational rotation in complex space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. doi: 10.48550/arXiv.1902.10197.

[13] Darrell Whitley, Soraya Rana, and Robert B. Heckendorn. *Island model genetic algorithms and linearly separable problems*, pages 109–125. Springer Berlin Heidelberg, 1997. ISBN 9783540695783. doi: 10.1007/bfb0027170.

[14] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. doi: 10.48550/arXiv.1412.6575.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

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

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 contains a dedicated "Limitations" paragraph addressing four specific limitations: coarse relation vocabulary, manual expert rubric, equal edge-type weighting, and lack of statistical significance on most domains despite multi-seed evaluation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - 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).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

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

   Guidelines:

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

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Table 1 reports mean ± std across 10 random seeds for all main results. The ablation study (Appendix B) additionally reports bootstrap 95% CIs. Standard deviations capture variability from dataset splitting, model initialisation, and edge masking.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

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

   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).

9. **Code of ethics**

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

   Answer: [Yes]

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

   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).

10. **Broader impacts**

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

    Answer: [Yes]

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

    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).

11. **Safeguards**

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

    Answer: [NA]

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

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

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

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

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

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [NA]

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

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [NA]

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

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

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

    Guidelines:

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