

COMPUTER SCIENCE

Bridging Data and Discovery: A Survey on Knowledge Graphs in AI for Science

Keyan Ding^{1,2,†}, Zhihui Zhu^{2,†}, Yuqi Tang³, Kehua Feng¹, Xiang Zhuang^{1,4}, Hongwei Wang³, Yi Yang¹, Huifang Du⁵, Zhangkai Ni⁶, Shiqi Wang⁷, Xiaohui Fan⁸, Huabin Xing⁹, Lei Bai^{4,*}, Qi Liu^{10,*}, Haofen Wang^{5,*}, Qiang Zhang^{2,3,*} and Huajun Chen^{1,2,*}

¹ College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China;

² ZJU-Hangzhou Global Scientific and Technological

Innovation Center, Zhejiang University, Hangzhou 311200, China;

³ ZJU-UIUC Institute, Zhejiang University, Haining 314400, China;

⁴ Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China;

⁵ College of Design and Innovation, Tongji University, Shanghai 200092, China;

⁶ College of Computer Science and Technology, Tongji University, Shanghai 201804, China;

⁷ Department of Computer Science, City University of Hong Kong, Hong Kong 999077, China;

⁸ College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;

⁹ College of Chemical and Biological Engineering, Zhejiang University, Hangzhou 310027, China;

¹⁰ School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

*Corresponding authors.

Email:

baisanshi@gmail.com,
qiliu@tongji.edu.cn,
haofen.wang@tongji.edu.cn,
qiang.zhang.cs@zju.edu.cn,
huajunsir@zju.edu.cn.

[†] Equally contributed to this work.

Received: XX XX Year;

Revised: XX XX Year;

Accepted: XX XX Year

ABSTRACT

Knowledge graphs have emerged as a powerful paradigm for structuring, organizing, and reasoning over complex scientific knowledge, and are increasingly recognized as catalysts for accelerating AI for science. This study provides a comprehensive survey of Scientific Knowledge Graphs (SciKGs), covering their construction methodologies and diverse applications across biology, chemistry, and materials science. We examine how SciKGs support tasks such as drug development, omics analysis, reaction prediction, and materials design, and highlight how the synergistic integration of SciKGs and large language models (LLMs) forms a knowledge- and language-driven framework for scientific discovery, in which SciKGs serve as the foundational knowledge infrastructure and LLMs act as dynamic semantic engines. We further identify key challenges and outline emerging opportunities toward building auditable, interoperable, and self-evolving SciKGs. Looking forward, we envision a new generation of SciKG-centered ecosystems where self-updating graphs, co-evolving with LLMs and embodied within AI scientists, become core infrastructures that autonomously drive, verify, and accelerate scientific discovery.

Keywords: Scientific Knowledge Graphs, AI for Science, Knowledge-driven Framework, Autonomous Scientific Discovery

1 INTRODUCTION

Scientific discovery is undergoing a paradigm shift from intuition-driven exploration to data-intensive, AI-powered inference. The deluge of high-throughput experiments, large-scale simulations, and multimodal sensing technologies has generated unprecedented volumes of heterogeneous, complex data across biology, chemistry, and materials science [1,2]. Yet, this data explosion has not been matched by a corresponding leap in our ability to synthesize, contextualize, and reason over it. Fragmentation across formats, terminologies, and domains leaves vast reservoirs of scientific knowledge underutilized – a “knowledge gap” that threatens to widen as data generation outpaces human interpretability [3]. Addressing this challenge requires computational frameworks capable of unifying, representing, and reasoning over large-scale knowledge.

Knowledge Graphs (KGs) have emerged as a powerful paradigm for organizing structured information by representing entities and their relations in a machine-interpretable form [4,5]. Generally, a KG can be defined as a directed, labeled graph where nodes represent entities and edges denote semantic relations among them. In scientific domains, KGs provide a unifying representation of diverse entities, such as genes, proteins, diseases, chemical compounds, and materials, capturing their intricate relationships across experimental and computational contexts [6,7]. Over the past decades, scientific knowledge graphs (SciKGs) have been applied to diverse problems such as drug repurposing, multi-omics analysis, chemical reaction modeling, and materials design [8–10], demonstrating their potential as engines of discovery.

However, constructing SciKGs remains technically demanding. Entity and relation extrac-

tion, ontology alignment, and knowledge integration must contend with unstructured scientific texts, inconsistent terminologies, and rapidly evolving knowledge. Traditional rule-based or ontology-driven approaches provide valuable structure but often lack scalability and adaptability in the face of scientific data complexity [3,11]. The integration of artificial intelligence (AI) techniques, particularly large language models (LLMs) [12], has begun to transform this landscape. LLMs can automate knowledge extraction from unstructured literature, enrich semantic representations, and predict missing links within graphs [13,14]. Conversely, SciKGs provide structured grounding for LLMs, improving factual reliability, contextual reasoning, and reducing hallucinations in generative scientific tasks [15,16]. This bidirectional synergy between SciKGs and LLMs is opening new opportunities for AI-driven scientific reasoning, hypothesis generation, and decision support.

Despite growing interest, most surveys to date have concentrated on general-purpose KGs [17–19], providing valuable overviews of graph construction techniques and applications but offering limited insight into the unique demands of scientific domains. Existing reviews of SciKGs [11,20] remain fragmented, often narrowing their scope to a single scientific field such as biomedicine. Moreover, they rarely explore the integration of SciKGs with LLMs, neglecting one of the most transformative developments in the field. What is still lacking is a unified, cross-disciplinary perspective that captures the full landscape of SciKGs (from construction and integration to application and evolution) and highlights their symbiosis with LLMs as a catalyst for accelerating discovery.

In this study, we fill this gap and provide a comprehensive survey of KGs in the fundamental scientific domains, particularly focusing on biology, chemistry, and materials science (Fig. 1). Specifically, we make four distinctive contributions. First, we systematically examine how SciKGs are constructed and applied across diverse scientific domains, highlighting their roles in advancing drug development, omics analysis, chemical synthesis, and materials discovery. Second, we place particular emphasis on the integration of SciKGs with LLMs for scientific discovery, showing how this emerging synergy opens new opportunities for knowledge extraction, reasoning, and generation. Third, we highlight unresolved challenges and propose concrete research directions to guide the development of

next-generation knowledge discovery systems in the LLM era. Fourth, we establish and actively maintain a curated, open-access repository for SciKGs at GitHub (<https://github.com/hicai-zju/scikgs>), which provides up-to-date resources including literature, datasets, and software. Together, these contributions establish this work as a comprehensive, living reference and a strategic roadmap for advancing scientific knowledge graphs in the era of AI-driven discovery.

Ultimately, this review aims to address a core question: How can knowledge graphs serve as the structured knowledge infrastructure for AI for Science, and how can SciKGs and LLMs synergize to enable future autonomous scientific discovery? By bridging the precision of symbolic knowledge with the generative power of large models, we argue that SciKGs are not merely data repositories but the essential deterministic substrate required to ground, validate, and accelerate the next generation of scientific intelligence.

The remainder of this review provides a roadmap for understanding, constructing, and leveraging SciKGs to accelerate AI for science (Fig. 2). Section 2 lays the conceptual foundation by defining SciKGs, outlining their key roles in organizing and reasoning over scientific knowledge, and tracing their historical evolution. Section 3 guides readers through the construction process, detailing strategies for integrating heterogeneous data, extracting entities and relations, aligning ontologies, and curating high-quality graphs. In Section 4, we chart the diverse domains where SciKGs are applied, with illustrative examples in biology, chemistry, and materials science, showing how these graphs enable interpretation, prediction, and generation. Section 5 discusses how SciKGs can be combined with large language models to accelerate scientific discovery, emphasizing their complementary roles in knowledge grounding and reasoning. Finally, Section 6 outlines the key challenges and opportunities that define the next stage in SciKG development, offering a forward-looking perspective on building robust knowledge infrastructures for LLM-driven autonomous AI system.

2 CONCEPTUAL FOUNDATIONS OF SCIENTIFIC KNOWLEDGE GRAPHS

Scientific Knowledge Graphs (SciKGs) provide a structured, semantically rich, and computable representation of scientific entities, their relationships, and contextual information across diverse disciplines. Unlike general-purpose

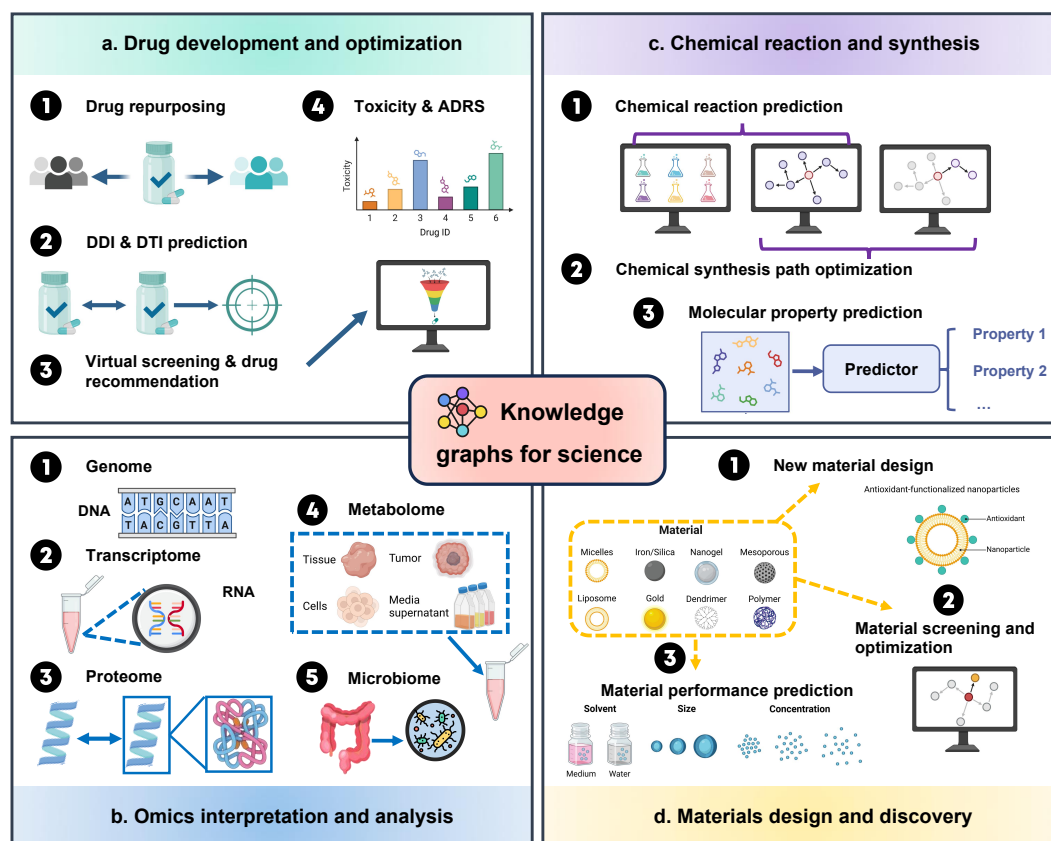


Figure 1. An overview of the research scope in this survey, covering four fundamental scientific tasks in biology, chemistry, and materials science: (a) drug development and optimization, (b) omics interpretation and analysis, (c) chemical reaction and synthesis, and (d) materials design and discovery.

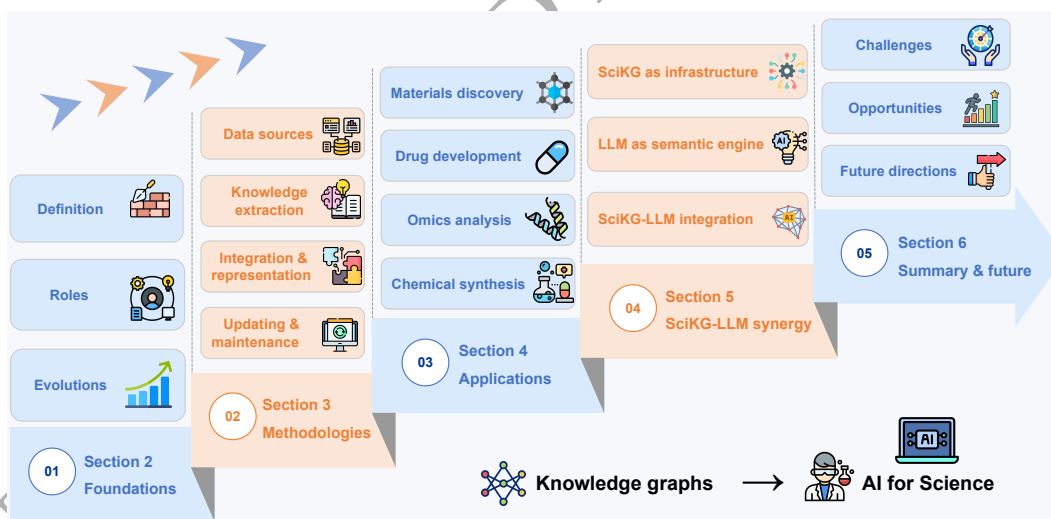


Figure 2. Structure of the survey. Our review is structured around the lifecycle of SciKGs: from their conceptual foundation and construction methodologies, to their applications and synergistic integration with LLMs for discovery, culminating in challenges, opportunities and future directions that envision SciKGs as engines for autonomous scientific discovery.

knowledge graphs that prioritize broad coverage and common-sense reasoning, SciKGs are purpose-built to encode domain-specific semantics, causal relationships, and contextual constraints inherent to scientific inquiry. In this sec-

tion, we introduce their definitions, roles, and evolutions in scientific discovery, laying the conceptual foundation for subsequent discussions on construction methodologies and applications.

