

Keyan Ding¹, Zhihui Zhu¹, Yuqi Tang¹, Kehua Feng¹, Xiang Zhuang^{1,2}, Hongwei Wang¹,
Yi Yang¹, Huifang Du³, Zhangkai Ni³, Shiqi Wang⁴, Xiaohui Fan¹, Huabin Xing¹, Lei Bai²,
Qi Liu³, Haofen Wang³, Qiang Zhang¹, and Huajun Chen^{1,2}

¹Zhejiang University

²Shanghai Artificial Intelligence Laboratory

³Tongji University

⁴City University of Hong Kong

November 21, 2025

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

Keyan Ding^{1*}, Zhihui Zhu^{1*}, Yuqi Tang¹, Kehua Feng¹, Xiang Zhuang^{1,2}, Hongwei Wang¹, Yi Yang¹, Huifang Du³, Zhangkai Ni³, Shiqi Wang⁴, Xiaohui Fan¹, Huabin Xing¹, Lei Bai², Qi Liu^{3✉}, Haofen Wang^{3✉}, Qiang Zhang^{1✉}, and Huajun Chen^{1,2✉}

¹Zhejiang University, Hangzhou, China

²Shanghai Artificial Intelligence Laboratory, Shanghai, China

³Tongji University, Shanghai, China

⁴City University of Hong Kong, Hong Kong (SAR), China

*Share the first authorship

✉Corresponding authors: qiliu@tongji.edu.cn, haofen.wang@tongji.edu.cn, qiang.zhang.cs@zju.edu.cn, huajunsir@zju.edu.cn

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.

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^{1–5}. 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^{6–9}. 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^{10–16}. 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^{17–20}. 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^{21–26}, demonstrating their potential as engines of discovery.

However, constructing SciKGs remains technically demanding. Entity and relation extraction, 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^{8, 27–31}. The integration of artificial intelligence (AI) techniques, particularly large language models (LLMs)^{32–36}, has begun to transform this landscape. LLMs can automate knowledge extraction from unstructured literature, enrich semantic representations, and predict missing links within graphs^{37–42}. Conversely, SciKGs provide structured grounding for LLMs, improving factual reliability, contextual reasoning, and reducing hallucinations in generative scientific tasks^{43–47}. This bidirectional synergy between SciKGs and LLMs is opening new opportunities for

