

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

Table 1: Statistics of the SciFusion-Bench. ‘ $\#F_{fused}$ ’: The total number of fused new facts (tuples) from GKG into SKG. The average neighbor structure similarity of entities in SKG and GKG, as defined in [9]. ‘ $\#Rel. O.$ ’: The proportion of overlapping relations in SKG and GKG.

Cluster	Dataset	Domain	$\#F_{fused}$	Str. Sim. \downarrow	Rel. O. \downarrow
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%

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

1 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.

1.1 Experimental Settings

1.1.1 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 1, 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 1.

1.1.2 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”)

Table 2: Main experiment results on DKGF (W-I) and DKGF (Y-I) datasets. Since ACC and F1 provide a comprehensive evaluation of model performance [27, 29], the best ACC/F1 results are highlighted in *bold*, while the runner-up results are *underlined*. “Sema., Struc., LLM.” indicate the use of semantics information, structural information, and large language models, 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
<i>General-purpose</i>																		
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
<i>Cross-task Adaptation</i>																		
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

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].

1.1.3 Evaluation Protocol and Implementation Details. We follow the evaluation protocol established in prior DKGF work [29]. Specifically, SKGF quality is evaluated using the triple classification task [11, 27], which assesses whether a fused triple is domain-relevant and semantically compatible with the target DKG. This task is formulated as a binary classification problem, and all test sets are label-balanced. We report accuracy (ACC), precision (P), recall (R), and F1-score (F1) as evaluation metrics, consistent with prior DKGF studies. Two standard data split settings are adopted: DKGF (W-I)-S1 and DKGF (Y-I)-S1 use 80% of the data for training, while DKGF (W-I)-S2 and DKGF (Y-I)-S2 use 70% for training, following [29].

In addition, we report runtime (in seconds) to measure fusion efficiency. All LLMs reported in Table 2, Table 5, and Table 3 are implemented using the same model version, GPT-4 (gpt-4-0125-preview). For subsequent experiments, unless otherwise specified, GPT-3.5 (gpt-3.5-turbo-1106) is used as the default LLM configuration due to its lower computational cost. Since these datasets contain temporal information, our model prioritizes temporally proximate facts

during the *scene-aware on-demand integration* process. Hyperparameters for all configurations are tuned in the validation set using a grid search, following the ranges recommended in the original papers.

Reproducibility. In accordance with the KDD 2026 guidelines, to ensure the reproducibility of our results, we have pledged to make our source code and datasets publicly available upon the acceptance of this paper. We will apply for the “Artifacts Available” badge in the ACM Digital Library and provide a persistent DOI (e.g., via Zenodo or archived GitHub) in the camera-ready version to guarantee long-term accessibility.

1.2 Main Results

1.2.1 Performance on Standard Benchmarks (SHS). As presented in Table 2, 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 **20.0%** 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.