2.1 Definitions

Formally, a SciKG can be defined as a directed, labeled graph $G = (V, E)$, where each node $v \in V$ represents a scientific entity (e.g., a gene, protein, compound, reaction, or material), and each edge $e \in E$ denotes a semantic relation between entities (e.g., activation, inhibition, binding, catalysis, or synthesis). In addition to structural connectivity, nodes and edges are often enriched with metadata such as provenance, experimental conditions, quantitative attributes, and links to external databases or literature references. This multi-layered representation transforms raw scientific data into an interconnected knowledge fabric that supports both human interpretability and automated reasoning. Moreover, SciKGs increasingly incorporate temporal, contextual, and multimodal dimensions. Temporal edges encode the evolution of knowledge over time, capturing how hypotheses, measurements, and discoveries emerge or are refuted. Contextual layers specify experimental settings, materials compositions, or biological environments in which relations hold true. Multimodal extensions integrate textual, numerical, and visual modalities, e.g., linking microscopy images or spectroscopy spectra to molecular entities, creating a richer and more expressive knowledge representation suitable for data-intensive science. Crucially, it is important to distinguish SciKGs from scientific data graphs. While data graphs (e.g., crystal structure graphs representing atomic connectivity, or raw protein interaction networks based on correlations) focus on the geometric or topological structure of specific data samples, SciKGs are fundamentally defined by their semantic backbone rooted in domain ontologies. In a multimodal SciKG, numerical, visual, and temporal data do not replace the semantic graph but serve as multimodal attributes or grounding evidence linked to specific entities, thereby enriching the symbolic knowledge with dense, computable representations.

2.2 Roles

SciKGs serve as a foundational infrastructure that bridges data, knowledge, and intelligence in scientific discovery. Their roles can be categorized along four axes:

1. *Knowledge Organization*: SciKGs unify heterogeneous data sources, spanning biological sequences, chemical structures, materials properties, and experimental records, under a consistent semantic schema. This unification mitigates data fragmentation, improves inter-

operability, and provides researchers with a single point of access for querying and integrating diverse knowledge.

2. *Knowledge Embedding*: SciKGs provide a scaffold for learning contextualized embeddings of scientific entities. Through knowledge graph embedding (KGE) approaches [20,21], entities and relations are projected into continuous vector spaces where geometric proximity encodes semantic relatedness. These representations enrich downstream tasks such as drug–target prediction, materials property estimation, or pathway inference by injecting structured scientific priors into model learning.
3. *Knowledge Inference*: By encoding relational dependencies, SciKGs enable various forms of reasoning such as link prediction, causal inference, and hypothesis generation. Graph algorithms and embedding-based approaches [22] allow the prediction of novel interactions (e.g., drug–target binding, gene–disease associations, or reaction pathways) that may not be explicitly observed in experimental data.
4. *Knowledge Interpretability*: Unlike black-box predictive models, SciKGs preserve explicit semantic relationships and traceable provenance information [23]. This transparency allows scientists to validate model predictions, interpret causal chains, and connect inferred results back to experimental evidence or literature sources, fostering trust and accountability in AI-driven discovery.

2.3 Evolutions

The development of SciKGs has undergone several transformative phases (Fig. 3), reflecting the co-evolution between knowledge representation technologies and scientific practices. Here we identify four key phases:

- *Cataloging Era (Pre-2000s)*: Early efforts focused on structured databases and controlled vocabularies (e.g., GenBank [24], PDB [25]). Knowledge was stored in relational tables with limited semantic expressivity, primarily supporting lookup and retrieval rather than reasoning.
- *Semantic Web Era (2000s–2010s)*: The introduction of the Resource Description Framework (RDF), Web Ontology Language (OWL), and SPARQL enabled the formal representation of scientific entities and relationships, giving rise to semantically interoperable knowledge systems. Initiatives such as Bio2RDF [26] and the Open Biological

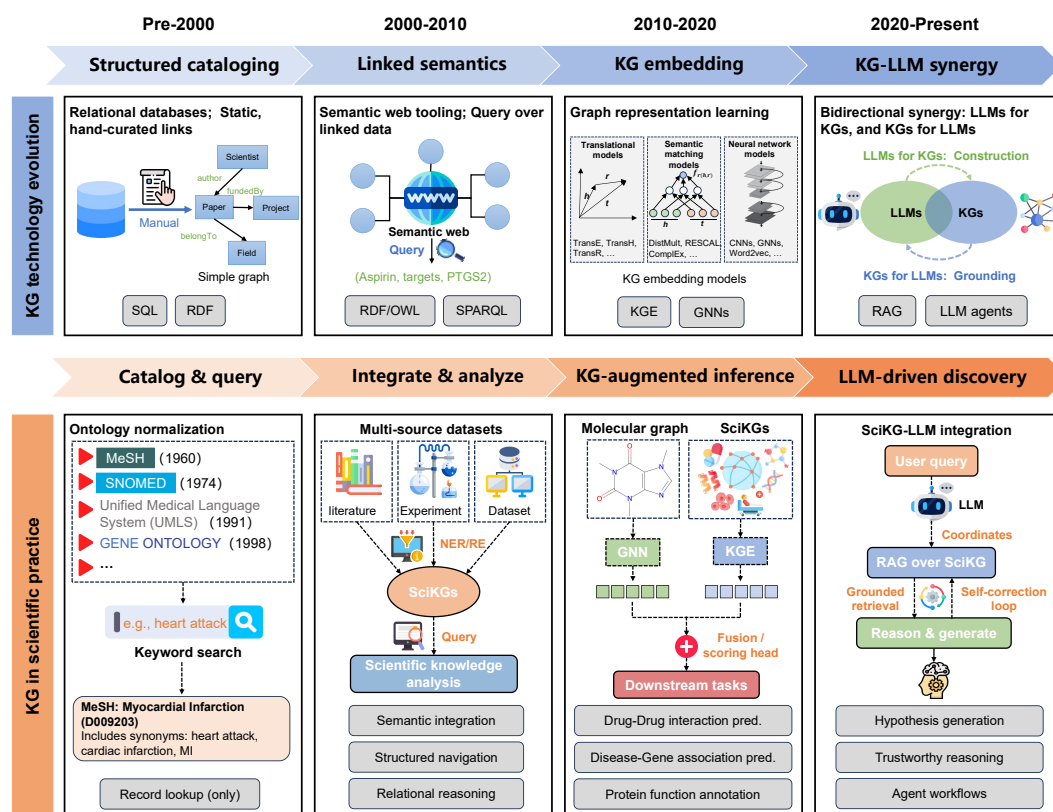


Figure 3. The co-evolution of knowledge graph technologies and their scientific practices. The technological evolution of KGs (top) has continually enabled new paradigms in SciKG applications (bottom). This progression has moved from static cataloging and manual integration to machine learning-driven inference, culminating in the current era of bidirectional synergy between LLMs and KGs. This synergy, leveraging tools such as RAG and AI agents, transforms SciKGs from static repositories into dynamic engines for generative scientific discovery. [Abbr., SQL: Structured Query Language; RDF: Resource Description Framework; OWL: Web Ontology Language; SPARQL: SPARQL Protocol and RDF Query Language; GNN: graph neural network; KGE: knowledge graph embedding; RAG: retrieval-augmented generation.]

and Biomedical Ontology (OBO) Foundry [27] exemplified this era, promoting cross-database reasoning and federated query capabilities.

- **Machine Learning Era (2010s–2020s):** With the emergence of graph embeddings and graph neural networks, SciKGs evolved into predictive engines capable of inferring new links and patterns from existing knowledge. Representation learning (e.g., TransE [28], GraphSAGE [29]) bridged the gap between symbolic knowledge and numerical computation, unlocking applications in drug repurposing, reaction prediction, and materials property estimation.
- **Large Language Model Era (2020s–present):** The integration of large language models has catalyzed a new paradigm. LLMs automate KG construction from literature (e.g., AutoKG [30]), generate hypotheses grounded in SciKGs (e.g., SciAgents [31]), and serve as natural-language interfaces for complex queries (e.g., DDI-GPT [32]). Conversely,

SciKGs mitigate LLM hallucinations via retrieval-augmented generation (RAG) and provide structured constraints for scientific plausibility. This bidirectional synergy transforms SciKGs from static knowledge storage to intelligent infrastructures.

Overall, the conceptual evolution of SciKGs mirrors the broader transformation of scientific inquiry: from static cataloging to semantic reasoning, and now toward autonomous, knowledge-augmented discovery. By bridging structured knowledge and generative intelligence, SciKGs lay the foundation for a new era of AI-driven scientific discovery.

3 METHODOLOGIES FOR CONSTRUCTING SCIENTIFIC KNOWLEDGE GRAPHS

The construction of SciKGs is a multi-stage process that involves integrating heterogeneous data sources, extracting entities and relations, aligning knowledge with existing ontologies, and ensuring the dynamic maintainability of the re-

sulting graphs (Fig. 4). Unlike general-purpose knowledge graphs, SciKGs face the additional complexity of representing domain-specific entities such as genes, proteins, and molecules, which often require fine-grained semantic modeling and contextual reasoning. In this section, we briefly review the major aspects of SciKG construction, including data sources, extraction techniques, integration strategies, and maintenance approaches. We also highlight emerging trends in multimodal SciKGs, which are increasingly critical for capturing the full complexity of scientific data. Tables S1 and S2 summarize the commonly used resources (including databases, software, and tools) for SciKG construction and management.

3.1 Data Sources

SciKGs are built from a wide range of scientific data sources, which can be broadly categorized into structured databases, unstructured text, and multimodal repositories.

Structured data sources form the semantic backbone of most domain-specific graphs. Well-curated repositories such as PubChem [33], UniProt [34], the Protein Data Bank (PDB) [25], and the Materials Project [35] provide standardized, machine-readable annotations of molecular structures, protein interactions, crystal lattices, and thermodynamic properties. These resources are typically developed under community-endorsed standards and employ persistent identifiers (e.g., DOIs, InChI, or UniProt accessions), ensuring interoperability and reproducibility. As such, they serve as stable and verifiable foundations for constructing large-scale SciKGs.

Unstructured textual sources (including scientific papers, patents, laboratory notebooks, and experimental reports) represent the most abundant yet least structured form of scientific knowledge. Massive text corpora such as PubMed, arXiv, and the USPTO database collectively encode millions of entities, relations, and claims expressed in natural language. Extracting structured knowledge from these heterogeneous materials requires advanced natural language processing (NLP) pipelines, encompassing named entity recognition, dependency parsing, event extraction, and relation detection.

Multimodal data sources are increasingly indispensable for capturing the quantitative and contextual complexity of modern science. These encompass a wide spectrum of experimental and computational modalities: Omics profiles (e.g., RNA-seq, proteomics, metabolomics) that quan-

tify molecular abundance; Imaging data such as electron microscopy, fluorescence microscopy, or X-ray diffraction patterns; Spectroscopic signals (e.g., NMR) that encode molecular fingerprints; Computational simulations, including molecular dynamics trajectories or density functional theory outputs; Time-series experimental measurements, such as thermal degradation curves or electrochemical cycling data. The integration of these modalities enables multimodal SciKGs, which enrich symbolic triples by anchoring numerical, visual, temporal, and topological evidence as attributes to the semantic entities. Crucially, these multimodal data serve as grounding context rather than replacing the ontological backbone of the SciKG.

3.2 Knowledge Extraction

Extracting entities, relations, and scientific events from heterogeneous data remains one of the core challenges in constructing SciKGs. Unlike general-purpose information extraction, scientific data involves highly specialized terminologies, nested relationships, and evolving conceptual frameworks that demand fine-grained semantic understanding and context-aware reasoning. Consequently, the design of SciKG extraction pipelines needs to reconcile three often conflicting requirements: precision, scalability, and adaptability to emerging knowledge.

Traditional rule- and ontology-based methods [36,37] represent the earliest efforts toward structured knowledge extraction in domains such as biomedicine and chemistry. These systems leverage domain-specific dictionaries, handcrafted rules, and curated ontologies (e.g., Gene Ontology) to identify entities and align them with controlled vocabularies. Their advantages lie in interpretability, reproducibility, and high precision within well-defined subdomains. However, they suffer from limited scalability, domain transferability, and poor adaptability to newly emerging concepts.