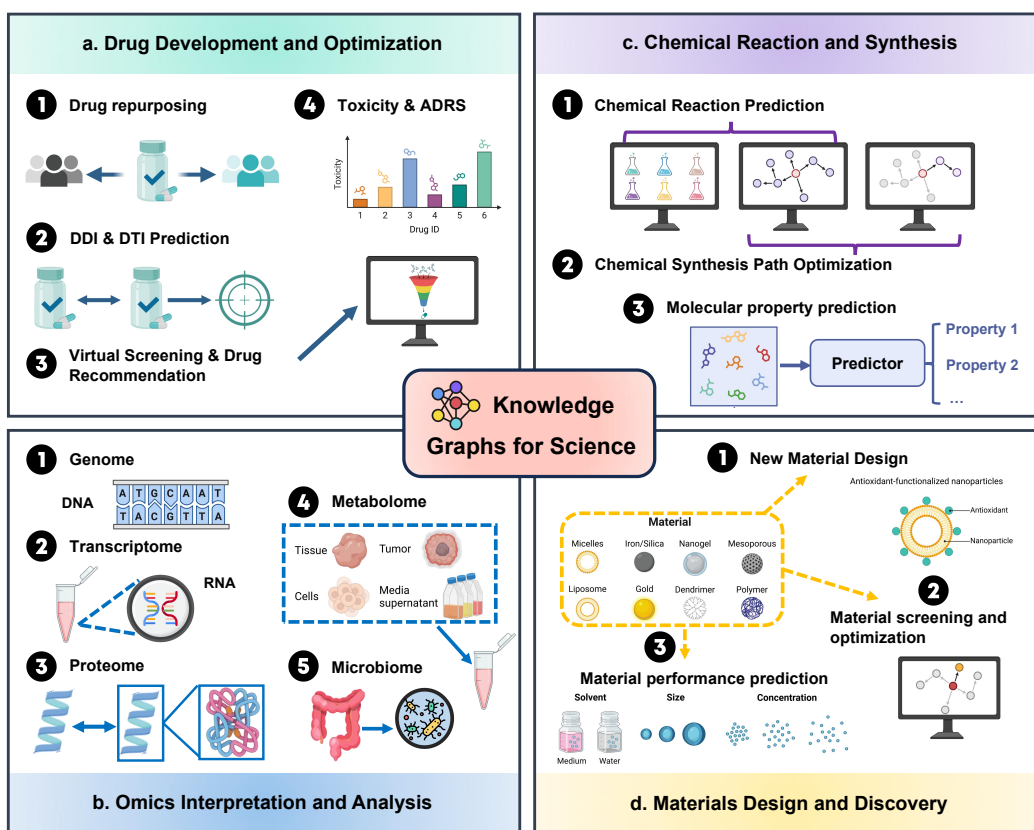


Figure 1. An overview of the research scope in this survey, covering four fundamental scientific tasks in biology, chemistry, and materials science: (a) drug development and optimization, (b) omics interpretation and analysis, (c) chemical reaction and synthesis, and (d) materials design and discovery.

AI-driven scientific reasoning, hypothesis generation, and decision support.

Despite growing interest, most surveys to date have concentrated on general-purpose KGs^{32,48–53}, providing valuable overviews of graph construction techniques and applications but offering limited insight into the unique demands of scientific domains. Existing reviews of SciKGs^{18,27,54–59} 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 (Figure 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 GitHub¹, 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.

The remainder of this review provides a roadmap for understanding, constructing, and leveraging SciKGs to accelerate scientific discovery (Figure 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

¹<https://github.com/hicai-zju/scikgs>

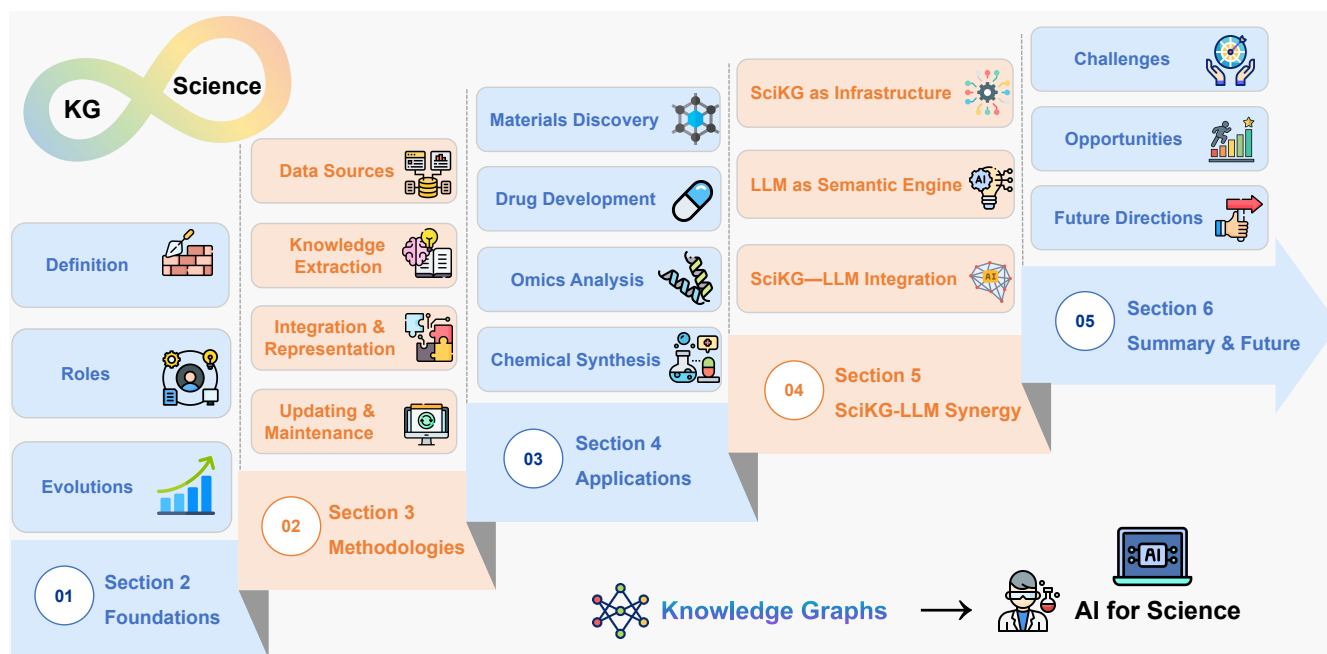


Figure 2. Structure of the survey. Our review is structured around the lifecycle of SciKGs: from their conceptual foundation and construction methodologies, to their applications and synergistic integration with LLMs for discovery, culminating in challenges, opportunities and future directions that envision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

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 scientific discovery.

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

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.