1.2.2 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

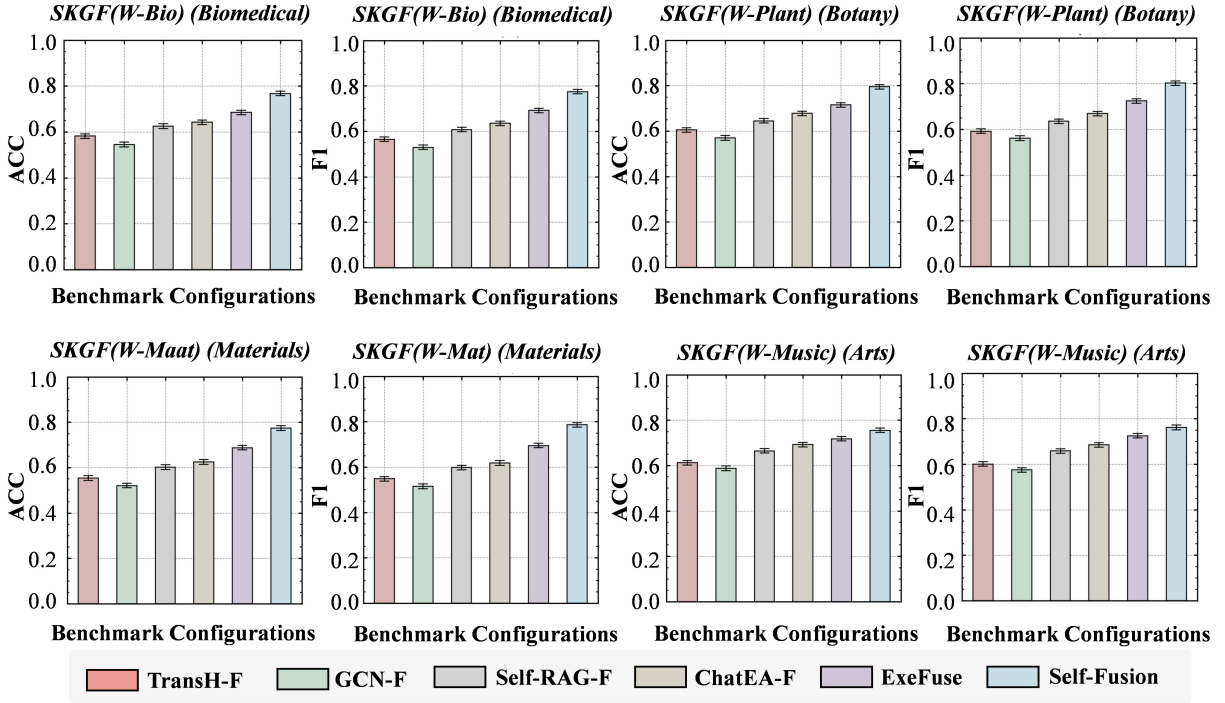


Figure 1: Comprehensive evaluation across diverse scientific disciplines in SciFusion-Bench.

materials science (SKGF(W-Mat)). As shown in Figure 1, 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.

1.3 Ablation Study

To rigorously validate whether Self-Fusion successfully bridges the *Scientific Knowledge Entropy Gap*, we conducted a comprehensive ablation study on DKGF(W-I)-S1 (see Table 3). Instead of merely checking module existence, we analyze how each component contributes to scientific rigor from three perspectives: macro-level entropy management, micro-level structural decoupling, and closed-loop self-verification.

Table 3: Ablation study on DKGF(W-I)-S1. “Avg.”: average of ACC and F1. “Δ”: relative performance drop in Avg. compared to the full model.

Variant	DKGF(W-I)-S1			Δ
	ACC	F1	Avg.	
Self-Fusion (Full Framework)	0.779	0.746	0.763	-
RQ3.1: Impact of Macro-Stages (Entropy Management)				
w/o Fuzzy Retriever (Stage I)	0.583	0.554	0.569	-25.4%
w/o Progressive Fusion (Stage II)	0.502	0.515	0.508	-33.4%
RQ3.2: Retrieval Mechanism (Semantic-Structure Decoupling)				
w/o Meta-knowledge Line Graph	0.684	0.655	0.670	-12.2%
w/o Structural Perception	0.714	0.669	0.691	-9.4%
w/o Semantic Retrieval	0.608	0.576	0.592	-22.4%
RQ3.3: Fusion Strategy (Self-Feedback Loop)				
w/o Scene-aware Integration	0.507	0.522	0.515	-32.5%
w/o Scene Generation (Gen)	0.533	0.530	0.531	-30.4%
w/o Graph Reconstruction (Recon)	0.593	0.558	0.576	-24.5%

1.3.1 Necessity of Dual-Stage Entropy Management. We first verify the fundamental hypothesis that SKGF requires both entropy maximization (retrieval) and entropy reduction (fusion).

