Enriching Scientific Knowledge Graph with Entropy-driven Progressive Self-Feedback Fusion (Technical Report)

Runhao Zhao
National University of Defense Technology
Changsha, China
runhaozhao@nudt.edu.cn

Weixin Zeng
National University of Defense Technology
Changsha, China
zengweixin13@nudt.edu.cn

Zhengpin Li
The Center for machine learning research, Peking University
Beijing, China
zpli@pku.edu.cn

Wentao Zhang
The Center for machine learning research, Peking University
Beijing, China
wentao.zhang@pku.edu.cn

Jiuyang Tang
National University of Defense Technology
Changsha, China
jiuyang_tang@nudt.edu.cn

Xiang Zhao
National University of Defense Technology
Changsha, China
xiangzhao@nudt.edu.cn

Abstract

To support the reproducibility and scientific rigor of "Enriching Scientific Knowledge Graph with Entropy-driven Progressive Self-Feedback Fusion", this technical report provides essential background and extended results. Key contents include in-depth descriptions of scientific knowledge fusion tasks, precise LLM prompt engineering strategies, and a broader array of experimental results across diverse scientific domains. This document acts as a primary reference for the implementation details and the expanded empirical evidence of the Self-Fusion methodology.

Keywords

AI for Science; Scientific Knowledge Graph Fusion; Entropy-driven Progressive Self-Feedback

A Methodology: Implementation Details

A.1 Entropy-driven Validity Estimation via Token Probability

In Section 4.3, we introduced the validity probability $P((v, r, \hat{v}) \mid \mathcal{G}^s_{local}(u^s_i))$ to filter high-entropy candidates. To ensure scientific rigor and avoid hallucinated confidence scores often associated with direct numeric output from LLMs, we utilize the model's intrinsic uncertainty via *output token log-probabilities*.

For a candidate fact $f \in C^g_{meta}(u^s_i)$ and the local context $C_{ctx} = \mathcal{G}^s_{local}(u^s_i)$, we construct a validity query prompt $\mathbf{x}$. We task the LLM to predict a binary verification token $y \in \{\text{"Yes"}, \text{"No"}\}$. The validity probability is calculated as the softmax-normalized probability of the positive token using the access to raw logits:

$$P(f \mid \mathcal{G}^s_{local}(u^s_i)) = \frac{\exp(\ell(\text{"Yes"}))}{\exp(\ell(\text{"Yes"})) + \exp(\ell(\text{"No"}))} \tag{1}$$

where $\ell(\cdot)$ denotes the logit value of the target token given the context $\mathbf{x}$.

Furthermore, during the *Fusion Scene Generation* (Section 4.3.2), we quantify the **Generation Entropy** to measure the model's hesitation in grounding the graph structure into natural language. Given the generated description sequence $\mathbf{y} = (y_1, \ldots, y_T)$, the sequence-level entropy is computed as: $\mathcal{H}_{gen}(\mathbf{y} \mid F_{new}) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{w \in \mathcal{V}} P(w \mid y_{<t}, F_{new}) \log P(w \mid y_{<t}, F_{new})$, where $P(w \mid \cdot)$ is the probability distribution over the vocabulary $\mathcal{V}$ at step t. A lower $\mathcal{H}_{gen}$ indicates that the fused scientific facts F_{new} are deterministic enough to induce a coherent and unambiguous scientific description. Facts resulting in $\mathcal{H}_{gen} > \tau_{ent}$ are flagged as high-entropy noise and discarded.

A.2 Core Prompt Details

To ensure reproducibility, we detail the core prompts used in the S elf-Fusion framework. We define a standardized prompt template to simulate scientific peer review.

Prompt 1: Validity Verification (Entropy Filter)

Role: You are an expert scientist in [Target Domain, e.g., Molecular Biology].
Context: We have a specialized Scientific Knowledge Graph (SKG) with the following local mechanism:
`{Local_SKG_Structure}`
Task: We have mined a candidate fact from a General Knowledge Graph:
`{Candidate_Fact}`
Instruction: Determine if this candidate fact is scientifically rigorous enough to be integrated into the SKG context.

(1) Does it align with the specific granularity of the context? (e.g., specific protein pathways vs. general associations).
(2) Does it contradict existing rigorous mechanisms?

Output: Answer strictly with "Yes" or "No".

Prompt 2: Scene Generation (Graph → Text)

Input: A set of scientific facts (Triples):
`{Fused_Facts_Set}`
Instruction: Synthesize these facts into a coherent, rigorous scientific abstract. The text must logically connect the entities and reflect the causal mechanisms implied by the relations. If the facts are disconnected or ambiguous, explicitly acknowledge the uncertainty in the text.