Data-driven NLP approaches [38,39] have since transformed the field by enabling automated, large-scale extraction from unstructured scientific text. Techniques such as named entity recognition, relation extraction, and event detection have been adapted to domain corpora using models like SciBERT [40] and domain-tuned transformers. These models outperform rule-based systems in recall and generalization, particularly when combined with weak supervision or self-training. More recently, LLMs have revolutionized scientific information extraction [41]. Through few-shot prompting and task-specific

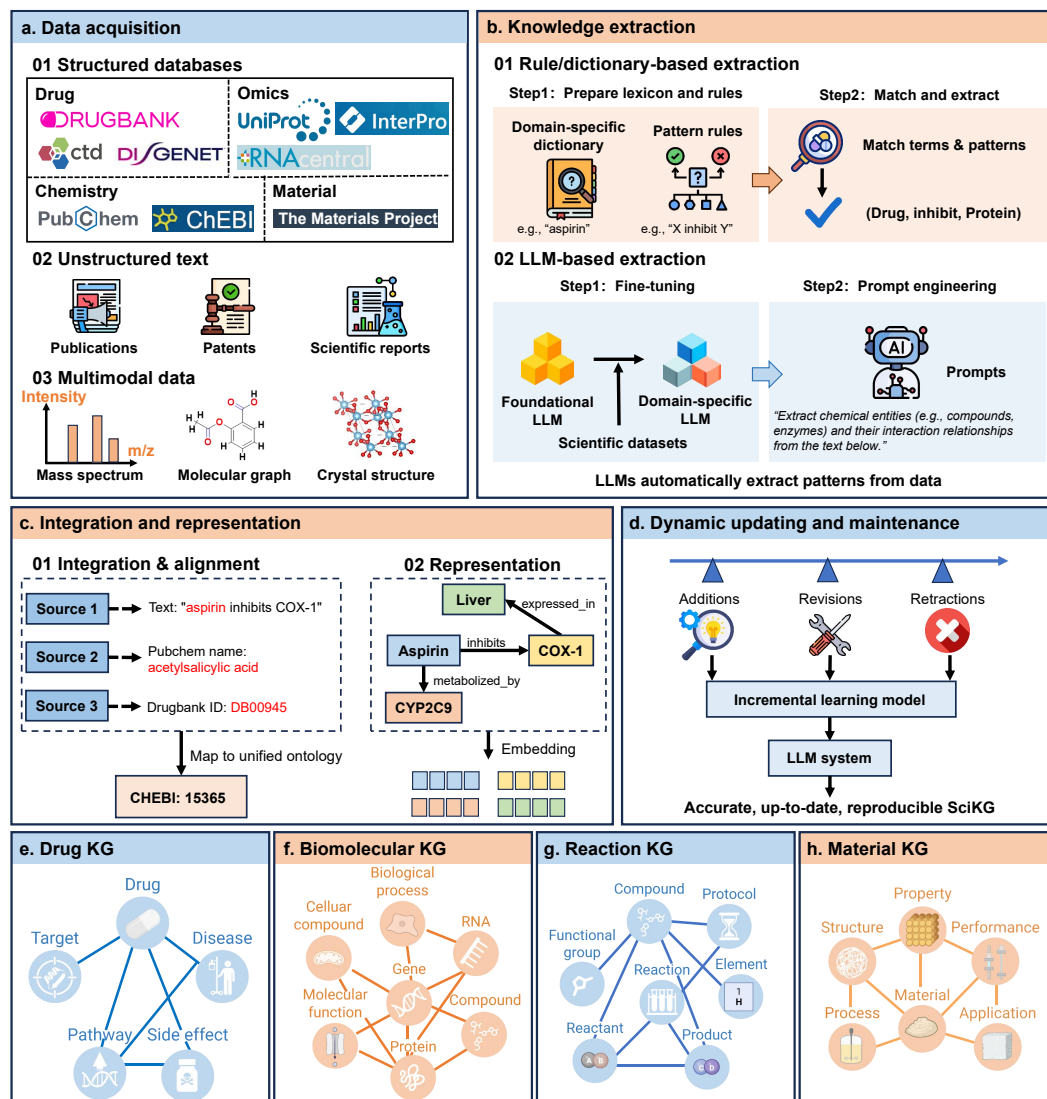


Figure 4. Construction and maintenance of SciKGs. (a) The foundation of SciKG construction involves integrating diverse data sources, including structured databases, unstructured text, and multimodal data. (b) Two main approaches for extracting entities and relations from the acquired data are illustrated: rule/dictionary-based extraction, which relies on predefined lexicons and rules, and LLM-based extraction, involving fine-tuning on scientific datasets and prompt engineering. (c) Ontology alignment integrates diverse representations of the same entity (e.g., aspirin), followed by graph embedding into a continuous vector space. (d) Dynamic updating through incremental learning and LLM-driven error correction ensures SciKGs remain accurate and up to date. (e-h) Sub-figures illustrate representative examples of specialized knowledge graphs for drugs, omics, chemicals, and materials, respectively.

fine-tuning, LLMs can recognize novel entities, infer implicit relations, and even generate structured hypotheses from textual evidence, bridging the gap between symbolic extraction and conceptual understanding.

Hybrid and semi-automated pipelines [42] are emerging as practical solutions to reconcile the conflicting requirements of precision, scalability, and adaptability (see Table S3 for a detailed comparative analysis of these paradigms). By combining the domain precision of ontology-based methods with the generalization of neural models, these frameworks address the intrinsic

limitations of individual approaches. For instance, rule-based prefilters can identify candidate entities with high confidence, which are then refined using transformer-based relation classifiers. Conversely, neural models can suggest new candidate relations or ontological extensions, which are validated against existing controlled vocabularies. This synergy reduces both annotation costs and error propagation, while maintaining interpretability, an essential requirement for scientific credibility and trustworthiness.

3.3 Integration and Representation

After the extraction of entities and relations, a critical challenge lies in transforming fragmented knowledge into a coherent and semantically consistent structure. Integration and representation serve as the foundation for building interoperable and computationally tractable SciKGs. To achieve semantic consistency across heterogeneous data sources, ontology alignment and schema matching techniques [43] are widely employed. These methods harmonize terminologies, reconcile conceptual discrepancies, and enable cross-domain reasoning. Recent frameworks increasingly adopt federated ontology mapping [44] and probabilistic schema alignment [45], which tolerate terminological uncertainty and facilitate large-scale integration across biomedical, chemical, and materials databases.

After achieving semantic interoperability, the next step is to encode the integrated graph into representations that support computation and reasoning. Representation learning has emerged as a key paradigm for capturing the relational and structural regularities within scientific knowledge. Graph-based embedding methods [20] project nodes, relations, and subgraphs into continuous vector spaces, preserving both topological proximity and semantic dependencies. These representations bridge the gap between symbolic integration and data-driven inference, enabling efficient similarity computation, link prediction, and hypothesis generation. However, the choice of representation architecture significantly impacts downstream utility. Shallow KGEs (e.g., TransE) are computationally efficient and effective for massive, relation-dense networks (e.g., drug interactions), but often struggle with inductive inference on unseen entities. In contrast, GNNs aggregate neighborhood features to support inductive reasoning, making them superior for structure-rich domains like molecular chemistry. Recently, LLM-based encodings have emerged to capture fine-grained semantic nuance from unstructured text, offering strong zero-shot capabilities albeit with higher inference latency compared to structural methods.

Moreover, scientific knowledge is often multimodal, with entity descriptions appearing in various forms such as text, microscopy images, molecular graphs, or time-series experimental measurements. To capture these characteristics, methods such as cross-modal embedding [46] are introduced to model interactions across heterogeneous data types effectively. These methods align heterogeneous modalities into shared

latent spaces, enabling unified reasoning over textual, visual, and structural data, and even support temporal inference over evolving scientific processes.

3.4 Updating and Maintenance

Scientific knowledge is continuously evolving, with new discoveries, revised findings, and retracted claims constantly reshaping the landscape. Thus, SciKGs must be designed as dynamic and adaptive systems rather than static repositories. Incremental learning approaches [47] allow for the seamless integration of new data while minimizing catastrophic forgetting of prior knowledge. Beyond algorithmic updating, community and human-in-the-loop mechanisms are essential for maintaining trustworthiness. Crowdsourced and expert-driven curation initiatives exemplified by biomedical resources such as UniProt and the Gene Ontology demonstrate that combining automated extraction with domain expertise yields more accurate and interpretable updates.

Meanwhile, automated verification and maintenance pipelines are increasingly powered by LLMs agents [48]. These agents can detect inconsistencies, contradictions, and obsolete links by comparing textual evidence, citation networks, or temporal trends. Automated correction mechanisms then propagate verified updates across dependent nodes and relations, improving graph coherence and reducing cumulative error. Integration with provenance metadata and version control frameworks further ensures reproducibility and traceability, which are core requirements for scientific accountability.

3.5 Evaluation

The utility of a SciKG fundamentally depends on its quality and reliability. Evaluation strategies typically span three granularity levels to ensure rigorous validation (see Table S4 for a comprehensive taxonomy of metrics and benchmarks):

(I) *Component-Level Evaluation*: This dimension focuses on the fidelity of the construction pipeline. The accuracy of entity and relation extraction is standardly assessed via Precision, Recall, and F1-scores against gold-standard corpora, while ontology alignment correctness is evaluated using reference mappings to ensure semantic consistency across heterogeneous sources.

(II) *Graph-Level Representation*: Beyond individual facts, the structural adequacy of the KG is critical for downstream inference. Metrics

such as graph density and connectivity are analyzed alongside embedding quality. Link prediction benchmarks are widely employed to measure how well the graph captures latent scientific associations using metrics like Mean Reciprocal Rank (MRR) and Hits@k.

(III) *Trustworthiness and Utility*: In the context of AI for Science, evaluation extends to trustworthiness, specifically provenance coverage and temporal consistency. Ultimately, the representational adequacy of a SciKG is validated extrinsically by its performance uplift in specific scientific tasks, such as drug repurposing or material property prediction.

3.6 Summary and Prospects

The construction of SciKGs is no longer a one-time curation effort but an ongoing, adaptive process that bridges structured databases, unstructured literature, and rich multimodal evidence. The rise of multimodal SciKGs marks a pivotal shift toward more holistic, quantitative, and interpretable scientific knowledge infrastructures. The integration of LLMs and advanced multimodal approaches has further accelerated KG construction by enabling automated entity recognition, relation extraction, and error correction at unprecedented scale and speed. Future directions include: (1) standardizing multimodal KG schemas across domains; (2) enabling real-time KG evolution via autonomous LLM agents; and (3) fostering open, community-governed platforms for collaborative KG curation. As SciKGs continue to evolve from static repositories to dynamic and adaptive knowledge infrastructures, these advances will directly improve the efficiency, accuracy, and comprehensiveness of KG construction, laying the foundation for more reliable and integrative scientific knowledge representation.

4 APPLICATIONS OF SCIENTIFIC KNOWLEDGE GRAPHS

Knowledge graphs organize multi-source scientific knowledge into linked, computable structures that support data-driven reasoning in complex scientific problems. In this section, we highlight representative applications of SciKGs in four domain tasks: drug development and optimization, omics interpretation and analysis, chemical reaction and synthesis, and materials design and discovery (Fig. 5). These applications demonstrate how SciKGs serve as scientific discovery engines by facilitating inference, prediction, and decision-making processes.

4.1 Drug Development and Optimization

The drug development process is prone to high attrition due to fragmented data, intricate biological mechanisms, and limited translational fidelity between preclinical and clinical phases. KGs address these challenges by integrating molecular, phenotypic, clinical, and literature-based information into semantically rich networks, enabling coherent reasoning across drugs, targets, and diseases. In drug repurposing, KGs uncover non-obvious drug-disease relationships by synthesizing multi-omics, literature, and pathway data, supporting mechanistic inference and rapid candidate identification for rare diseases and epidemic contexts [49,50]. For instance, the TxGNN model [49], pre-trained on a large-scale medical KG, enables zero-shot prediction across more than 17,000 diseases, demonstrating how KG-derived representations can generalize to diseases with no known treatments. For drug-drug interaction prediction, heterogeneous graphs capture chemical similarity, shared targets, and physiological effects, while subgraph learning and knowledge-enhanced models reveal underlying mechanisms [32,51]. As exemplified by the DDI-GPT framework [32], the integration of KG-derived features with LLMs not only achieves high predictive accuracy but also provides biologically interpretable explanations for the predicted interactions. In drug-target interaction tasks, KGs integrate sequence, structural, and semantic information, employing attention-based graph neural networks to capture topological dependencies and uncover actionable drug-target pairs [52,53]. The DTINet framework [52] addresses this by learning low-dimensional representations from a heterogeneous network to predict novel drug-target interactions, with several predictions for cyclooxygenase inhibitors subsequently receiving experimental validation. Beyond interaction prediction, KGs enhance virtual screening, drug recommendation, and toxicity assessment by linking patient-specific records, molecular profiles, and clinical outcomes, enabling personalized, mechanism-driven decision-making and reducing reliance on costly experimental assays [54,55]. For example, frameworks like GAMENet [54] integrate DDI knowledge graphs with patient records to recommend safe, personalized medication combinations. Meanwhile, in toxicity assessment, dedicated resources such as ReproTox-KG [55] profile and predict compound-induced birth defect risks by constructing a specialized knowledge graph. Overall, KGs provide a structured, interpretable,

and integrative framework that accelerates drug discovery, optimizes therapeutic strategies, and improves safety evaluation throughout the drug development pipeline.

4.2 Omics Interpretation and Analysis

Omics research, encompassing genomics, transcriptomics, proteomics, metabolomics, and microbiomics, underpins systems biology by elucidating molecular architectures of health and disease through multi-scale datasets. The intrinsic complexity and fragmentation of omics data have historically impeded mechanistic insight, but KGs provide a semantically rigorous framework to model entities (e.g., genes, proteins, metabolites, microbes) and their contextual relationships (e.g., regulatory cascades, functional associations, pathophysiological links), enabling cross-disciplinary reasoning, integrative analysis, and interpretable hypothesis generation. In genomics, KGs integrate genetic variants, regulatory elements, and phenotypic data to move beyond pre-defined candidate lists towards systems-level inference of gene regulatory mechanisms and pathogenic variants [56–58]. This capability is demonstrated by Phe-noKG [58], which leverages graph neural networks to directly infer causative genes from patient phenotypes, providing a powerful framework for rare disease diagnosis without relying on pre-curated gene panels. Transcriptomic KGs model spatial and regulatory intricacies, capturing intercellular signaling and RNA-mediated regulation through structured representations of ligand-receptor-target pathways and RNA-interaction networks [59,60]. The RNA-KG resource [60] exemplifies this approach, integrating over 60 databases into an ontology-grounded framework that enables systematic exploration of the “RNA world” and its functional implications. Proteomics benefits from graph-based meta-path analyses linking proteins to disease-associated risk genes, facilitating functional annotation, biomarker validation, and therapeutic target prioritization [61,62]. The CKG [62] serves as a prime example of a scalable platform for this purpose, integrating millions of relationships from public databases and literature to statistically contextualize clinical proteomics data, thereby accelerating the interpretation of biomarker studies and informing clinical decision-making. In metabolomics, KGs contextualize metabolite perturbations within biological networks, uncovering associations with disease phenotypes and supporting biomarker discovery. The FORUM KG [63] addresses the

central challenge of biological interpretation in metabolomics by semantically integrating disparate data sources, using ontological reasoning to infer novel metabolite-disease associations and generate testable biological hypotheses from experimental signatures. Microbiomic KGs map ecological and functional interactions between microbial populations, metabolites, and host physiology, informing microbe-based therapeutic strategies, as seen in resources like MMiKG [64] and MiKG [65]. In multi-omics research, KGs unify transcriptomic, proteomic, and metabolomic layers into cohesive networks, enabling systems-level modeling of biological processes, for instance, predicting cancer metastasis by integrating graph models with physics-informed constraints [66]. This integrative approach transforms disjointed omics profiles into interpretable, mechanism-driven models that accelerate translational discovery.