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 interoperability, 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^{60–64}, 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^{65–69} 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^{70–74}. 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.

Evolutions The development of SciKGs has undergone several transformative phases (Figure 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., GenBank⁷⁵, PDB⁷⁶). 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 Bio2RDF⁷⁷ and the Open Biological and Biomedical Ontology (OBO) Foundry⁷⁸ 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 networks⁷⁹, SciKGs evolved into predictive engines capable of inferring new links and patterns from existing knowledge. Representation learning (e.g., TransE⁸⁰, GraphSAGE⁸¹) 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., AutoKG⁸²), generate hypotheses grounded in SciKGs (e.g., SciAgents⁸³), and serve as natural-language interfaces for complex queries (e.g., DDI-GPT⁸⁴). 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 resulting graphs (Figure 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.

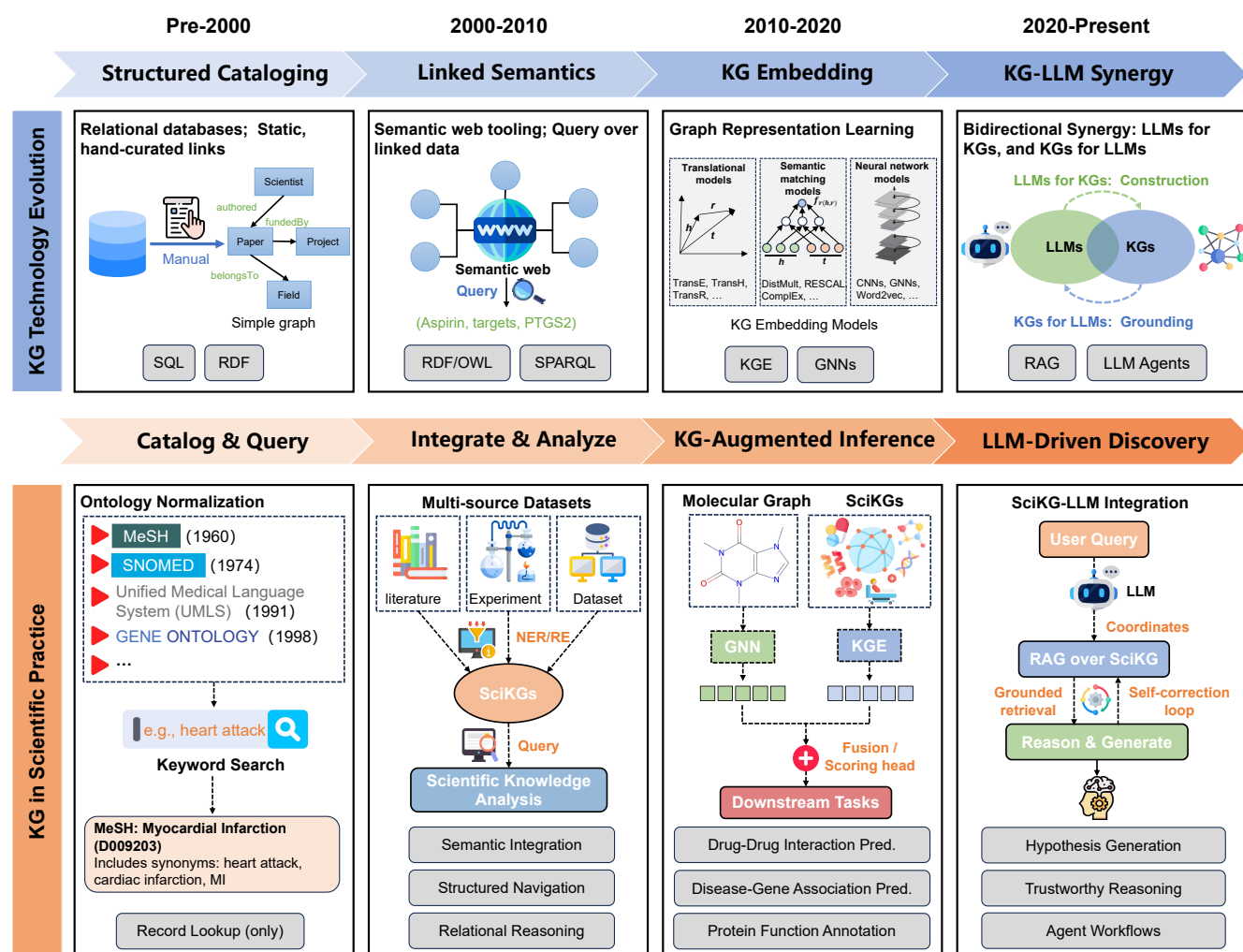


Figure 3. The co-evolution of knowledge graph technologies and their scientific practices. The technological evolution of KGs (top) has continually enabled new paradigms in SciKG applications (bottom). This progression has moved from static cataloging and manual integration to machine learning-driven 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.]