Algorithm 1: Entropy-driven Progressive Self-Feedback Fusion

Input : Scientific KG G^s, General KG G^g, Thresholds τ, θ
Output : Fused Facts F_{final}

Initialize $F_{final} \leftarrow \emptyset$
Transform G^s, G^g to Meta-knowledge Line Graphs $\mathcal{G}^s, \mathcal{G}^g$
for *each fact node* $u_i^s \in \mathcal{G}^s$ **do**
  /* Phase 1: Fuzzy Retriever (Entropy Maximization) */
  Compute Semantic Candidates $C_{node}^{sema}(u_i^s)$
  Compute Structural Candidates $C_{meta}^{g}(u_i^s)$
  /* Phase 2: Progressive Fusion (Entropy Reduction) */
  Initialize feedback constraints $\mathcal{I}_{neg} \leftarrow \emptyset$
  $F_{new} \leftarrow \emptyset$
  while *not converged* ***and*** $t < MaxIter$ **do**
    // Step 2.1: Entropy-driven Filtering
    for $f \in C_{meta}^{g}(u_i^s)$ **do**
      Calculate $P(f \mid \mathcal{G}_{local}^s, \mathcal{I}_{neg})$ via Token Logprobs
      if $P(f) \geq \tau$ **then**
        $F_{new} \leftarrow F_{new} \cup \{f\}$
    // Step 2.2: Scene Generation (Graph → Text)
    $\mathcal{S}_{desc} \leftarrow \mathrm{SceneGen}(F_{new}, \mathcal{G}_{local}^s)$
    // Step 2.3: Reconstruction (Text → Graph)
    $F_{recon} \leftarrow \mathrm{Recon}(\mathcal{S}_{desc})$
    // Step 2.4: Cycle Consistency Check
    $F_{valid} \leftarrow F_{recon} \cap F_{new}$
    $F_{mismatch} \leftarrow F_{recon} \setminus F_{new}$
    if $F_{mismatch} == \emptyset$ **then**
      $F_{final} \leftarrow F_{final} \cup F_{valid}$
      break // Entropy minimized
    else
      $\mathcal{I}_{neg} \leftarrow \mathrm{UpdateConstraints}(F_{mismatch})$
      $\tau \leftarrow \tau + \delta$ // Increase strictness
return F_{final}

Prompt 3: Reconstruction (Text → Graph)

Input: A scientific abstract:
`{Generated_Scene_Text}`
Instruction: Extract all deterministic scientific facts from the text. Return them strictly as triples (Head, Relation, Tail). Do not infer information not present in the text.

A.3 Self-Fusion Workflow

The overall training procedure of Self-Fusion is summarized in Algorithm 1. The framework alternates between the fuzzy retrieval phase (Exploration) and the progressive self-feedback phase (Crystallization).

Table B1: Statistics of the `SciFusion-Bench`. "$\#F_{fused}$": Fused facts from GKG to SKG. "*Str. Sim.*": average entity neighbor structure similarity[9]. "*#Rel. O.*": The proportion of overlapping relations in SKG and GKG.

Cluster	Dataset	Domain	$\#F_{fused}$	Str. Sim. ⇓	Rel. O. ⇓
LS	`SKGF(W-Bio)`	Biomedical	12,430	8.2%	0%
	`SKGF(W-Plant)`	Botany	8,120	13.2%	0%
NS	`SKGF(W-Mat)`	Materials	11,440	5.7%	0%
SHS	`DKGF(W-I)`	Politics	796,254	15.4%	0%
	`DKGF(Y-I)`	Politics	451,158	14.0%	0%
	`SKGF(W-Music)`	Musicology	5,057	16.8%	0%

B Supplementary Experiments

In this section, we present a comprehensive evaluation of our framework. We begin by detailing the experimental setup, followed by an in-depth analysis of the results to demonstrate the superiority of the proposed framework in multiple scientific scenarios.

B.1 Experimental Settings

B.1.1 Benchmark Datasets. To evaluate the robustness of Self-Fusion across diverse scientific disciplines, we construct `SciFusion-Bench`, a multi-domain benchmark encompassing six standardized datasets. As illustrated in Table B1, each dataset follows the prior criteria [29], where general knowledge from Wikipedia (via Wikidata or YAGO) is selectively integrated into a specialized SKG. The benchmark is categorized into three major scientific clusters:

- **Life Science (LS):** This cluster covers biological systems and molecular mechanisms. `SKGF(W-Bio)` integrates general molecular attributes and genetic metadata from Wikidata into PrimeKG [4], a precision medicine KG. `SKGF(W-Plant)` fuses Wikipedia's botanical descriptions into the specialized taxonomy and trait records of AgroLD [19].
- **Physical Science (PS):** This cluster emphasizes the properties of physical matter. `SKGF(W-Mat)` identifies chemical nomenclature and physical constants from Wikidata to enrich the experimental crystal structure data in the materials project [14].
- **Social & Humanistic Science (SHS):** This cluster captures human events and culture. It includes `DKGF(W-I)` and `DKGF(Y-I)`, which selectively fuse background entity facts from Wikidata/YAGO into the temporal political event streams of ICEWS [29]. `SKGF(W-Music)` integrates artist biographies from Wikipedia into the musicology structures of MusicBrainz [12, 23].

Detailed statistics of the entities and fused facts are provided in Table B1.