4.3 Chemical Reaction and Synthesis

Chemical reaction mechanism elucidation and compound synthesis are increasingly driven by data- and knowledge-centered approaches. KGs structure chemical entities, reactions, intermediates, and properties into graph-based networks, enabling reasoning that accelerates hypothesis generation, predictive modeling, and synthesis planning. For reaction prediction, molecules and reactions are represented as nodes and edges, capturing structural similarity, reactivity principles, and catalytic dependencies; graph inference and semantic learning facilitate the identification of feasible synthetic routes, prediction of products, intermediates, and by-products, and improved reaction classification and yield estimation [67,68]. For instance, the ReaKE framework [68] constructs a chemical synthesis KG to enable contrastive learning that improves reaction classification and product prediction by capturing functional group transformations. Similarly, CatKG [69] integrates structured reaction databases to model reactant-catalyst-product relations. Through word embeddings and masked language modeling, it supports both analogical reasoning and direct catalyst prediction. In synthesis pathway optimization, KGs integrate data on reactant availability, catalysts, and yields, allowing multi-criteria reasoning to identify cost-effective, efficient, and sustainable routes. Knowledge-enhanced algorithms and language models are integrated to support dynamic adaptation to new reaction data [8,10,70]. The work by Li et al. [70] demonstrates this by building a reaction network KG

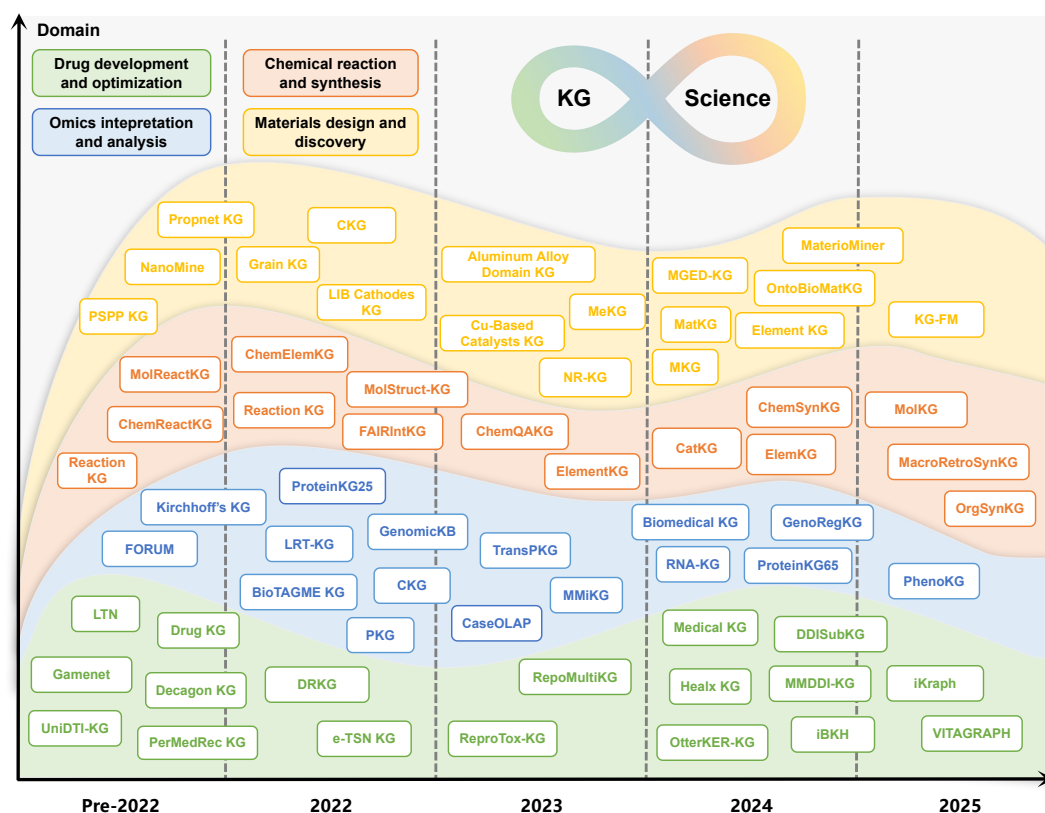


Figure 5. A taxonomy of existing SciKGs and their applications in science. More details of these SciKGs are represented in Supplementary Tables 5-8. An extended comparison of SciKGs across six dimensions (construction methodologies, benchmarking performance, practical usability, scalability, reproducibility, and data quality) is available at our open-access repository: <https://github.com/hicai-zju/scikgs>.

from historical data to quantify synthetic accessibility, providing a knowledge-based filter for prioritizing synthesizable compounds in molecular design. Moreover, molecular property prediction benefits from situating molecules within networks that link structural features, functional groups, and experimental properties, enabling enhanced representation of non-bonding interactions and contextual reasoning for solubility, reactivity, and bioactivity, even with limited datasets [5,71,72]. The GODE framework [71] exemplifies this approach by fusing molecular graphs with biochemical KGs through contrastive learning, significantly enhancing property prediction accuracy by leveraging structured domain knowledge. Overall, these KG-based methods provide mechanistic insights and data-driven decision-making in both fundamental and applied chemistry.

4.4 Materials Design and Discovery

Material discovery aims to uncover intrinsic links between composition, microstructure, and functional properties to accelerate design, performance tuning, and scalable application of

advanced materials. KGs address the challenge of integrating multi-scale and heterogeneous data by structuring entities and their relationships into semantically rich networks, converting disparate information into actionable knowledge for targeted design, accurate property prediction, and efficient screening [73,74]. In new material design, KGs guide innovative strategies by embedding fundamental principles such as atomic bonding, crystal symmetry, and structure-property correlations, enabling the identification of promising candidates in energy, catalysis, and nanocomposite materials [75–77], and supporting multi-agent reasoning frameworks for biomimetic materials [31]. For material performance prediction, KGs integrate elemental, structural, and processing information to infer unknown properties through graph traversal and logical reasoning, outperforming conventional simulations in predicting key indicators such as thermal conductivity, mechanical strength, bandgap, and formation energy [78,79]. In screening and optimization, KGs consolidate millions of entities across databases, literature, and experiments to prioritize candidates with desired characteristics, guide experimental de-

sign, and balance performance with production feasibility in piezoelectric materials, ultra-high-performance concrete, or COFs for gas storage [80–83]. In summary, KGs accelerate materials research by unifying heterogeneous data, enabling predictive performance modeling, and supporting systematic design, screening, and optimization of advanced functional materials.

4.5 Summary and Prospects

Across the diverse domains of drug discovery, omics analysis, chemical synthesis, and materials design, SciKGs have demonstrated their pivotal role as the connection of modern scientific intelligence. As summarized in Fig. 6, SciKGs provide a unified framework for knowledge organization, transforming fragmented, multi-source scientific data into coherent, machine-interpretable structures that enable holistic reasoning across molecular, biological, and material hierarchies. Through knowledge embedding, SciKGs capture both symbolic semantics and latent correlations, bridging structured ontologies with continuous representations to support predictive and generative modeling. Their graph-based topology further enables causal inference and discovery, allowing automated identification of hidden relationships, such as drug repurposing candidates that would be difficult to infer from isolated datasets. Beyond predictive performance, SciKGs enhance interpretability and transparency, grounding model outputs in explicit knowledge pathways and causal relationships, thus aligning data-driven predictions with scientific rationale.

While this survey primarily focuses on domains where discrete entities and mature data infrastructures have facilitated early adoption, the underlying graph-structuring paradigm is broadly extensible to the wider AI for Science landscape. Promising frontiers include Physics, where researchers structure physical laws as graph constraints for neural operators, and Earth Science, which involves integrating geochemical and atmospheric data to model complex climate feedbacks. These emerging fields introduce distinct challenges, such as handling data sparsity and integrating continuous physical fields with symbolic graphs. Addressing these hurdles will not only broaden the application of SciKGs but also drive the technical evolution of more robust and generalizable knowledge infrastructures.

Looking forward, the next evolution of SciKGs will hinge on cross-disciplinary integration, adaptive intelligence, and embodied reasoning. From a systems perspective, future

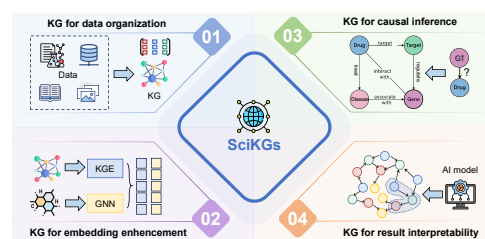


Figure 6. Summary of core functions of SciKGs in diverse scientific tasks. SciKGs serve as a foundational infrastructure that: (1) organizes heterogeneous scientific data into structured knowledge; (2) enhances representation learning via graph embedding; (3) enables causal and relational inference for hypothesis generation; and (4) improves AI model interpretability by grounding predictions in traceable, evidence-based knowledge paths.

SciKGs should transcend domain boundaries, linking domain-specific knowledge, ranging from atomic interactions to macroscopic physical systems, into interoperable meta-graphs that capture the full continuum. This cross-domain integration addresses critical data scarcity issues by constructing a unified semantic space. Theoretically underpinned by schema alignment and structural isomorphism, this approach enables hypothesis migration, where logic from data-rich domains (e.g., pharmaceuticals) guides discovery in data-scarce domains (e.g., materials). Recent works like SciAgents [31] have demonstrated this by leveraging biological mechanisms to inform the generative design of sustainable composites. Furthermore, integrating multi-modal and mechanistic knowledge will foster a new generation of explainable and scientific AI. This convergence of symbolic knowledge, statistical learning, and experimental validation points toward a unified paradigm of knowledge-centric scientific discovery, where SciKGs become both the foundation and the evolving fabric of intelligent, interpretable, and autonomous science.

5 SYNERGIZING KNOWLEDGE GRAPHS AND LARGE LANGUAGE MODELS

The advancement of scientific discovery increasingly depends on combining knowledge bases with generative intelligence (e.g., LLMs) [31,32, 84]. KGs provide explicit representations of entities, relations, and domain knowledge, while LLMs offer powerful capabilities in reasoning, abstraction, and summary [6,7]. While the integration of KGs and LLMs has been actively explored in general AI contexts, we herein propose a specialized taxonomy tailored for scientific discovery. By synthesizing recent literature, we conceptualize a unified collaborative framework where SciKGs serve as the foundational