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 PubChem⁸⁵, UniProt⁸⁶, the Protein Data Bank (PDB)⁷⁶, and the Materials Project⁸⁷ 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,

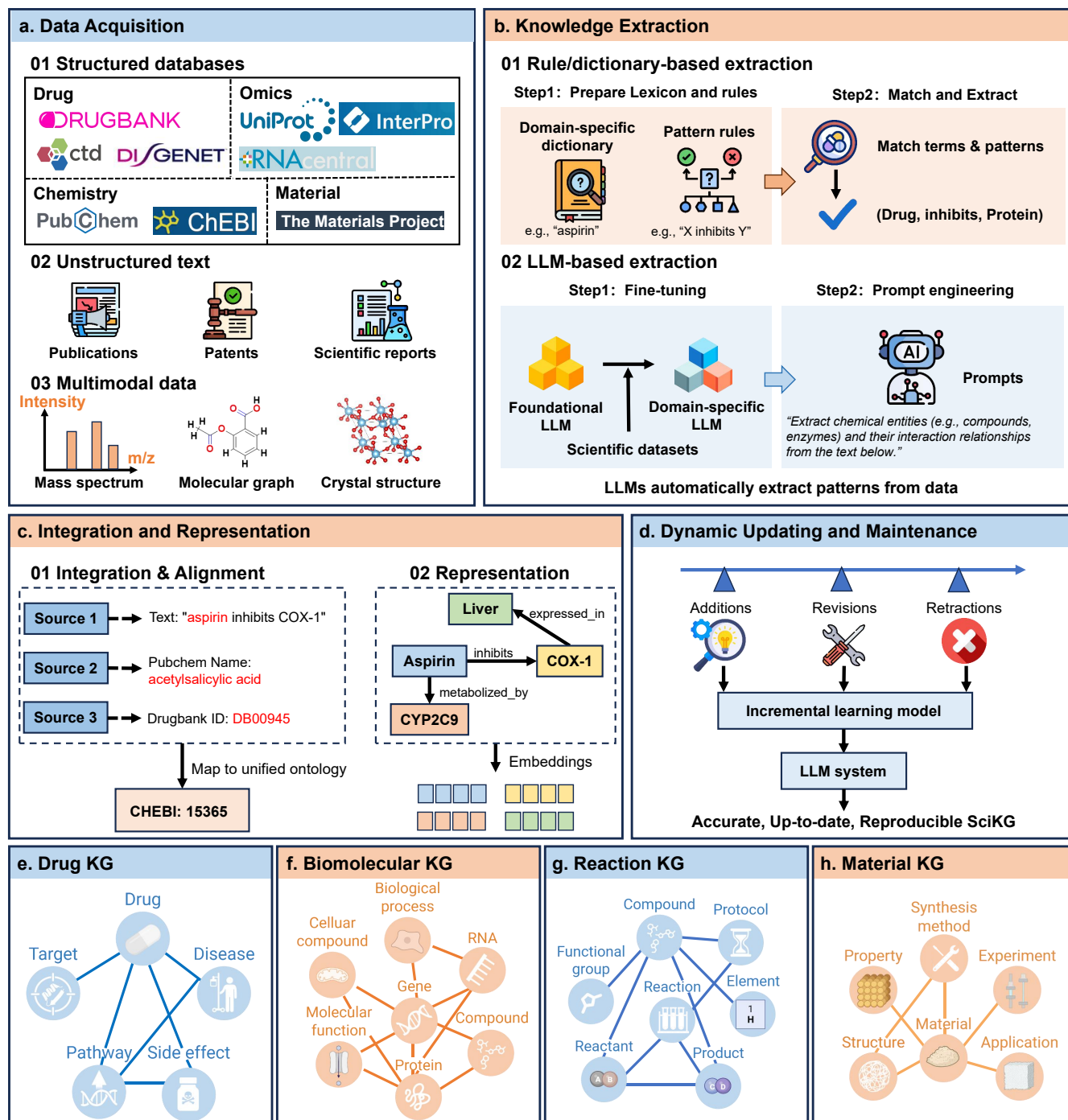


Figure 4. Construction and maintenance of SciKGs. (a) The foundation of SciKG construction involves integrating diverse data sources, including structured databases, unstructured text, and multimodal data. (b) Two main approaches for extracting entities and relations from the acquired data are illustrated: 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.