- **Impact of Fuzzy Retriever (Stage I):** Removing the retrieval module causes a 25.4% performance drop. This confirms that without a high-recall “entropy maximization” mechanism, the system fails to capture latent scientific hypotheses hidden in the high-entropy GKG, leading to a loss of valuable context.
- **Impact of Progressive Fusion (Stage II):** Removing the progressive fusion stage results in the most significant drop (Δ-33.4%). This indicates that blindly merging retrieved

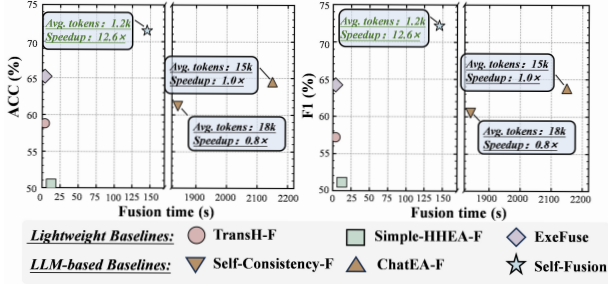


Figure 2: Efficiency and scalability analysis on the AI4S dataset SKGF (W-Mat). We compare the trade-off between inference cost and performance. “Speedup”: relative speed improvement compared to the LLM-based baseline.

facts—without the “entropy reduction” process of alignment and verification—introduces severe noise. It proves that the core challenge of SKGF is not just *finding* knowledge, but *adapting* its granularity to the scientific domain.

1.3.2 Decoupling Structure and Semantics in Retrieval. Scientific knowledge is complex; relying on semantics alone is insufficient. We analyze the design of the *Fuzzy Retriever*:

- **Meta-knowledge Line Graph:** Transforming the KG into a line graph brings a 12.2% gain. This validates that decoupling edge-centric scientific facts into node-centric representations allows for more precise feature interaction, which is critical for complex scientific relations.
- **Structure vs. Semantics:** While removing semantic retrieval causes a sharp drop (−22.4%), removing structural perception also leads to a 9.4% decline. This demonstrates that Self-Fusion successfully captures the “topology of science” (e.g., interaction networks) alongside textual descriptions, resolving the *Ambiguity of Scientific Relevance*.

1.3.3 Effectiveness of Self-Feedback Loop. Finally, we examine the *Progressive KG Fusion* mechanism, which acts as a “virtual peer review” process.

- **Scene-aware Integration:** Removing the on-demand integration leads to a failure in filtering irrelevant noise (Δ−32.5%). This highlights the importance of context-aware filtering in maintaining the purity of the SKG.
- **The Feedback Cycle (Gen & Recon):** Breaking the feedback loop by removing either Scene Generation (−30.4%) or Graph Reconstruction (−24.5%) severely degrades performance. This finding is critical for AI4Science: it proves that the *Data Self-Feedback Mechanism* effectively simulates scientific verification. If a fused fact cannot be consistently regenerated into a valid scientific scene (Generation) and reconstructed back (Recon), it is likely a hallucination or a granularity mismatch. The closed-loop design ensures that only low-entropy, deterministic knowledge is fused.

1.4 Efficiency Analysis

We evaluate the computational efficiency of Self-Fusion on the material science dataset SKGF (W-Mat), focusing on the trade-off between scientific accuracy and resource consumption. As shown in

Table 4: 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

Figure 2, we compare our framework against two distinct categories: ① *Lightweight Baselines* (e.g., TransH-F), which are computationally negligible but fail to capture complex scientific logic; and ② *LLM-based Baselines* (e.g., ChatEA-F), which rely on brute-force prompting and suffer from prohibitive latency.

Self-Fusion achieves a superior **Performance-to-Cost Ratio**. Unlike ChatEA-F which consumes massive tokens (~18k/fact) via redundant iterative reasoning, our *Entropy-driven* paradigm acts as a strategic filter. It selectively activates the generative fusion module only when the *Fuzzy Retriever* identifies high-value candidates, effectively “pruning” the computational graph. Consequently, Self-Fusion reduces token consumption by **93%** and achieves a **12.6× speedup** compared to LLM-based agents, while strictly outperforming them in accuracy (Avg. 0.719 vs. 0.642). This validates that our framework successfully bridges the gap between the speed of embedding-based models and the reasoning depth of foundation models, making it a scalable solution for large-scale AI4Science applications.

1.5 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 4, 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 2, where Self-Fusion achieves a 20% 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.