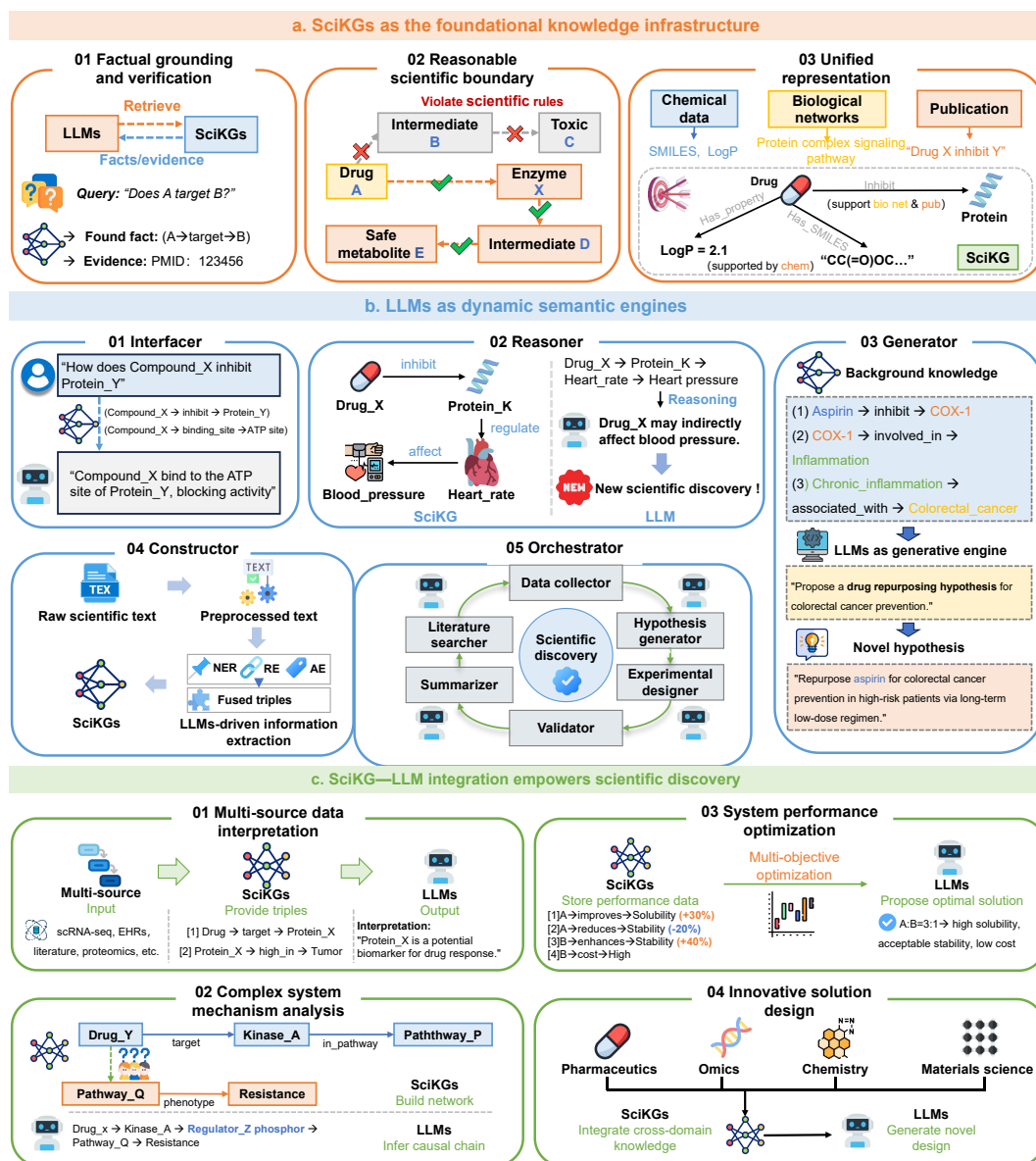


Figure 7. Synergistic integration of SciKGs and LLMs for knowledge-driven scientific discovery. (a) SciKGs serve as the foundational knowledge infrastructure by ensuring factual grounding and verification, defining reasonable scientific boundaries, and enabling unified representation of heterogeneous data. (b) LLMs act as dynamic semantic engines through five core functions: semantic interface for knowledge access, analytical reasoner for inference, generative engine for hypothesis design, constructor for knowledge curation, and orchestrator for workflow automation. (c) The SciKG-LLM integration empowers four key scientific discovery tasks: multi-source data interpretation, complex system mechanism analysis, system performance optimization, and innovative solution design.

knowledge infrastructure and LLMs as dynamic semantic engines (Fig. 7, Table 1). This synergy enables knowledge-grounded, interpretable, and adaptive solutions to complex scientific problems. In this section, we dissect this complementary relationship and its application across the scientific discovery pipeline.

5.1 SciKGs as the Foundational Knowledge Infrastructure

Traditional LLMs are prone to hallucinations during scientific reasoning, such as generating non-existent drug-target interactions, which leads to outputs lacking factual support [85]. Moreover, LLMs struggle with grounding in physical constraints, reliable generation of symbolic structures, and robustness to domain shifts, limitations that are particularly concerning in safety-critical applications like drug discov-

ery. Leveraging their explicit entity-relationship structure, SciKGs constrain LLM reasoning from three key perspectives to ensure the reliability of scientific decision-making. First, SciKGs ensure factual grounding, verification, and precise retrieval. They serve as authoritative benchmarks against which LLM-generated hypotheses can be validated, directly countering the hallucination and factual inconsistency. For instance, the KNOWNET framework [86] extracts triples from LLM outputs and maps them to validated evidence in external KGs, providing a visual interface to trace and verify claims. Similarly, FactFinder [87] augments LLMs with a medical KG through a structured retrieval-and-generation pipeline, which retrieves precise sub-graphs to focus LLM attention, demonstrating significant improvements in the accuracy and completeness of responses for critical tasks like target identification. By accessing pre-stored knowledge of established scientific mechanisms, these systems assess the plausibility of proposed ideas and provide traceable evidence, directly countering the opaque nature of black-box LLM reasoning [88].

The efficacy of these systems, however, is inherently tied to the completeness and veracity of the KG itself; missing or erroneous facts in the KG can lead to false validation or missed detections. By accessing pre-stored knowledge of established scientific mechanisms, these systems assess the plausibility of proposed ideas and provide traceable evidence, directly countering the opaque nature of black-box LLM reasoning [88]. Second, SciKGs define reasonable boundaries for explainable causal reasoning. They prevent the generation of schemes that violate scientific principles, thus addressing the LLM's lack of intrinsic grounding in physical laws and domain constraints. The Graph-Constrained Reasoning (GCR) framework [89] integrates the KG structure directly into the LLM's decoding process, ensuring that every step of the reasoning path is faithful to the graph to prevent attention drift, and achieving zero reasoning hallucination on knowledge graph question-answering tasks. In chemistry, the Synergizing KG and LLM approach [90] for relay catalysis uses a detailed catalysis KG (Cat-KG) to apply expertise-informed scoring rules, ensuring that only chemically plausible multi-step reaction pathways are recommended, thereby constraining the LLM's generative space to scientifically valid outcomes. Finally, advanced multi-modal SciKGs integrate heterogeneous data into a unified framework, allowing LLMs to perform cross-modal reasoning

and holistic analyses and mitigating their brittleness to domain shifts. Systems like DDI-GPT [32] exemplify this by constructing a multimodal KG that fuses drug-related chemical, substructure, and molecular data. This rich, structured context enables the LLM to not only predict drug-drug interactions with high accuracy but also to generate explainable insights grounded in the multifaceted evidence from the graph, providing a scaffold for more reliable symbolic and structured reasoning. Beyond these functional capabilities, SciKGs offer a fundamental advantage in dynamic knowledge evolution. Unlike foundation models, which require computationally expensive retraining to absorb new findings, SciKGs support incremental, low-cost updates, enabling the real-time integration of emerging discoveries. In the era of powerful foundation models, SciKGs are therefore not optional but essential infrastructure. They act as the irreplaceable deterministic substrate that complements the probabilistic nature of LLMs, and they provide the explicit provenance, symbolic logic, multimodal consistency, and dynamic evolvability required for rigorous AI-driven science.

5.2 LLMs as Dynamic Semantic Engines

Despite their strengths in structured representation, SciKGs are inherently static, presenting a fundamental challenge for dynamic scientific exploration [91]. LLMs bridge this gap by serving as dynamic semantic engines that transform static knowledge into actionable scientific intelligence. This transformation is achieved through several core capabilities. First, LLMs act as semantic interfaces, parsing complex SciKGs and converting structured scientific data into intuitive natural language summaries and precise formal queries. Systems like HeCiX [92] demonstrate this by integrating a biomedical KG with GPT-4 through LangChain, creating a natural language interface that enables researchers to efficiently query complex clinical trial and biological data. This dramatically lowers the barrier to cross-domain knowledge acquisition and facilitates interdisciplinary collaboration. Second, they function as analytical reasoners, performing complex inference and prediction tasks based on the rich relational structures of SciKGs to uncover novel mechanistic insights. The DDI-GPT framework [32] exemplifies this capability, where an LLM enhanced with a multimodal drug KG not only predicts drug-drug interactions with high accuracy but also generates explainable insights by capturing contextual dependencies between biomedical entities. Third,

they serve as generative engines for scientific innovation, producing novel, plausible hypotheses, experimental strategies, and design solutions that are grounded in structured knowledge. SciAgents [31] showcases this through a multi-agent system that autonomously generates and refines research hypotheses for bioinspired materials discovery by leveraging large-scale ontological knowledge graphs. Similarly, the automated retrosynthesis planning system by Ma et al. [10] demonstrates how LLMs can design novel synthesis pathways for macromolecules by extracting and reasoning over chemical reaction data stored in KGs. Furthermore, LLMs undertake a constructive role by building, curating, and maintaining SciKGs from raw scientific literature and data. The comprehensive biomedical KG iKraph [9] exemplifies this, where an LLM-powered information extraction pipeline processes all PubMed abstracts to construct a large-scale KG that matches human expert annotations. The KG-RAG framework [84] further demonstrates how LLMs can optimize knowledge extraction from biomedical KGs in a token-efficient manner, while systems like UpToDate [93] show how LLMs can automatically validate and update KG facts to maintain currency. Finally, in their most advanced role, LLMs orchestrate complex scientific workflows, managing multi-step reasoning processes and coordinating multi-agent systems. The ESCAR-GOT agent [94] exemplifies this by combining LLMs with a dynamic Graph of Thoughts and biomedical KGs to significantly outperform standard retrieval-augmented generation methods in open-ended biomedical questions. These functions position LLMs as active collaborators in scientific discovery, transcending mere information retrieval.

5.3 SciKG–LLM Integration Empowers Scientific Tasks

Built on the complementary roles of factual anchors and semantic engines, the SciKG–LLM synergy framework can systematically address four core tasks in scientific discovery, covering the full workflow from fundamental cognition to applied breakthroughs (Table 1). During multi-source data interpretation, SciKGs convert massive datasets into structured triples, and LLMs extract interpretable knowledge to unlock the latent value of data accumulation [51,94]. For complex system mechanism analysis, SciKGs integrate multi-source data to construct entity-relationship networks, and LLMs infer causal chains based on these networks to address the

limitation of traditional methods that prioritize phenomenological description over causal modeling [86,87,95]. In system performance optimization, SciKGs store quantitative variable–performance correlations, and LLMs generate multi-objective optimal solutions by incorporating domain constraints to overcome the local optimization trap of trial-and-error iteration [32,84,90]. For innovative scheme design, SciKGs integrate cross-domain knowledge, and LLMs generate new schemes that integrate multi-disciplinary principles through analogical reasoning to break the innovation lag caused by domain barriers [9,10]. Overall, these four tasks constitute a self-reinforcing discovery feedback loop. The process begins with multi-source data interpretation, which feeds structured knowledge into the SciKG. This enriched graph enables deeper complex system mechanism analysis, whose insights guide system performance optimization. The optimized systems and understood mechanisms then fuel innovative solution design, which, when experimentally validated, generates new data and knowledge, thus closing the loop and beginning the cycle anew.

5.4 Toward Autonomous Scientific Discovery Paradigm

The synergistic integration of SciKGs and LLMs heralds a paradigm shift in scientific methodology, from human-driven hypothesis–validation cycles to AI-augmented autonomous discovery loops. In this emerging paradigm, LLMs continuously generate and refine hypotheses from massive multi-modal data; SciKGs evaluate and ground these hypotheses against existing knowledge; and validated results are automatically integrated back into the SciKG, forming an ever-growing knowledge flywheel. This closed feedback loop enables a self-evolving scientific ecosystem capable of accelerating discovery at scale. The concrete embodiment of this framework is the *AI Scientist Copilot*: an autonomous system that embeds the SciKG–LLM synergy within a perception-cognition-action loop to assist and augment the scientific process (Fig. 8). Such a copilot is capable of sensing, reasoning, and acting across the full discovery pipeline:

(I) *From Data to Knowledge*: LLMs act as perceptual organs that “understand” scientific literature and multimodal data, while SciKGs provide the structured schema to organize this extracted information. This synergy transforms raw data into computable knowledge across domains. For instance, in biomedicine, iKraph [9] employs LLMs to extract entities and relations

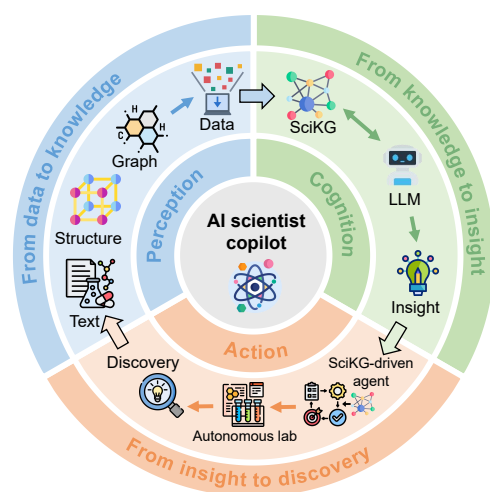


Figure 8. The autonomous scientific discovery flywheel driven by LLM agents and SciKGs.

from PubMed abstracts, constructing a comprehensive biomedical KG comparable to human curation. Similarly, in chemistry, LLMs parse reaction literature to populate KGs with precise synthesis conditions [70], while in materials science, frameworks like MatKG [74] utilize LLMs to integrate heterogeneous data from publications and simulations into unified graphs. In these systems, LLMs unlock the information, and SciKGs store it in a rigorous, machine-interpretable format.

(II) From Knowledge to Insight: SciKGs function as long-term memory and logical engines, while LLMs perform analogical reasoning and path inference. Their collaboration enables verifiable reasoning chains: LLMs hypothesize causal links, and SciKGs validate or refine these paths with literature evidence or alternative mechanisms. Such collaborative reasoning drives insights across domains: it generates biologically grounded explanations for drug-drug interactions (e.g., DDI-GPT [32]), identifies chemically plausible reaction pathways in catalysis (e.g., Cat-KG [90]), and contextualizes property predictions for new materials (e.g., MechGPT [96]). This bidirectional flow, where LLMs propose and SciKGs ground, ensures that generated insights are both innovative and scientifically sound.