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 quantify 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 go beyond symbolic triples to embed numerical, visual, and temporal evidence directly into the graph structure.

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^{88–92} 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^{93–99} 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 SciBERT¹⁰⁰ 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^{101–107}. Through few-shot prompting and task-specific 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^{108,109} are emerging as practical solutions that combine the domain precision of ontology-based methods with the flexibility of neural models. 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^{110,111} are widely employed. These methods harmonize terminologies, reconcile conceptual discrepancies, and enable cross-domain reasoning. Recent frameworks increasingly adopt federated ontology mapping^{112–115} and probabilistic schema alignment^{116–119}, 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^{56,120–124} 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.

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^{103,104,125–127}. To capture these characteristics, methods such as cross-modal embedding^{128–131} 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^{132–136} 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^{137–141}. 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 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 (Figure 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^{24,142–144}. For instance, the TxGNN model¹⁴⁴, 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^{84,145–147}. As exemplified by the DDI-GPT framework⁸⁴, 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^{148–150}. The DTINet framework¹⁴⁸ 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^{121,151–153}. For example, frameworks like GAMENet¹⁵¹ integrate DDI knowledge graphs with patient records to recommend safe, personalized medication combinations. Meanwhile, in toxicity assessment, dedicated resources such as ReproTox-KG¹⁵² 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.