B.2 Dataset Construction Details

To ensure the reproducibility of `SciFusion-Bench` and to facilitate future research in AI for Science, we provide a detailed elaboration on the dataset construction protocol. Following the methodology proposed in [29], our construction process adheres to a rigorous pipeline: *Source Selection → Entity Alignment → Scientific Relevance Filtering → Fusion Set Generation.*

Table B2: Main results on `DKGF(W-I)` and `DKGF(Y-I)`. The best ACC/F1 results are highlighted in *bold*, while the runner-up results are *underlined*. " *Sema., Struc., LLM.*" indicate the use of semantics, structural information, and LLMs, respectively.

Benchmark Configurations		Settings			DKGF(W-I)-S1				DKGF(Y-I)-S1				DKGF(W-I)-S2				DKGF(Y-I)-S2			
		Sema.	*Struc.*	*LLM.*	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1
General-purpose																				
Rule.	StringMatch-F	✓			0.496	0.160	0.002	0.004	0.498	0.286	0.003	0.005	0.498	0.192	0.002	0.003	0.498	0.308	0.003	0.005
	TF-IDF-F	✓			0.498	0.461	0.025	0.047	0.507	0.598	0.043	0.081	0.498	0.451	0.023	0.043	0.508	0.598	0.046	0.086
Trans.	TransE-F	✓	✓		0.649	0.668	0.596	0.630	0.588	0.782	0.244	0.372	0.637	0.662	0.560	0.607	0.575	0.772	0.212	0.332
	TransH-F	✓	✓		0.640	0.683	0.520	0.591	0.583	0.764	0.240	0.365	0.632	0.685	0.489	0.570	0.568	0.768	0.195	0.311
	DistMult-F	✓	✓		0.554	0.547	0.629	0.585	0.525	0.524	0.548	0.536	0.536	0.532	0.590	0.560	0.532	0.530	0.562	0.545
	ComplEx-F	✓	✓		0.553	0.546	0.632	0.586	0.526	0.525	0.567	0.545	0.544	0.539	0.609	0.572	0.515	0.514	0.546	0.530
GNN.	GCN-F		✓		0.485	0.487	0.572	0.526	0.501	0.501	0.793	0.614	0.493	0.494	0.582	0.534	0.498	0.499	0.796	0.613
	TransGNN-F	✓	✓		0.490	0.491	0.505	0.498	0.465	0.465	0.477	0.471	0.486	0.487	0.507	0.497	0.468	0.468	0.481	0.474
	Graph-Memba-F	✓	✓		0.485	0.487	0.572	0.526	0.503	0.502	0.794	0.615	0.493	0.494	0.582	0.534	0.499	0.501	0.796	0.615
Generative.	BERT-F	✓			0.532	0.531	0.549	0.540	0.586	0.599	0.523	0.558	0.530	0.531	0.507	0.519	0.569	0.561	0.631	0.594
	ICL-F	✓		✓	0.550	0.553	0.528	0.540	0.583	0.573	0.649	0.609	0.494	0.494	0.523	0.508	0.488	0.489	0.523	0.505
	Self-Consistency-F	✓	✓	✓	0.592	0.607	0.522	0.561	0.590	0.584	0.621	0.602	0.531	0.533	0.507	0.520	0.566	0.559	0.628	0.591
	Self-RAG-F	✓	✓	✓	0.576	0.589	0.506	0.544	0.594	0.623	0.477	0.540	0.544	0.544	0.545	0.545	0.577	0.595	0.481	0.532
Cross-task Adaptation																				
EA.	SimpleHHEA-F	✓	✓		0.490	0.492	0.507	0.499	0.493	0.493	0.503	0.498	0.492	0.492	0.507	0.500	0.477	0.477	0.481	0.479
	ChatEA-F	✓	✓	✓	0.596	0.551	0.581	0.566	0.649	0.727	0.477	0.576	0.592	0.611	0.507	0.554	0.592	0.619	0.481	0.541
KGC.	KG-BERT-F	✓	✓		0.523	0.522	0.546	0.534	0.516	0.515	0.531	0.523	0.507	0.507	0.515	0.511	0.503	0.503	0.481	0.492
	KG-LLaMA-F	✓	✓	✓	0.512	0.512	0.531	0.521	0.539	0.537	0.558	0.547	0.529	0.527	0.529	0.528	0.521	0.521	0.514	0.517
	KoPA-F	✓	✓	✓	0.558	0.556	0.577	0.566	0.561	0.556	0.602	0.578	0.549	0.547	0.564	0.556	0.541	0.543	0.522	0.532
	PRGC-F	✓	✓		0.503	0.503	0.508	0.505	0.507	0.507	0.515	0.511	0.504	0.504	0.507	0.506	0.503	0.503	0.481	0.492
RTE.	NoGen-BART-F	✓	✓		0.516	0.515	0.531	0.523	0.511	0.511	0.523	0.517	0.508	0.508	0.517	0.512	0.509	0.509	0.491	0.500
	NoGen-T5-F	✓	✓		0.510	0.510	0.522	0.516	0.507	0.507	0.515	0.511	0.509	0.509	0.519	0.514	0.504	0.504	0.509	0.506
D.	ExeFuse	✓	✓		0.680	0.673	0.708	0.690	0.661	0.615	0.717	0.662	0.655	0.627	0.683	0.654	0.633	0.634	0.682	0.657
	Self-Fusion (Ours)	✓	✓	✓	**0.779**	0.877	0.649	**0.746**	**0.750**	0.815	0.647	**0.721**	**0.753**	0.878	0.586	**0.703**	**0.691**	0.664	0.772	**0.714**