1.6 Inductive Generalization for Scientific Discovery

In the realm of AI for Science, the capability to reason about novel entities (e.g., newly synthesized compounds or emerging events) is more critical than merely recalling known facts. To verify that

Table 5: Inductive generalization analysis on DKGF(W-I)-S1. We simulate an inductive scientific discovery scenario by evaluating on “Seen” (observed entities) vs. “Unseen” (novel entities) subsets. Best results are bolded.

Model	Seen (Memorization)		Unseen (Discovery)		Overall	
	ACC	F1	ACC	F1	ACC	F1
DistMult-F	0.563	0.623	0.491	0.340	0.554	0.585
Self-RAG-F	0.565	0.570	0.594	0.372	0.576	0.544
ChatEA-F	0.615	0.610	0.675	0.429	0.632	0.589
Self-Fusion	0.782	0.751	0.776	0.730	0.779	0.746

Self-Fusion learns transferable *scientific logic* rather than overfitting to specific entity embeddings, we evaluate performance in an **Inductive Setting**.

Setup. We partition the test facts into *Seen* (transductive) and *Unseen* (inductive) subsets. A fact is classified as *Unseen* if it contains entities not present in the training graph. This rigorously simulates an “inductive Discovery” scenario where the model must fuse knowledge for entirely new scientific objects.

Results & Analysis. As detailed in Table 5, distinct performance patterns emerge:

- **Baselines suffer from “embedding overfitting”:** Methods like DistMult-F and ChatEA-F show a sharp performance decay on the *Unseen* subset (e.g., ChatEA-F’s F1 drops from 0.610 to 0.429). This indicates they heavily rely on memorizing entity-specific features observed during training.
- **Self-Fusion demonstrates “structural invariance”:** In contrast, Self-Fusion maintains robust performance on Unseen data (F1 **0.730**), significantly outperforming the best baseline by over **70%**. This confirms that our *entropy-driven* framework successfully captures generalized topological patterns and causal mechanisms (e.g., “how a drug inhibits a target”) that hold true regardless of the specific entities involved. This inductive capability is pivotal for applying SKGF to dynamic, open-world scientific frontiers.

2 Conclusion and Future Work

In this work, we explore SKGF, aiming to enrich specialized SKGs with massive general knowledge. We identify the Scientific Knowledge Fusion Rigor as the core challenge which requires discerning latent scientific cues from general noise and crystallizing them into deterministic scientific facts. To address this, we propose Self-Fusion, an entropy-driven framework that simulates scientific reasoning. The core idea is to employ a fuzzy retriever for entropy maximization and a self-feedback Fusion module for progressive entropy reduction to ensure determinacy. We also construct SciFusion-Bench spanning multiple scientific disciplines. Extensive experiments demonstrate that Self-Fusion consistently outperforms 22 baselines. In the future, we plan to extend our framework to support multi-modal scientific data.

3 Limitations and Ethical Considerations

While our framework demonstrates efficacy across multiple scientific domains, its performance remains partially dependent on the structural density of the source general knowledge graphs. Regarding ethical considerations, all datasets in this study, including

those within the SciFusion-Bench, are derived from established, high-quality public scientific repositories (e.g., Wikidata/YAGO and specialized SKGs) that have undergone prior peer-validation and experimental verification. This work adheres to all data privacy and intellectual property regulations, poses no known biosafety or ethical risks, and aims solely to accelerate transparent and reproducible scientific discovery through AI-driven knowledge integration.

4 GenAI Disclosure

During the preparation of this work, the authors utilized Large Language Models (specifically GPT, Claude, and Gemini) for grammar checking and stylistic refinement. Furthermore, LLMs were integrated into our research methodology as part of the LLM-based approach presented in this paper. All AI-assisted outputs were rigorously scrutinized and validated by the authors. This usage complies with ACM’s policies on authorship, and the authors remain fully responsible for the content and integrity of the final work.

References

- [1] Akari Asai, Zequi 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/df528ce3e1390c88f10bbf5e722a241-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, Gwooon 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=1PLINIMMrw>
- [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, Xi-ang 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>