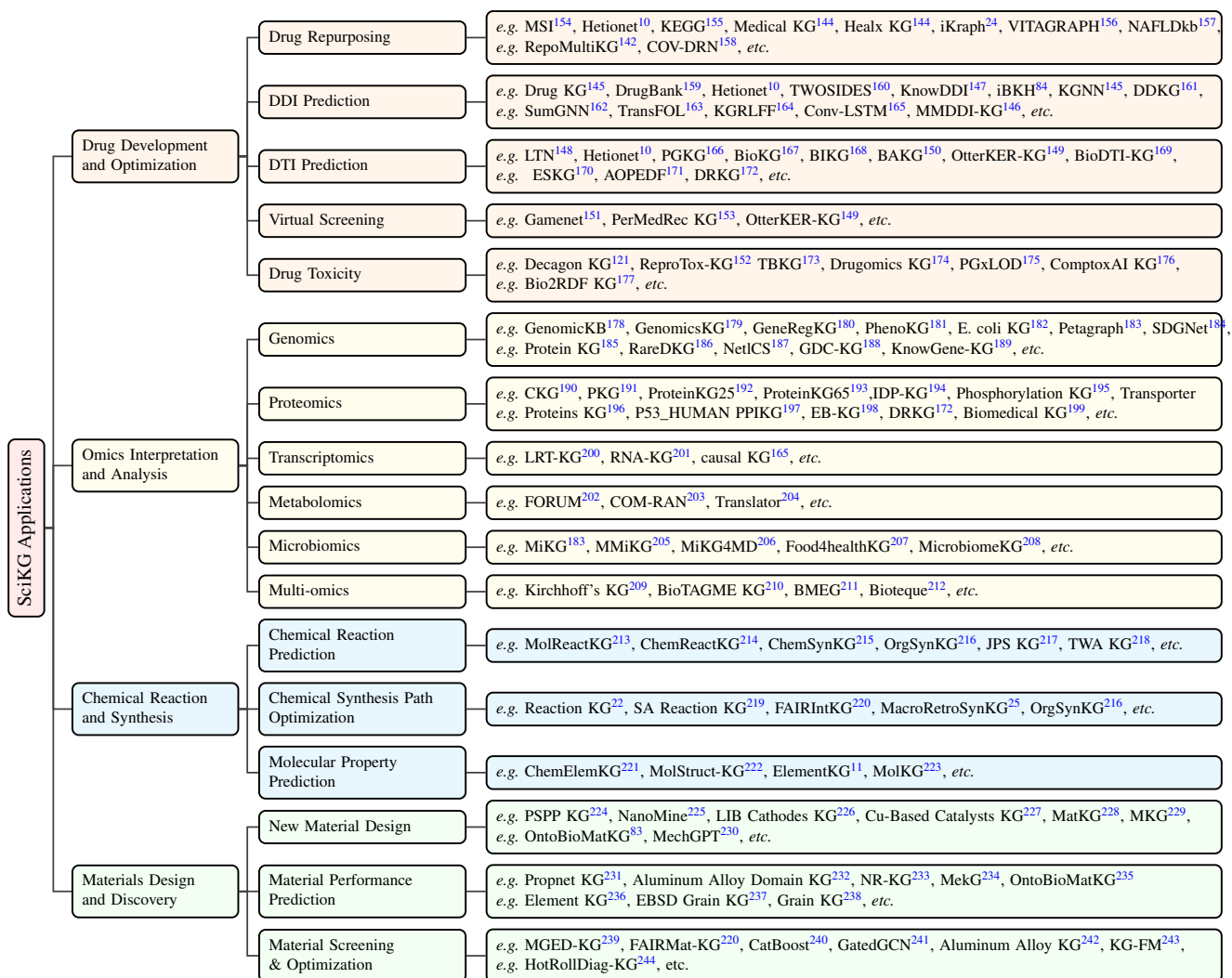


Figure 5. A taxonomy of existing SciKGs and their applications in science. More details of representative SciKGs are provided in Table S3, S4, S5, and S6.

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^{178,180,181}. This capability is demonstrated by PhenoKG¹⁸¹, 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^{200,201}. The RNA-KG resource²⁰¹ 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^{190–193,199}. The CKG¹⁹⁰ 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 KG²⁰² 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 MMiKG²⁰⁵ and MiKG²⁴⁵. 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 constraints²⁰⁹. 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^{66,213–215}. For instance, the ReaKE framework²¹⁵ constructs a chemical synthesis KG to enable contrastive learning that improves reaction classification and product prediction by capturing functional group transformations. 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^{22,25,219}. The work by Li et al.²¹⁹ demonstrates this by building a reaction network KG 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^{11,221–223,246}. The GODE framework²²³ 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^{97,247,248}. 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^{228,229}. 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^{225–227}, and supporting multi-agent reasoning frameworks for biomimetic materials⁸³. 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^{231,236}. In screening and optimization, KGs consolidate millions of entities across databases, literature, and experiments to prioritize candidates with desired characteristics, guide experimental design, and balance performance with production feasibility in piezoelectric materials, ultra-high-performance concrete, or COFs for gas storage^{239–241,243}. 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 Figure 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

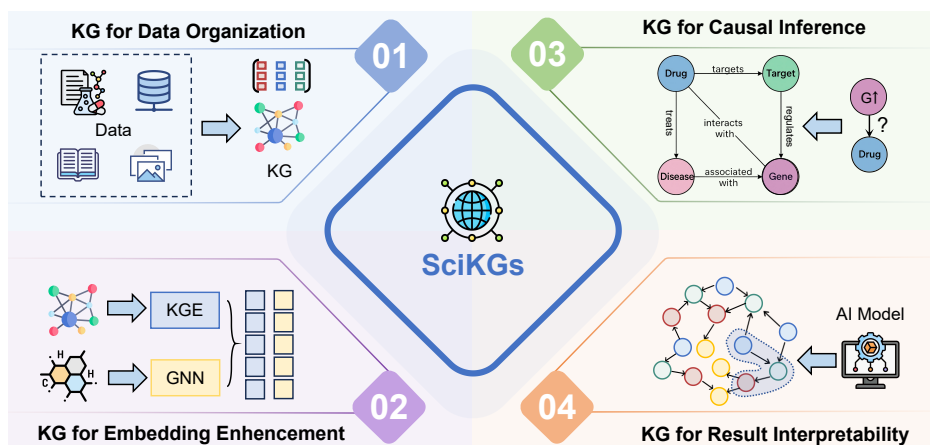


Figure 6. Summary of core functions of SciKGs in diverse scientific tasks. SciKGs serve as a foundational infrastructure that: (1) organizes heterogeneous scientific data into structured knowledge; (2) enhances representation learning via graph embedding; (3) enables causal and relational inference for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

transparency, grounding model outputs in explicit knowledge pathways and causal relationships, thus aligning data-driven predictions with scientific rationale.

Looking forward, the next evolution of SciKGs will hinge on cross-disciplinary integration, adaptive intelligence, and embodied reasoning. From a systems perspective, future SciKGs should transcend domain boundaries, linking molecular, biological, chemical, and material knowledge into interoperable meta-graphs that capture the full continuum from atomic interactions to macroscopic functions. This will enable knowledge transfer across fields, such as leveraging protein–ligand binding knowledge to guide catalyst design or applying material defect models to understand biological aging. 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 Scientific Knowledge Graphs and Large Language Models

The advancement of scientific discovery increasingly depends on combining knowledge bases with generative intelligence (e.g., LLMs)^{83, 146, 249–254}. KGs provide explicit representations of entities, relations, and domain knowledge, while LLMs offer powerful capabilities in reasoning, abstraction, and summary^{17, 18}. Their synergistic integration enables knowledge-grounded, interpretable, and adaptive solutions to complex scientific problems. In this section, we explore the integration of SciKGs and LLMs in scientific discovery, with a focus on dissecting their core complementary relationship of factual constraint and intelligent reasoning during scientific problem-solving (Figure 7).