B.2.1 Construction Protocol. **Phase 1: Domain-Specific Source Selection.** We selected three representative authoritative scientific knowledge bases as the backbone for our SKGs:

- **Life Science (PrimeKG):** A precision medicine graph focusing on drug-disease-phenotype interactions. We utilize the subgraph related to *pharmacokinetics*.
- **Physical Science (Materials Project):** A database containing computed properties of inorganic crystals. We focus on the *crystal structure* and *thermodynamic stability* subsets.
- **Social Science (ICEWS):** As described in Section B, we use the temporal political event graph.

Phase 2: Cross-Graph Entity Alignment. To bridge the specialized SKGs with General KGs (Wikidata/YAGO), we performed Entity Alignment (EA). Unlike traditional EA tasks that assume a 1-to-1 mapping, we adopted a *high-precision filtering strategy*. We used curated scientific identifiers (e.g., PubChem CID for chemicals, DOI for papers, GeoNames for locations) to establish rigid anchor links. For entities lacking explicit IDs, we employed a strict string matching protocol constrained by type consistency (e.g., a node in SKG must be of type "Protein" to match a Wikidata node).

Phase 3: Iterative Degree-based Masking (IDS). To rigorously simulate data scarcity in scientific domains, we employed the *Iterative Degree-based Sampling (IDS)* strategy [29]. Specifically, we systematically masked 50% of the aligned entities and their **incident triples** within the SKG. This procedure establishes a challenging "incompleteness gap", thereby compelling the model to actively retrieve and fuse external knowledge to restore scientific connectivity.

Phase 4: Scientific Relevance Filtering. A naive integration of Wikidata introduces massive noise (e.g., a scientist's birth place is irrelevant to their research output). To construct the ground truth F_{fused}, we manually defined a set of *scientifically valid relation types* for each domain (e.g., for `SKGF(W-Mat)`, we strictly retained relations like `has_band_gap`, `crystal_system`, excluding generic ones like `instance_of`).

B.2.2 Benchmark Configurations. Since SKGF is a nascent task with no pre-existing specialized scientific solutions, a primary contribution of this work is to establish a comprehensive and standardized evaluation suite to facilitate future research. Following [29], we systematically adapted 22 representative methods from related fields (e.g., KGC, EA) to the SKGF setting using a unified fusion scoring protocol. These configurations are categorized as follows:

- **General-purpose Configurations.** To establish benchmark performance using universal architectures, this category employs broadly applicable methods (e.g., TransE, GNN, BERT, LLMs) rather than task-specific designs. This category of configurations is mainly based on current advanced and classic general methods, including rule-based methods (i.e., "Rule."), such as StringMatch-F [7] and TF-IDF-F [16]; Translation-based methods (i.e., "Trans.") such as TransE-F [3], TransH-F [17], DistMult-F [22], and ComplEx-F [18]; GNN-based methods (i.e., "GNN") like GCN-F [5, 28], TransGNN-F [26], and Graph-Memba-F [2];

Generative methods (i.e., "Generative.") like BERT-F [6], ICL-F [10, 21], Self-Consistency-F [15, 20], and Self-RAG-F [1];

- **Cross-task Adaptation Configurations.** This category of configuration mainly involves improving representative methods from current related research tasks to adapt to the DKGF task, including entity alignment (i.e., "EA."), such as SimpleHHEA-F [9] and ChatEA-F [8]; knowledge graph completion (i.e., "KGC."), such as KG-BERT-F [24, 27], KG-LLaMA-F [25, 27], KoPA-F [27] and PRGC-F [13]; and relation triple extraction (i.e., "RTE."), like NoGen-BART-F [11], and NoGen-T5-F [11]; and domain-specific knowledge graph fusion (i.e., "D."), like `ExeFuse` [29].

B.2.3 Baseline Implementation. To ensure fair comparison, we adapted all baselines to the SKGF task under a **Unified Information Setting**.

Adaptation for Embedding Models (e.g., TransE-F, GCN-F). Since these models are designed for Link Prediction within a single graph, we merged the SKG and GKG into a unified graph $\mathcal{G}_{unified} = \mathcal{G}^s \cup \mathcal{G}^g$ via the alignment anchors. The models were trained to score the plausibility of triples (h^s, r^s, t^g) or (h^g, r^s, t^s).