(III) From Insight to Discovery: LLMs operate as strategy planners, designing experiments, synthesis routes, or material compositions; SciKGs act as feasibility filters, ensuring that proposed actions comply with known scientific principles. This is advanced by systems like SciAgents [31], where a multi-agent framework reasons over a materials knowledge

graph for bioinspired design; Automated Retrosynthesis Planning [10] that uses reaction KGs to constrain and validate LLM-generated synthesis pathways; and SciToolAgent [97], which leverages a scientific tool knowledge graph to automatically orchestrate the execution of complex analytical workflows (e.g., protein design, chemical reactivity prediction, and MOF materials screening). The physical realization of this loop is now being demonstrated by autonomous robotic platforms, such as multi-agent driven robotic AI chemists that conduct closed-loop chemical research on demand [98]. Coupled with automated laboratory systems, this architecture can close the loop from computational inference to physical experimentation.

5.5 Summary and Prospects

In summary, SciKG–LLM synergy marks a conceptual leap from knowledge utilization to knowledge evolution. By combining SciKGs with the generative and reasoning power of LLMs, future scientific ecosystems may operate as continuously learning systems, capable of generating, testing, and consolidating knowledge without constant human supervision.

Realizing this vision, however, requires a clear-eyed assessment of the paradigm's inherent fragility. The reliability of SciKG–LLM systems is co-dependent: the veracity of the SciKG (subject to incompleteness, noise, and ontological misalignment) directly grounds or misguides LLM reasoning; conversely, LLM imperfections (hallucinations, biases, over-generalization) can propagate errors during knowledge extraction, curation, and generative design, potentially corrupting the very knowledge infrastructure they rely on. This bidirectional risk is amplified in autonomous discovery loops, where errors may compound. Developing robust evaluation protocols to measure factuality, reasoning robustness, and performance gains, which are summarized in Table S9, is therefore a critical step toward trustworthy autonomous discovery.

These inherent vulnerabilities become particularly acute as systems evolve from passive retrieval toward active, autonomous generation. As these systems evolve toward autonomy, a critical frontier lies in developing guardrails for generative science. Existing works have yet to fully resolve how to prevent plausible but scientifically invalid hallucinations in experimental design. Future research must focus on establishing a tiered constraint framework: (1) *Feasibility Checks*, where SciKGs act as physical logic filters (e.g., verifying reagent compatibility); (2)

Safety & Ethical Compliance, integrating toxicity and biosafety protocols directly into graph attributes; and (3) *Scientist-in-the-Loop Governance*, ensuring that high-stakes autonomous decisions trigger expert-review checkpoints. Addressing these validation challenges is a prerequisite for deploying trustworthy AI scientist copilots.

The next frontier lies in building these robust, safety-aware, LLM-based scientist copilots that embody this closed-loop intelligence, integrating real-time experimental feedback with cognitive models to achieve a new era of autonomous scientific discovery.

6 LIMITATIONS, CHALLENGES, OPPORTUNITIES, AND FUTURE DIRECTIONS

Despite the growing success, SciKGs remain in an early stage of development. Their evolution is shaped not only by technical challenges but also by fundamental epistemological constraints and practical barriers. In this section, we first discuss the inherent limitations of the SciKG paradigm itself. We then discuss four major technical challenges: data quality, interoperability, dynamism, and trustworthy reasoning, and highlight promising opportunities for advancing SciKGs (Fig. 9). We further propose three complementary directions for next-generation SciKGs, aimed at enhancing their role as actionable knowledge infrastructures for scientific discovery.

6.1 Inherent Limitations of SciKGs

Oversimplification of Continuous Processes. The discrete triple structure (subject–predicate–object) inherent to KGs often fails to capture the continuous, dynamic nature of scientific processes. For example, a simple “drug–inhibits–protein” relation may overlook critical contextual factors such as dosage dependence, temporal dynamics, spatial localization, or reaction kinetics. This structural limitation can lead to a loss of mechanistic detail, restricting the graph’s ability to fully represent complex biological, chemical, or physical phenomena.

Epistemic Uncertainty and Retraction Handling. Scientific knowledge is continually revised as new evidence emerges, yet most SciKGs treat extracted relations as static facts. Current frameworks lack built-in mechanisms to represent confidence scores, conflicting hypotheses, or negative results. Moreover, when published findings

are retracted or corrected, propagating those updates through the graph while ensuring downstream models disregard invalidated triples remains an unresolved challenge. This can lead to persistent error propagation and reduce the reliability of KG-driven predictions.

Domain Imbalance and Bias. SciKGs constructed from literature inevitably inherit the severe research biases present in scientific publishing. Well-studied entities (e.g., certain disease-associated genes or high-performance materials) accumulate dense connections, while under-researched areas remain sparse. This “rich-get-richer” topology can skew graph algorithms, such as embedding methods or link prediction models, toward highly connected nodes, potentially overlooking novel interactions or under-representing emerging scientific domains.

6.2 Technical Challenges

Data Quality and Completeness. The effectiveness of SciKGs depends critically on the quality, consistency, and coverage of the underlying data. Scientific data are often incomplete, noisy, or biased, reflecting variations in experimental protocols, reporting standards, and publication practices. For example, biomedical databases may lack negative results, while materials datasets may disproportionately emphasize high-performing compounds. Integrating heterogeneous sources further compounds these issues, as differences in terminology, granularity, and measurement standards lead to inconsistencies that propagate through the graph. Ensuring robust data quality requires advances in automated curation, normalization, and error detection, as well as community-wide efforts to establish minimal reporting standards and promote the publication of negative and null results. Ultimately, improving data quality and completeness will determine whether SciKGs can provide trustworthy substrates for scientific reasoning.

Interoperability and Integration. A second challenge lies in the integration of knowledge across diverse scientific disciplines. Existing SciKGs are often domain-specific, relying on bespoke ontologies that impede interoperability across biology, chemistry, and materials science. Beyond bespoke schemas, integration faces dual hurdles: technical heterogeneity and practical access barriers. Technically, scientific concepts often suffer from cross-disciplinary polysemy (e.g., “nucleus” in cell biology vs. physics) and topolog-

Table 1. Representative SciKG–LLM integration systems, categorized by their core functionality, highlighting the complementary roles of LLMs (as semantic engines) and SciKGs (as knowledge infrastructures) in addressing scientific tasks.

Model	Domains	Roles of LLMs	Roles of SciKG	Task	Application
KnowNET [86] (2024)	Drug	Semantic Interface (Query Generation)	Grounding (Factual Verification)	M	Guide health information seeking
FactFinder [87] (2024)	Drug	Semantic Interface (Query Generation)	Grounding (Factual Retrieval)	M	Life-science question answering
DDI-GPT [32] (2024)	Drug	Reasoner (Prediction & Explanation)	Representation (Semantic Enhancement)	C	Explainable prediction of drug-drug interactions
Soman et al. [84] (2024)	Drug, Omics	Constructor, Interface (KG Construction, Text Generation)	Grounding (Knowledge Base & Traceability)	M, C	Drug repurposing and medical QA
BioLORD [99] (2024)	Drug, Omics	Reasoner (Semantic Representation Optimization)	Grounding (Knowledge Base & Semantic Support)	M	Enhance biomedical semantic similarity
HeCiX [92] (2024)	Drug, Omics	Semantic Interface (Format Conversion)	Grounding (Knowledge Base)	M	Enhance clinical trial research
KRAGEN [95] (2024)	Drug, Omics	Orchestrator (Plan Generation & Execution)	Grounding (Knowledge Base & Visualization)	M	Visualized biomedical QA system
MechGPT [96] (2024)	Material	Constructor, Reasoner, Orchestrator (KG Construction, Explanation, Multi-agent)	Grounding, Reasoning Constraints (Knowledge & Explainability)	C, S, I	Materials analysis and design
SciAgents [31] (2024)	Material	Constructor, Reasoner, Generator (KG Construction, Analytical Reasoning, Hypothesis Generation)	Grounding (Knowledge Base)	M, I	Automated discovery in biomaterials science
MKG [51] (2024)	Material	Constructor (KG Construction & Maintenance)	Grounding (Knowledge Base)	I	Multidisciplinary materials science discovery
OpenTCM [100] (2025)	Drug	Interface, Reasoner, Constructor (Retrieval, Diagnosis, KG Construction)	Reasoning Constraints (Knowledge Retrieval Enhancement)	M	Traditional Chinese Medicine diagnosis
iKraph [9] (2025)	Drug	Constructor (KG Construction)	Grounding (Knowledge Base)	S	Biomedical Research
KGT [15] (2025)	Drug, Omics	Interface, Reasoner (Query Generation & Reasoning Output)	Grounding, Reasoning Constraints (Fact Checking & Path Constraint)	S, M	Drug repositioning, Framework for pan-cancer QA
ESCARGOT [94] (2025)	Drug, Omics	Generator, Orchestrator (Strategy & Code Generation)	Grounding (Knowledge Base)	S, I	Biomedical AI agent
Cat-KG [90] (2025)	Chemistry	Constructor, Reasoning, Interface (KG Construction, Path Reasoning & Explanation)	Grounding, Reasoning Constraints (Explainability & Path Constraint)	C, M	Relay catalysis pathway recommendation
Ma et al. [10] (2025)	Chemistry	Constructor, Generator (KG Construction & Path Recommendation)	Grounding (Structured Knowledge Management)	S	Automated Retrosynthesis Planning of Macromolecules
KG-FM [82] (2025)	Material	Constructor, Reasoner (Multimodal Extraction, QA & Reasoning)	Grounding (Knowledge Base & Visualization)	M	Improve LLM QA in framework materials
SciToolAgent [97] (2025)	Comprehensive	Orchestrator (Multi-agent Collaboration)	Grounding (Tool Knowledge Base)	S, M, I	Scientific agent for multi-tool integration

Abbr: M: Multi-source Data Interpretation; C: Complex System Mechanism Analysis; S: System Performance Optimization; I: Innovative Solution Design.

ical disparities (e.g., dense interaction networks vs. sparse property tables), complicating schema alignment. Practically, significant barriers arise from restrictive data governance. Many high-value datasets are locked behind proprietary paywalls or restrictive licenses, creating "data silos" that effectively block the construction of comprehensive public graphs. Even when data is accessible, inconsistent usage policies further complicate automated federation. This siloed development undermines the potential of SciKGs to support cross-disciplinary reasoning, for instance, linking protein interaction networks with materials-based drug delivery systems. Addressing these multifaceted challenges requires a synergistic approach: deploying technical solutions such as automated ontology alignment and federated query architectures to resolve heterogeneity, while simultaneously advocating for standardized open data licensing to break down proprietary silos. Such advances would enable truly cross-domain SciKGs that capture the interconnected nature of scientific discovery.

Dynamic and Temporal Knowledge. Scientific knowledge is inherently dynamic, with new discoveries, revised hypotheses, and retracted claims constantly reshaping the research landscape. Traditional KGs, however, are largely static, making them ill-suited to capture the temporal and evolving nature of science. This mismatch raises both technical and epistemic challenges: how can SciKGs be continuously updated without sacrificing reproducibility? How should they represent uncertainty, competing hypotheses, or retracted findings? Current frameworks lack mechanisms to encode confidence scores, conflicting evidence, or the provenance of retracted claims. Probabilistic knowledge graphs (modeling triples with associated probabilities) offer a promising direction to represent uncertainty, but their scalable integration and efficient reasoning remain open problems. Incremental learning and temporal graph modeling offer promising solutions, enabling knowledge graphs to evolve in tandem with scientific progress. At the same time, reproducibility concerns highlight the need for version-controlled and provenance-aware SciKGs, ensuring that dynamic updates remain transparent and traceable.

Trustworthy and Explainable Reasoning. As SciKGs are increasingly coupled with LLMs, questions of trust, transparency, and bias become paramount. Automated reasoning over incomplete or biased data risks producing

misleading conclusions, with potentially serious implications in sensitive domains such as drug development or clinical decision-making. The lack of formal uncertainty representation (e.g., confidence scores for triples) and mechanisms to handle retractions or conflicting findings undermines the trustworthiness of KG-driven inferences. Moreover, the opaque nature of many AI models undermines interpretability and hinders adoption by domain experts. Building trustworthy SciKGs requires mechanisms for explainable reasoning, bias detection and mitigation, and transparent provenance tracking. Ethical and societal considerations must also be addressed, including issues of data privacy, intellectual property, and equitable access to knowledge infrastructures. Establishing trustworthiness is not merely a technical challenge but a prerequisite for integrating SciKGs into the scientific process.

6.3 Opportunities

Building Standards, Benchmarks, and Validation Pipelines. To address data quality at scale, there is a pressing need for community-driven standards, benchmarking frameworks, and automated validation pipelines. Establishing minimal information standards across domains ensures consistent and transparent reporting. Standardized ontologies and interoperable data formats enable harmonization across repositories, while benchmark suites can evaluate SciKG performance in capturing domain-specific knowledge, such as chemical reaction mechanisms in chemistry or gene regulatory networks in biology. Furthermore, automated curation tools powered by natural language processing and machine learning can detect anomalies, impute missing values, and flag potential biases. Together, these measures create a feedback loop of continuous quality assessment and improvement, transforming SciKGs into auditable and trustworthy knowledge infrastructures.