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^{255–259}. 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 and verification. They serve as authoritative benchmarks against which LLM-generated hypotheses can be validated. For instance, the KNOWNET framework²⁶⁰ extracts triples from LLM outputs and maps them to validated evidence in external KGs, providing a visual interface to trace and verify claims. Similarly, FactFinder²⁶¹ augments LLMs with a medical KG through a structured retrieval-and-generation pipeline, 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^{204, 262}. Second, SciKGs define reasonable boundaries for LLM reasoning and prevent the generation of schemes that violate scientific principles. The Graph-Constrained Reasoning (GCR) framework²⁶³ integrates the KG structure directly into the LLM’s decoding process, ensuring that every step of the reasoning path is faithful to the graph and achieving zero reasoning hallucination on knowledge graph question-answering tasks. In chemistry, the Synergizing KG and LLM approach²⁶⁴ for relay catalysis uses a detailed catalysis KG (Cat-KG) to

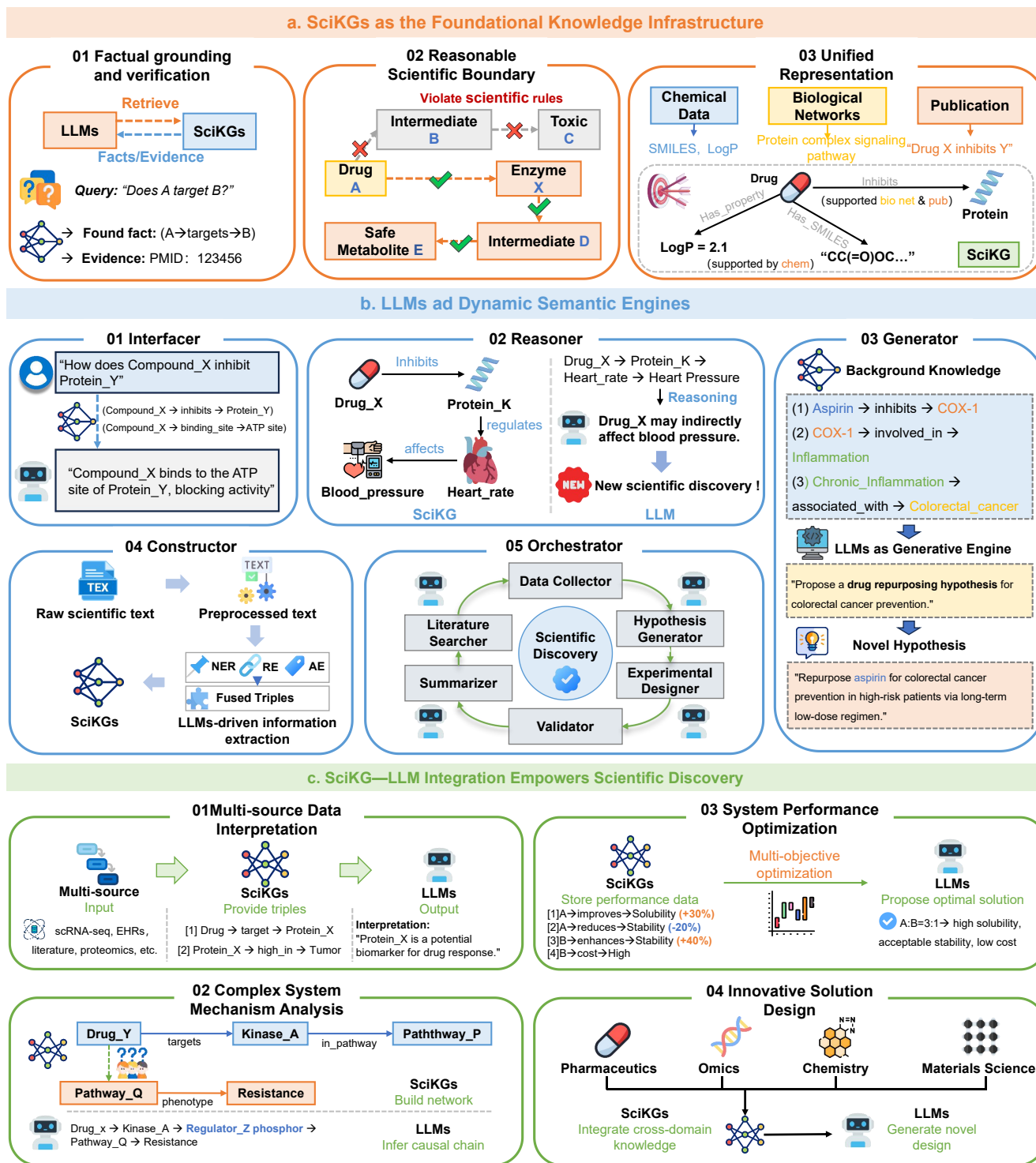


Figure 7. Synergistic integration of SciKGs and LLMs for knowledge-driven scientific discovery. (a) SciKGs serve as the foundational knowledge infrastructure by ensuring factual grounding and verification, defining reasonable scientific boundaries, and enabling unified representation of heterogeneous 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.

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. Systems like DDI-GPT⁸⁴ 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.

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^{51,265–269}. 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 HeCiX²⁷⁰ 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 framework⁸⁴ 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. SciAgents⁸³ 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.²⁵ 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 iKraph²⁴ 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 framework²⁴⁹ further demonstrates how LLMs can optimize knowledge extraction from biomedical KGs in a token-efficient manner, while systems like UpToDate²⁷¹ 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 ESCARGOT agent²⁷² 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^{146,272–274}. 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^{260,261,275–278}. 