- **Hyperparameters:** We performed a grid search for the embedding dimension $d \in \{128, 256, 512\}$ and learning rate $\eta \in \{1e-3, 5e-4, 1e-4\}$. The batch size was fixed at 1024.
- **Negative Sampling:** We used strict negative sampling where we corrupted the head or tail entity with a random entity from the *same domain* to enforce domain constraints.

Adaptation for LLM Baselines (e.g., ChatEA, SelfRAG). For method relying on Large Language Models, we constructed prompts that included the local neighborhood subgraph as context.

- **Model Version:** Unless otherwise stated in the ablation, we used gpt-3.5-turbo-1106 for efficient inference and gpt-4-0125-preview for the reported "Best" numbers.
- **Temperature:** Set to 0.2 to reduce randomness while allowing slight creativity for reasoning.

B.3 Main Results

B.3.1 Performance on Standard Benchmarks (SHS). As presented in Table B2, Self-Fusion consistently outperforms all 22 baseline configurations on the widely-used ICEWS datasets (`DKGF(W-I)` and `DKGF(Y-I)`). Specifically, Self-Fusion achieves a substantial improvement of **15.0%** in F1-score compared to the strongest generative baseline, Self-RAG-F. A closer inspection reveals that general-purpose methods (e.g., TransE-F, GCN-F) struggle to bridge the semantic gap due to their reliance on shallow structural features. Similarly, LLM-based cross-task adaptation methods (e.g., ChatEA-F), while semantically powerful, often introduce high-entropy noise due to hallucinations. In contrast, Self-Fusion effectively mitigates these issues by coupling the *fuzzy retriever* with *progressive self-feedback*, ensuring that only scientifically relevant and structurally compatible facts are fused.

B.3.2 Generalization across Scientific Disciplines (LS & NS). To further validate the *scientific rigor* of our framework in "AI for science" scenarios, we extended the evaluation to the diverse disciplines in `SciFusion-Bench`, including biomedical (`SKGF(W-Bio)`), and materials science (`SKGF(W-Mat)`). As shown in Figure B1, Self-Fusion exhibits superior generalization capabilities across all domains.

Insight: "Hard Science" Demands Entropy Reduction. Crucially, we observe that the performance gap between Self-Fusion and LLM-based baselines (e.g., ChatEA-F) widens in "hard science" domains (LS & NS) compared to humanities (SHS).

- On **`SKGF(W-Mat)`** (materials science), Self-Fusion achieves a **13.1%** gain over the runner-up. material discovery requires strict adherence to physical laws; vague associations from GKGs (e.g., matching a material solely by name) often lead to erroneous crystal structure predictions. Our *entropy-driven* mechanism effectively filters this "scientific noise".
- On **`SKGF(W-Bio)`** (biomedical), where precision is paramount to avoid false biological pathways, our model outperforms Self-RAG-F by over **27%**.

This suggests that while LLMs can handle the ambiguity of social sciences (e.g., `SKGF(W-Music)`), they fail to capture the *deterministic mechanisms* required by rigorous scientific domains without the explicit entropy reduction constraints imposed by Self-Fusion.

Table B3: Diagnostic analysis on the auxiliary task "Relevant Scientific Entity Discovery (RSED)".

Models	DKGF(W-I)-S1			DKGF(Y-I)-S1		
	ACC	F1	Avg.	ACC	F1	Avg.
DistMult-F	0.975	0.835	0.905	0.946	0.787	0.866
Self-RAG-F	0.970	0.785	0.878	0.936	0.721	0.829
ChatEA-F	0.973	0.806	0.889	0.938	0.719	0.828
ExeFuse	0.982	0.871	0.927	0.958	0.823	0.890
Self-Fusion	**0.986**	**0.892**	**0.939**	**0.965**	**0.848**	**0.907**

B.4 Relevant Scientific Entity Discovery (RSED)

Accurate identification of specialized, scientific domain-relevant entities serves as the cornerstone for effective knowledge fusion. We evaluate this capability through the diagnostic **RSED** task [29]. As shown in Table B3, Self-Fusion consistently outperforms both the retrieval baseline (Self-RAG-F) and the previous SOTA domain fusion method (`ExeFuse`).

Notably, while `ExeFuse` achieves high accuracy, Self-Fusion yields a significant improvement in F1-score (e.g., **0.892** vs. 0.871 on `DKGF(W-I)`). This indicates that our *Entropy-driven Fuzzy Retriever* effectively balances precision and recall, capturing latent structural isomorphisms that rigid alignment methods miss. Crucially, this superior capability in "finding the right needle in the haystack" directly correlates with the main results in Table B2, where Self-Fusion achieves a 15.0% performance gain. The *Data Self-Feedback* loop plays a pivotal role here, iteratively filtering out high-entropy noise (irrelevant entities) that would otherwise propagate errors to the downstream fusion stage.