Deeper Integration with Multimodal Foundation Models. To bridge disciplinary divides, SciKGs must evolve into multimodal, semantically unified knowledge backbones that integrate diverse data types and modalities. Foundational multimodal LLMs can act as powerful intermediaries for cross-modal alignment and semantic translation. For example, LLMs can extract and normalize entity mentions from scientific texts across domains, while molecular encoders standardize chemical representations. When

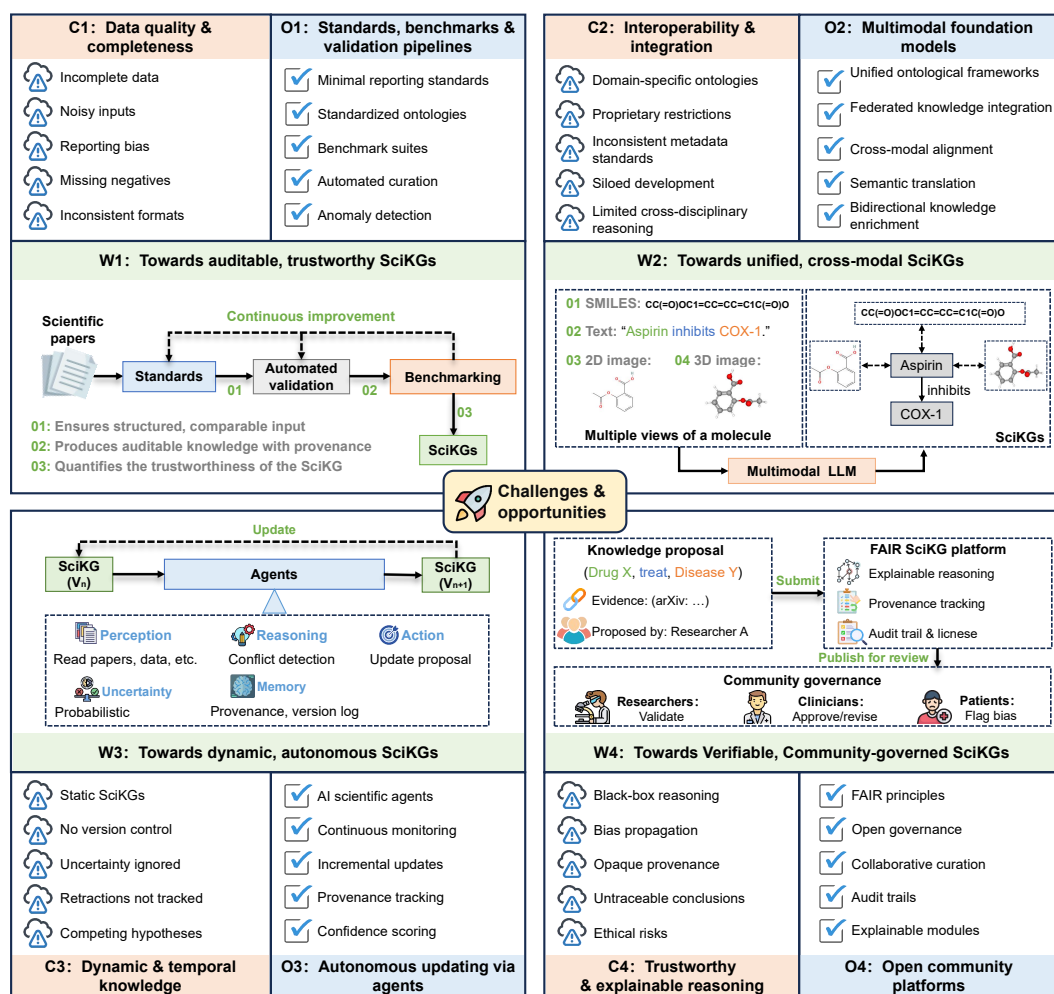


Figure 9. Challenges and Opportunities of SciKGs. This figure illustrates the major challenges (C1-C4) facing SciKGs, including data quality and completeness, interoperability and integration, dynamic and temporal knowledge, and trustworthy and explainable reasoning. Each challenge is paired with corresponding opportunities (O1-O4) for advancement, such as building standards and benchmarks, integrating multimodal foundation models, autonomous updating via agents, and developing community-driven platforms. The green sections depict workflows (W1-W4) that enable these opportunities, highlighting a path towards more auditable, unified, dynamic, and community-governed SciKGs.

grounded in a shared SciKG schema, these models enable knowledge fusion across text, tables, images, and structured databases. This bidirectional integration allows foundation models to enrich SciKGs with newly mined knowledge, while SciKGs provide symbolic, interpretable constraints that improve the factual accuracy and reasoning fidelity of generative models. The result is a synergistic architecture that supports truly interdisciplinary knowledge synthesis.

Autonomous Updating and Correcting Knowledge Graphs via LLM Agents. To keep pace with the evolving nature of science, SciKGs must transition from static repositories to adaptive, self-updating systems. Autonomous scientific agents capable of reading literature, analyzing data, generating hypotheses, and even design-

ing experiments can serve as intelligent curators that continuously monitor, update, and validate knowledge graphs. These agents can perform incremental updates, flag anomalies, resolve contradictions using probabilistic reasoning, and maintain versioned histories of assertions with full provenance. They can also attach confidence scores to triples based on evidence strength and implement procedures for deprecating knowledge linked to retracted publications. For instance, in genomics, an agent could detect conflicting annotations about a gene's function, assess the credibility of sources, and dynamically update the graph with confidence scores. Similarly, in materials science, agents could ingest newly published alloy properties and suggest plausible performance predictions. By embedding temporal logic and uncertainty model-

ing, such agent systems transform SciKGs into evolving knowledge ecosystems.

Developing Open SciKG Platforms. To establish trust, SciKGs must be developed and governed through open, inclusive, and community-led platforms grounded in the FAIR principles: Findability, Accessibility, Interoperability, and Reusability. Such platforms empower diverse stakeholders to collaboratively build, validate, and govern knowledge graphs. Crucially, to dismantle the practical barriers of proprietary restrictions (as highlighted in the Challenges section), the community must adopt transparent open data licenses (e.g., CC-BY) and develop sustainable business models for maintaining public knowledge infrastructures. Transparent provenance tracking, open licensing, and audit trails ensure accountability, while modular, explainable reasoning modules allow users to trace how conclusions are derived. For example, global biomedical consortia could co-develop a shared SciKG integrating clinical trial data, omics profiles, and real-world patient outcomes, enabling transparent, reproducible translational research. By democratizing access and participation, these platforms not only enhance trustworthiness but also foster equitable innovation across regions and disciplines.

6.4 Future Directions

SciKG Self-Evolving Framework. Future SciKGs should be designed as a self-evolving framework capable of autonomously ingesting new knowledge, detecting inconsistencies, and refining existing entities, relations, and attributes. Realizing such self-evolution can be conceptualized as a multi-agent system, where specialized agents handle complementary tasks: one agent continuously mines and extracts new knowledge from publications, preprints, or experimental logs; another agent performs consistency checking, conflict resolution, and uncertainty quantification; yet another updates embeddings and temporal representations while maintaining provenance and version control. These agents communicate and coordinate to ensure that the knowledge graph evolves in a coherent and reproducible manner. Methodologically, incremental learning, temporal graph modeling, and probabilistic graph models underpin agent operations, while automated pipelines powered by LLMs enable the ingestion of unstructured text and multimodal data. For example, in genomics, an extraction agent could identify new functional an-

notations for genes, a validation agent reconciles conflicts with existing evidence, and an update agent adjusts confidence scores for prior assertions. By framing self-evolution as a coordinated multi-agent system, SciKGs achieve adaptive knowledge management, supporting longitudinal studies and scalable, automated curation across scientific domains.

SciKG-LLM Co-Evolution System. A co-evolutionary framework between SciKGs and large language models (LLMs) envisions a tightly coupled system in which structured and unstructured knowledge continuously inform and refine one another through an iterative, bidirectional pipeline. In this paradigm, LLMs equipped with domain-specific prompting, retrieval-augmented generation, and self-verification modules autonomously extract new entities, relations, and hypotheses from scientific literature, experimental logs, and multimodal datasets. The extracted triples are then verified by a knowledge validation agent that applies probabilistic reasoning and schema alignment to ensure consistency and novelty before being merged into the SciKG via incremental updates with full provenance tracking. Conversely, SciKGs serve as interpretable priors that ground and constrain LLM inference, reducing hallucinations and enhancing domain fidelity through techniques such as graph-based retrieval augmentation, neural-symbolic reasoning, and contrastive knowledge alignment that integrate KG embeddings directly into the LLM's representation space. Over time, co-adaptive feedback mechanisms allow both components to improve jointly: the evolving SciKG provides structured supervision for continual fine-tuning or reinforcement learning of the LLM, while the LLM reorganizes and corrects graph regions showing inconsistency or conceptual drift. This closed feedback loop enables SciKGs to grow richer and more precise while LLMs become more grounded and interpretable, forming a foundation for more reliable and explainable scientific reasoning.

SciKG-Driven AI Scientist Agents. The ultimate vision is to embed SciKGs within autonomous AI scientist agents that operate in a closed-loop “perception–cognition–execution–feedback” cycle. In this paradigm, agents perceive experimental or computational data, encode it into the SciKG, cognitively reason over the integrated knowledge (using both symbolic reasoning and generative LLM inference), and plan subsequent

actions, including designing new experiments or simulations. Key components include reinforcement learning for action selection, thinking and reasoning frameworks to handle uncertainty and conflicting evidence, and automated experiment execution interfaces (e.g., robotic lab platforms). For instance, in materials discovery, the agent could propose a new alloy composition, simulate its thermodynamic stability, update the SciKG with predicted properties, and decide the next set of experiments based on expected information gain. The closed-loop integration of real-time data ingestion, knowledge graph updates, and adaptive action planning enables a continuously learning system, where the SciKG serves not only as a repository of knowledge but as a dynamic decision-making substrate that informs, constrains, and amplifies scientific exploration.

In summary, these directions envision SciKGs not merely as static repositories but as the dynamic, reasoning core of future scientific ecosystems. The progression from self-evolving frameworks to co-evolution with LLMs, and ultimately to embodiment within AI scientist agents, charts a course toward autonomous discovery systems. By pursuing this roadmap, we can transform SciKGs from passive knowledge bases into active partners in the scientific process, capable of guiding, accelerating, and ultimately redefining the very frontiers of scientific exploration.

FUNDING

This work is funded by the New Generation Artificial Intelligence - National Science and Technology Major Project (2025ZD0122801, H.C.), the Zhejiang Provincial “Jianbing” “Lingyan” Research and Development Program of China (2025C01097, K.D. and Q.Z.), the National Natural Science Foundation of China (NSFC62301480: K.D.; NSFC62302433: Q.Z.; NSFCU23A20496: Q.Z.; NSFCU23B2055: H.C.), and the Fundamental Research Funds for the Central Universities (226-2023-00138, H.C.).

AUTHOR CONTRIBUTIONS

K.D. and Z.Z.: Conceptualization, Investigation, Visualization, Writing – original draft. Y.T.: Investigation, Writing – original draft. K.F. and X.Z.: Writing – original draft (Sections on SciKG applications). H.W., Y.Y., H.D., Z.N., S.W., X.F., H.X., and L.B.: Writing – review & editing. Q.L., H.W., Q.Z., and H.C.: Supervision, Writing – review & editing, Funding ac-

quisition. All authors discussed the framework and approved the final manuscript.

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

1. Shiflet AB and Shiflet GW. *Introduction to Computational Science: Modeling and Simulation for the Sciences*. Princeton University Press: Princeton, 2014.
2. Deshpande D, Chhugani K, Ramesh T *et al*. The evolution of computational research in a data-centric world. *Cell* 2024; **187**: 4449–57.
3. Fillinger S, de la Garza L, Peltzer A *et al*. Challenges of big data integration in the life sciences. *Anal Bioanal Chem* 2019; **411**: 6791–800.
4. Himmelstein DS, Lizée A, Hessler C *et al*. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* 2017; **6**: e26726.
5. Fang Y, Zhang Q, Zhang N *et al*. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nat Mach Intell* 2023; **5**: 542–53.
6. MacLean F. Knowledge graphs and their applications in drug discovery. *Expert Opin Drug Discov* 2021; **16**: 1057–69.
7. Zheng X, Wang B, Zhao Y *et al*. A knowledge graph method for hazardous chemical management: Ontology design and entity identification. *Neurocomputing* 2021; **430**: 104–11.
8. Jeong J, Lee N, Shin Y *et al*. Intelligent generation of optimal synthetic pathways based on knowledge graph inference and retrosynthetic predictions using reaction big data. *J Taiwan Inst Chem Eng* 2022; **130**: 103982.
9. Zhang Y, Sui X, Pan F *et al*. A comprehensive large-scale biomedical knowledge graph for ai-powered data-driven biomedical research. *Nat Mach Intell* 2025; 1–13.
10. Ma Q, Zhou Y and Li J. Automated retrosynthesis planning of macromolecules using large language models and knowledge graphs. *Macromol Rapid Commun* 2025; 2500065.
11. Chen J, Dong H, Hastings J *et al*. Knowledge graphs for the life sciences: Recent developments, challenges and opportunities [preprint]. arXiv: 2309.17255.
12. Pan S, Luo L, Wang Y *et al*. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans Knowl Data Eng* 2024; **36**: 3580–99.
13. Yasunaga M, Bosselut A, Ren H *et al*. Deep bidirectional language-knowledge graph pretraining. *Adv Neural Inf Process Syst* 2022; **35**: 37309–23.
14. Kim J, Kwon Y, Jo Y *et al*. KG-GPT: A general framework for reasoning on knowledge graphs using large language models [preprint]. arXiv: 2310.11220.
15. Feng Y, Zhou L, Ma C *et al*. Knowledge graph-based thought: A knowledge graph-enhanced llm framework for pan-cancer question answering. *GigaScience* 2025; **14**: giae082.
16. Wu J, Zhu J, Qi Y *et al*. Medical Graph RAG: Towards safe medical large language model via graph retrieval-augmented generation [preprint]. arXiv: 2408.04187.