B.5 Performance in Low-Resource Settings

Scientific data is often scarce. We evaluated Self-Fusion and key baselines on `SKGF(W-Bio)` by varying the training data ratio from

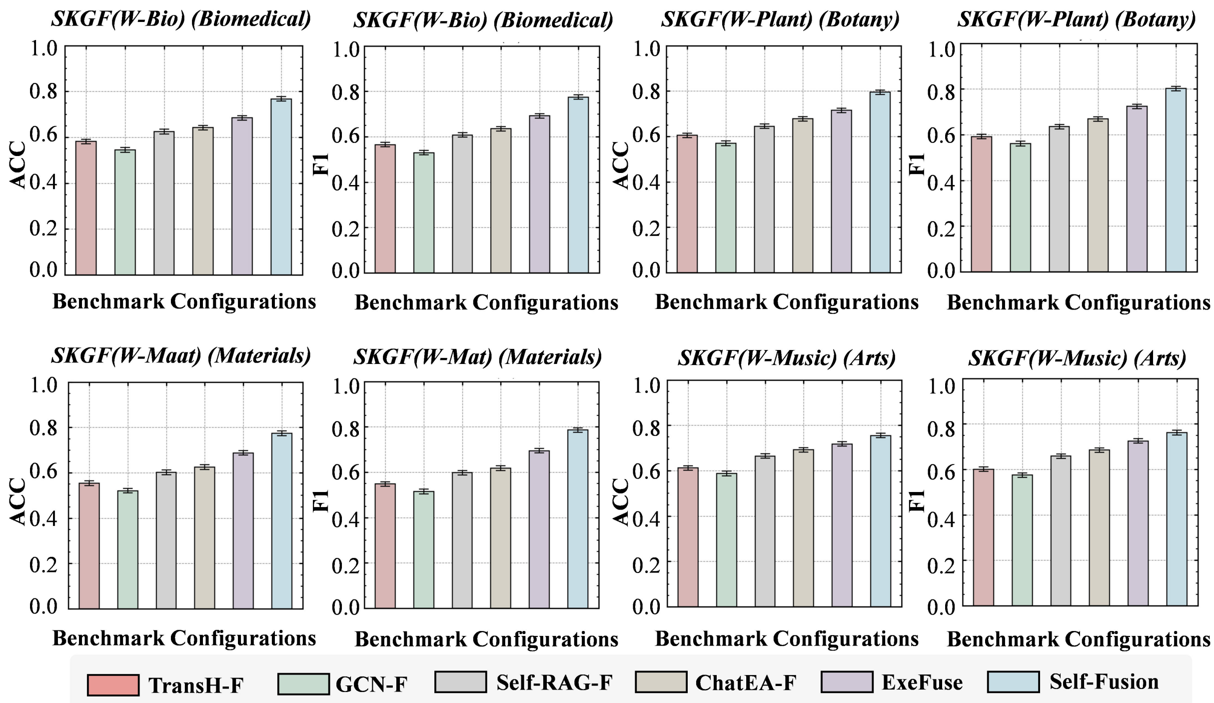

Figure B1: Comprehensive evaluation across diverse scientific disciplines in `SciFusion-Bench`.

20% to 100%. As shown in Table B4, Self-Fusion maintains high performance even with limited data. Notably, at 15.0% training data, Self-Fusion surpasses ChatEA-F (trained on 100% data), demonstrating that our *entropy-driven* mechanism learns generalized scientific logic rather than relying on memorization.

Table B4: Performance (F1-score) w.r.t. Training Data Ratio on `SKGF(W-Bio)`. "Δ" denotes improvement over the best baseline.

Method	20%	40%	60%	80%	100%
TransH-F	0.352	0.415	0.488	0.542	0.565
GCN-F	0.320	0.395	0.462	0.510	0.530
ChatEA-F	0.485	0.542	0.589	0.620	0.635
Self-Fusion	**0.654**	**0.702**	**0.738**	**0.765**	**0.775**
Improvement (Δ)	*+34.8%*	*+29.5%*	*+25.3%*	*+23.4%*	*+22.0%*

B.6 Impact of LLM Backbones

To verify that our performance gain stems from the proposed framework rather than the underlying LLM capability, we tested Self-Fus ion with different backbones on `SKGF(W-Mat)`. Table B5 shows that Self-Fusion powered by the open-source `Llama-3-70B` outperforms ChatEA-F powered by the proprietary `GPT-4`. This confirms the effectiveness of our *Progressive Self-Feedback* design.

B.7 Parameter Sensitivity Analysis

We investigate two critical hyperparameters in Self-Fusion: the entropy threshold τ (which controls fusion strictness) and the structural weight α.

Table B5: Ablation of LLM Backbones on `SKGF(W-Mat)` (F1-score).

Backbone	ChatEA-F	Self-RAG-F	Self-Fusion
GPT-3.5-Turbo	0.582	0.565	**0.742**
Llama-3-8B (Open)	0.545	0.532	**0.695**
Llama-3-70B (Open)	0.605	0.588	**0.768**
GPT-4-Turbo	<u>0.638</u>	<u>0.615</u>	**0.786**

Table B6 (represented as table data for clarity) illustrates the F1-score variations on `DKGF(W-I)`.