17. Ji S, Pan S, Cambria E *et al.* A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans Neural Netw Learn Syst* 2021; **33**: 494–514.
18. Peng C, Xia F, Naseriparsa M *et al.* Knowledge graphs: Opportunities and challenges. *Artif Intell Rev* 2023; **56**: 13071–102.
19. Zhong L, Wu J, Li Q *et al.* A comprehensive survey on automatic knowledge graph construction. *ACM Comput Surv* 2023; **56**: 1–62.
20. Wang C, Yang Y, Song J *et al.* Research progresses and applications of knowledge graph embedding technique in chemistry. *J Chem Inf Model* 2024; **64**: 7189–213.
21. Cao J, Fang J, Meng Z *et al.* Knowledge graph embedding: A survey from the perspective of representation spaces. *ACM Comput Surv* 2024; **56**: 1–42.
22. Kulkarni A, Alotaibi F, Zeng X *et al.* Scientific hypothesis generation and validation: Methods, datasets, and future directions [preprint]. arXiv: 2505.04651.
23. Wehner C, Iliopoulou C and Besold TR. From latent to lucid: Transforming knowledge graph embeddings into interpretable structures [preprint]. arXiv: 2406.01759.
24. Benson DA, Cavanaugh M, Clark K *et al.* Genbank. *Nucleic Acids Res* 2012; **41**: D36–D42.
25. Berman HM, Westbrook J, Feng Z *et al.* The protein data bank. *Nucleic Acids Res* 2000; **28**: 235–42.
26. Belleau F, Nolin MA, Tourigny N *et al.* Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008; **41**: 706–16.
27. Smith B, Ashburner M, Rosse C *et al.* The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007; **25**: 1251–55.
28. Bordes A, Usunier N, Garcia-Duran A *et al.* Translating embeddings for modeling multi-relational data. *Adv Neural Inf Process Syst* 2013; **26**.
29. Hamilton W, Ying Z and Leskovec J. Inductive representation learning on large graphs. *Adv Neural Inf Process Syst* 2017; **30**.
30. Chen B and Bertozzi AL. AutoKG: Efficient automated knowledge graph generation for language models. In: 2023 IEEE International Conference on Big Data 2023: 3117–3126. IEEE.
31. Ghafarollahi A and Buehler MJ. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Adv Mater* 2024; 2413523.
32. Xu C, Bulusu KG, Pan H *et al.* DDI-GPT: Explainable prediction of drug-drug interactions using large language models enhanced with knowledge graphs [preprint]. bioRxiv.
33. Kim S, Chen J, Cheng T *et al.* Pubchem 2023 update. *Nucleic Acids Res* 2023; **51**: D1373–D80.
34. UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res* 2019; **47**: D506–D15.
35. Jain A, Ong SP, Hautier G *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater* 2013; **1**.
36. Hirschman L, Yeh A, Blaschke C *et al.* Overview of biocreative: Critical assessment of information extraction for biology. *BMC Bioinformatics* 2005; **6**: S1.
37. Degtyarenko K, De Matos P, Ennis M *et al.* ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2007; **36**: D344–D50.
38. Fabian B, Edlich T, Gaspar H *et al.* Molecular representation learning with language models and domain-relevant auxiliary tasks [preprint]. arXiv: 2011.13230.
39. Sung M, Jeong M, Choi Y *et al.* BERN2: An advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* 2022; **38**: 4837–39.
40. Beltagy I, Lo K and Cohan A. SciBERT: A pre-trained language model for scientific text [preprint]. arXiv: 1903.10676.
41. Shamsabadi M, D'Souza J and Auer S. Large language models for scientific information extraction: An empirical study for virology [preprint]. arXiv:2401.10040.
42. Choi S and Jung Y. Knowledge graph construction: Extraction, learning, and evaluation. *Appl Sci* 2025; **15**: 3727.
43. Silva MC, Faria D and Pesquita C. Matching multiple ontologies to build a knowledge graph for personalized medicine. In: *The Semantic Web. ESWC 2022*; **13261**: 461–77.
44. Kokash N, Wang L, Gillespie TH *et al.* Ontology- and llm-based data harmonization for federated learning in healthcare [preprint]. arXiv: 2505.20020.
45. Osman I, Pileggi SF and Yahia SB. Uncertainty in automated ontology matching: Lessons from an empirical evaluation. *Applied Sciences* 2024; **14**: 4679.
46. Chen Z, Zhang Y, Fang Y *et al.* Knowledge graphs meet multi-modal learning: A comprehensive survey [preprint]. arXiv: 2402.05391.
47. Tian Z, Zhang D and Dai HN. Continual learning on graphs: A survey [preprint]. arXiv: 2402.06330.
48. Zhao X, Blum M, Yang R *et al.* Agentigraph: An interactive knowledge graph platform for llm-based chatbots utilizing private data [preprint]. arXiv: 2410.11531.
49. Huang K, Chandak P, Wang Q *et al.* A foundation model for clinician-centered drug repurposing. *Nat Med* 2024; **30**: 3601–13.
50. Bang D, Lim S, Lee S *et al.* Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. *Nat Commun* 2023; **14**: 3570.
51. Wu D, Sun W, He Y *et al.* MKG-FENN: A multimodal knowledge graph fused end-to-end neural network for accurate drug-drug interaction prediction. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 2024; **38**: 10216–24.
52. Luo Y, Zhao X, Zhou J *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun* 2017; **8**: 573.
53. Hoang TL, Sbodio ML, Galindo MM *et al.* Knowledge enhanced representation learning for drug discovery. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 2024; **38**: 10544–10552.
54. Shang J, Xiao C, Ma T *et al.* Gamenet: Graph augmented memory networks for recommending medication combination. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 2019; **33**: 1126–33.
55. Evangelista JE, Clarke DJB, Xie Z *et al.* Toxicology knowledge graph for structural birth defects. *Commun Med* 2023; **3**: 98.

56. Mulero-Hernández J, Mironov V, Miñarro-Giménez JA *et al.* Integration of chromosome locations and functional aspects of enhancers and topologically associating domains in knowledge graphs enables versatile queries about gene regulation. *Nucleic Acids Res* 2024; **52**: e69.
57. Feng F, Tang F, Gao Y *et al.* GenomicKB: A knowledge graph for the human genome. *Nucleic Acids Res* 2023; **51**: D950–D56.
58. Zaripova K, Özsoy E, Navab N *et al.* PhenoKG: Knowledge graph-driven gene discovery and patient insights from phenotypes alone [preprint]. arXiv: 2506.13119.
59. Shao X, Li C, Yang H *et al.* Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with SpaTalk. *Nat Commun* 2022; **13**: 4429.
60. Cavalleri E, Cabri A, Soto-Gomez M *et al.* An ontology-based knowledge graph for representing interactions involving RNA molecules. *Sci Data* 2024; **11**: 906.
61. Binder J, Ursu O, Bologa C *et al.* Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. *Commun Biol* 2022; **5**: 125.
62. Santos A, Colaço AR, Nielsen AB *et al.* A knowledge graph to interpret clinical proteomics data. *Nat Biotechnol* 2022; **40**: 692–702.
63. Delmas M, Filangi O, Paulhe N *et al.* Building a knowledge graph from public databases and scientific literature to extract associations between chemicals and diseases. *Bioinformatics* 2021; **37**: 3896–904.
64. Sun H, Song Z, Chen Q *et al.* MMiKG: A knowledge graph-based platform for path mining of microbiota–mental diseases interactions. *Brief Bioinform* 2023; **24**: bbad340.
65. Liu T, Pan X, Wang X *et al.* Exploring the microbiota–gut–brain axis for mental disorders with knowledge graphs. *J Artif Intell Med Sci* 2021; **1**: 30–42.
66. Jha A, Khan Y, Sahay R *et al.* Metastatic site prediction in breast cancer using omics knowledge graph and pattern mining with Kirchhoff's law traversal [preprint]. bioRxiv.
67. McDermott MJ, Dwaraknath SS and Persson KA. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nat Commun* 2021; **12**: 3097.
68. Xie J, Wang Y, Rao J *et al.* Self-supervised contrastive molecular representation learning with a chemical synthesis knowledge graph. *J Chem Inf Model* 2024; **64**: 1945–54.
69. Zhang Z, Ma S, Zheng S *et al.* Semantic knowledge graph as a companion for catalyst recommendation. *Natl Sci Open* 2024; **3**: 20230040.
70. Li B and Chen H. Prediction of compound synthesis accessibility based on reaction knowledge graph. *Molecules* 2022; **27**: 1039.
71. Jiang P, Xiao C, Fu T *et al.* Bi-level contrastive learning for knowledge-enhanced molecule representations. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 2025; **39**: 352–60.
72. Fang Y, Zhang Q, Yang H *et al.* Molecular contrastive learning with chemical element knowledge graph. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 2022; **36**: 3968–76.
73. Ye Y, Ren J, Wang S *et al.* Construction and application of materials knowledge graph in multidisciplinary materials science via large language model. *Adv Neural Inf Process Syst* 2024; **37**: 56878–97.
74. Venugopal V and Olivetti E. MatKG: An autonomously generated knowledge graph in material science. *Sci Data* 2024; **11**: 217.
75. Nie Z, Zheng S, Liu Y *et al.* Automating materials exploration with a semantic knowledge graph for Li-ion battery cathodes. *Adv Funct Mater* 2022; **32**: 2201437.
76. Gao Y, Wang L, Chen X *et al.* Revisiting electrocatalyst design by a knowledge graph of Cu-based catalysts for CO₂ reduction. *ACS Catal* 2023; **13**: 8525–34.
77. McCusker JP, Keshan N, Rashid S *et al.* Nanomine: A knowledge graph for nanocomposite materials science. In: *The Semantic Web – ISWC* 2020; **12507**: 144–59.
78. Mrdjenovich D, Horton MK, Montoya JH *et al.* PropNet: A knowledge graph for materials science. *Matter* 2020; **2**: 464–80.
79. Huang C, Chen C, Shi L *et al.* Material property prediction with element attribute knowledge graphs and multimodal representation learning [preprint]. arXiv: 2411.08414.
80. Anand A, Kumari P and Kalyani AK. High throughput screening of new piezoelectric materials using graph machine learning and knowledge graph approach. *Comput Mater Sci* 2025; **246**: 113445.
81. Guo P, Meng W and Bao Y. Knowledge graph-guided data-driven design of ultra-high-performance concrete (UHPC) with interpretability and physicochemical reaction discovery capability. *Constr Build Mater* 2024; **430**: 136502.
82. Bai X, He S, Li Y *et al.* Construction of a knowledge graph for framework material enabled by large language models and its application. *npj Comput Mater* 2025; **11**: 51.
83. Zhang Y, Chen F, Liu Z *et al.* A materials terminology knowledge graph automatically constructed from text corpus. *Sci Data* 2024; **11**: 600.
84. Soman K, Rose PW, Morris JH *et al.* Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics* 2024; **40**: btae560.
85. Sriramanan G, Bharti S, Sadasivan VS *et al.* LLM-Check: Investigating detection of hallucinations in large language models. *Adv Neural Inf Process Syst* 2024; **37**: 34188–216.
86. Yan Y, Hou Y, Xiao Y *et al.* Knownet: Guided health information seeking from LLMs via knowledge graph integration. *IEEE Trans Vis Comput Graph* 2024; .
87. Steinigen D, Teucher R, Ruland TH *et al.* Fact finder – enhancing domain expertise of large language models by incorporating knowledge graphs [preprint]. arXiv: 2408.03010.
88. Qin G, Narsinh K, Wei Q *et al.* Generating biomedical knowledge graphs from knowledge bases, registries, and multiomic data [preprint]. bioRxiv.
89. Luo L, Zhao Z, Haffari G *et al.* Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models [preprint]. arXiv: 2410.13080.
90. Fu F, Li QQ, Wang F *et al.* Synergizing knowledge graph and large language model for relay catalysis pathway recommendation. *Natl Sci Rev* 2025; **12**: nwaf271.
91. Liang W, Meo PD, Tang Y *et al.* A survey of multi-modal knowledge graphs: Technologies and trends. *ACM Comput Surv* 2024; **56**: 1–41.

92. Kulkarni PS, Jain M, Sheshanarayana D *et al.* HeCiX: Integrating knowledge graphs and large language models for biomedical research [preprint]. arXiv: 2407.14030.
93. Hatem S, Khoriba G, Gad-Elrab MH *et al.* Up to date: Automatic updating knowledge graphs using LLMs. *Procedia Comput Sci* 2024; **244**: 327–34.
94. Matsumoto N, Choi H, Moran J *et al.* ESCARGOT: An AI agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for enhanced reasoning. *Bioinformatics* 2025; **41**: btaf031.
95. Matsumoto N, Moran J, Choi H *et al.* KRAGEN: A knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. *Bioinformatics* 2024; **40**: btae353.
96. Buehler MJ. Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design. *ACS Eng Au* 2024; **4**: 241–77.
97. Ding K, Yu J, Huang J *et al.* SciToolAgent: A knowledge graph-driven scientific agent for multi-tool integration. *Nat Comput Sci* 2025; .
98. Song T, Luo M, Zhang X *et al.* A multiagent-driven robotic AI chemist enabling autonomous chemical research on demand. *J Am Chem Soc* 2025; **147**: 12534–45.
99. Remy F, Demuynck K and Demeester T. BioLORD-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights. *J Am Med Inform Assoc* 2024; **31**: 1844–55.
100. He J, Guo Y, Lam LK *et al.* OpenTCM: A GraphRAG-empowered LLM-based system for traditional chinese medicine knowledge retrieval and diagnosis [preprint]. arXiv: 2504.20118.