- **Entropy Threshold τ:** Performance peaks at $\tau = 0.7$. Lower values ($\tau < 0.5$) introduce noise (high recall, low precision), while aggressive filtering ($\tau > 0.85$) discards valid scientific cues.
- **Structural Weight α:** The optimal α is around 0.4, indicating that structural isomorphism is slightly more important than semantic similarity in defining scientific validity.

Table B6: Sensitivity Analysis on `DKGF(W-I)` (F1-score).

Metric	0.1	0.3	0.5	0.7	0.9
Threshold τ	0.582	0.645	0.712	**0.746**	0.685
Struct. Weight α	0.665	0.735	**0.746**	0.710	0.625

B.8 Case Study

To intuitively understand the "Scientific Rigor," we present a case from `SKGF(W-Bio)`.

- **Query:** Drug *Metformin*.

- **Candidate from Wikidata:** *(Metformin, treats, Cancer).*
- **Baseline (ChatEA-F):** Accepts the fact. (Result: Too generic, low scientific value).
- **Self-Fusion:**
  (1) *Fuzzy Retrieval:* Identifies structural isomorphism between Metformin's pathway and the mTOR pathway.
  (2) *Entropy Check:* The direct link "treats" has high entropy.
  (3) *Self-Feedback:* Reconstructs the chain: *(Metformin, activates, AMPK)* $\rightarrow$ *(AMPK, inhibits, mTOR).*
  (4) *Final Fusion:* Fuses the specific mechanism chain instead of the generic edge.

This demonstrates how `Self-Fusion` crystallizes vague general knowledge into deterministic scientific mechanisms.

B.9 Extended Results on Inductive Generalization

Building upon the transductive and inductive generalization analysis presented in Section 5.5 of the main text, we further evaluate the structural invariance of Self-Fusion across additional scientific domains within the `SciFusion-Bench`. In the context of AI for Science, the ability to generalize to unseen entities is paramount; for instance, applying learned topological mechanisms to newly discovered inorganic crystals in materials science or predicting novel protein functions in bioinformatics without requiring model retraining.

Table B7 details the inductive generalization performance on the `SKGF(W-Bio)` (Life Science) and `SKGF(W-Mat)` (Physical Science) datasets. We partition the test sets into *Seen* and *Unseen* subsets following the identical protocol used for `DKGF(W-I)` in the main paper.

Table B7: Extended inductive generalization analysis on `SKGF(W-Bio)` and `SKGF(W-Mat)`. The *Unseen* subset evaluates the model's capability to discover facts involving entirely novel scientific entities.

Dataset	Model	Seen (Memorization)		Unseen (Discovery)		Overall	
		ACC	F1	ACC	F1	ACC	F1
`SKGF(W-Bio)`	Self-RAG-F	0.640	0.628	0.595	0.568	0.625	0.608
	ChatEA-F	0.665	0.665	0.596	0.575	0.642	0.635
	Self-Fusion	**0.772**	**0.782**	**0.760**	**0.761**	**0.768**	**0.775**
`SKGF(W-Mat)`	Self-RAG-F	0.620	0.622	0.566	0.550	0.602	0.598
	ChatEA-F	0.650	0.662	0.575	0.554	0.625	0.618
	Self-Fusion	**0.780**	**0.795**	**0.762**	**0.768**	**0.774**	**0.786**

The empirical results reveal consistent trends across both biological and physical science domains. Baseline methods, including the LLM-based ChatEA-F, exhibit severe performance degradation on the *Unseen* subsets, confirming their susceptibility to embedding overfitting and their reliance on memorizing specific biochemical or physical entities during training. Conversely, Self-Fusion sustains high F1-scores on the *Unseen* sets (e.g., 0.761 on `SKGF(W-Bio)` and 0.768 on `SKGF(W-Mat)`), underscoring that the entropy-driven progressive self-feedback mechanism successfully distills generalized, domain-agnostic scientific principles. This robust inductive reasoning capacity confirms the framework's readiness for dynamic, open-world scientific environments.

References

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=hSyW5go0v8

[2] Ali Behrouz and Farnoosh Hashemi. 2024. Graph Mamba: Towards Learning on Graphs with State Space Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, Ricardo Baeza-Yates and Francesco Bonchi (Eds.). ACM, 119–130. https://doi.org/10.1145/3637528.3672044

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Adv Neural Inf Process Syst* 26 (2013).

[4] Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data* 10, 1 (2023), 67.

[5] Ziang Chen, Jialin Liu, Xiaohan Chen, Xinshang Wang, and Wotao Yin. 2024. Rethinking the Capacity of Graph Neural Networks for Branching Strategy. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/dff528ce3e1390c88f10bbf5e722a241-Abstract-Conference.html

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/V1/N19-1423

[7] Yongkun Du, Zhineng Chen, Caiyan Jia, Xieping Gao, and Yu-Gang Jiang. 2025. Out of Length Text Recognition with Sub-String Matching. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh, Julie Shah, and Zico Kolter (Eds.). AAAI Press, 2798–2806. https://doi.org/10.1609/AAAI.V39I3.32285

[8] Xuhui Jiang, Yinghan Shen, Zhichao Shi, Chengjin Xu, Wei Li, Zixuan Li, Jian Guo, Huawei Shen, and Yuanzhuo Wang. 2024. Unlocking the Power of Large Language Models for Entity Alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 7566–7583. https://aclanthology.org/2024.acl-long.408

[9] Xuhui Jiang, Chengjin Xu, Yinghan Shen, Yuanzhuo Wang, Fenglong Su, Zhichao Shi, Fei Sun, Zixuan Li, Jian Guo, and Huawei Shen. 2024. Toward Practical Entity Alignment Method Design: Insights from New Highly Heterogeneous Knowledge Graph Datasets. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee (Eds.). ACM, 2325–2336. https://doi.org/10.1145/3589334.3645720

[10] Vignesh Kothapalli, Hamed Firooz, and Maziar Sanjabi. 2025. CoT-ICL Lab: A Synthetic Framework for Studying Chain-of-Thought Learning from In-Context Demonstrations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 14620–14642. https://aclanthology.org/2025.acl-long.712/

[11] You Li, Xupeng Zeng, Yixiao Zeng, and Yuming Lin. 2024. Enhanced Packed Marker with Entity Information for Aspect Sentiment Triplet Extraction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 619–629. https://doi.org/10.1145/3626772.3657734

[12] Pasquale Lisena, Manel Achichi, Pierre Choffé, Cécile Cecconi, Konstantin Todorov, Bernard Jacquemin, and Raphaël Troncy. 2018. Improving (re-) usability of musical datasets: An overview of the doremus project. *Bibliothek Forschung und Praxis* 42, 2 (2018), 194–205.

[13] Xin Lv, Yankai Lin, Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. Do Pre-trained Models Benefit Knowledge Graph Completion? A Reliable Evaluation and a Reasonable Approach. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 3570–3581. https://doi.org/10.18653/V1/2022.FINDINGS-ACL.282

[14] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature* 624, 7990 (2023), 80–85.

[15] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.* 36, 7 (2024), 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100

[16] Derek Paulsen, Yash Govind, and AnHai Doan. 2023. Sparkly: A Simple yet Surprisingly Strong TF/IDF Blocker for Entity Matching. *Proc. VLDB Endow.* 16, 6 (2023), 1507–1519. https://doi.org/10.14778/3583140.3583163

[17] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR (Poster)*. OpenReview.net.

[18] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *ICML (JMLR Workshop and Conference Proceedings, Vol. 48)*. JMLR.org, 2071–2080.

[19] Aravind Venkatesan, Gildas Tagny Ngompe, Nordine El Hassouni, Imene Chentli, Valentin Guignon, Clement Jonquet, Manuel Ruiz, and Pierre Larmande. 2018. Agronomic Linked Data (AgroLD): A knowledge-based system to enable integrative biology in agronomy. *PLoS One* 13, 11 (2018), e0198270.

[20] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=1PL1NIMMrw

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[22] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR (Poster)*.

[23] Linyan Yang, Shiqiao Zhou, Jingwei Cheng, Fu Zhang, Jizheng Wan, Shuo Wang, and Mark Lee. 2025. DAEA: Enhancing Entity Alignment in Real-World Knowledge Graphs Through Multi-Source Domain Adaptation. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 5890–5901. https://aclanthology.org/2025.coling-main.393/

[24] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for Knowledge Graph Completion. *CoRR* abs/1909.03193 (2019). arXiv:1909.03193 http://arxiv.org/abs/1909.03193

[25] Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. 2025. Exploring Large Language Models for Knowledge Graph Completion. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*. IEEE, 1–5. https://doi.org/10.1109/ICASSP49660.2025.10889242

[26] Peiyan Zhang, Yuchen Yan, Xi Zhang, Chaozhuo Li, Senzhang Wang, Feiran Huang, and Sunghun Kim. 2024. TransGNN: Harnessing the Collaborative Power of Transformers and Graph Neural Networks for Recommender Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 1285–1295. https://doi.org/10.1145/3626772.3657721

[27] Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Wen Zhang, and Huajun Chen. 2024. Making Large Language Models Perform Better in Knowledge Graph Completion. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (Eds.). ACM, 233–242. https://doi.org/10.1145/3664647.3681327

[28] Zhanqiu Zhang, Jie Wang, Jieping Ye, and Feng Wu. 2022. Rethinking Graph Convolutional Networks in Knowledge Graph Completion. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini (Eds.). ACM, 798–807. https://doi.org/10.1145/3485447.3511923

[29] Runhao Zhao, Weixin Zeng, Wentao Zhang, Chong Chen, Zhengpin Li, Xiang Zhao, and Lei Chen. 2026. Panning for Gold: Expanding Domain-Specific Knowledge Graphs with General Knowledge. arXiv:2601.10485 [cs.AI] https://arxiv.org/abs/2601.10485