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^{84,249,264,279,280}. 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^{24,25}. 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. Such a paradigm naturally gives rise to the *AI Scientist Copilot*, a SciKG-empowered autonomous assistant 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 ontological alignment and entity linking to ensure factual consistency (e.g., resolving synonyms via PubChem IDs in chemistry). This is evidenced by tools like BioBERT³⁸ and SciBERT¹⁰⁰ for domain-specific text understanding. Furthermore, the paradigm of constructing task-specific knowledge substrates from multimodal sources is exemplified by works like MKG-FENN¹⁴⁶, which builds a multimodal knowledge graph by integrating drug-chemical, drug-substructure, drug-drug, and molecular structural knowledge to provide a comprehensive relational context for predicting drug-drug interactions.

(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. This process not only mitigates hallucination but also strengthens interpretability and scientific rigor. Works such as DDI-GPT⁸⁴ illustrate this for predicting and explaining drug-drug interactions, and Synergizing KG and LLM for Relay Catalysis²⁶⁴ demonstrates it for inferring complex reaction pathways.

(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 SciAgents⁸³, where a multi-agent framework reasons over a materials knowledge graph for bioinspired design; Automated Retrosynthesis Planning²⁵ that uses reaction KGs to constrain and validate LLM-generated synthesis pathways; and SciToolAgent²⁸¹, which leverages a scientific tool knowledge graph to automatically orchestrate the execution of complex analytical workflows. Coupled with automated laboratory systems, this architecture can close the loop from computational inference to physical experimentation.

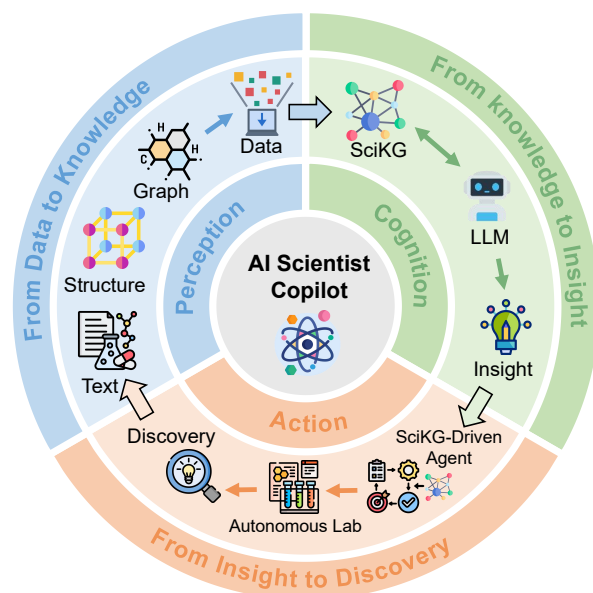


Figure 8. The autonomous scientific discovery flywheel driven by LLM agents and SciKGs.

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.

The next frontier lies in building 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 Challenges, Opportunities, and Future Directions

Despite the growing success, SciKGs remain in an early stage of development, with substantial technical and conceptual challenges that must be addressed before they can serve as reliable foundations for AI-driven discovery. At the same time, these challenges open opportunities for new research directions that could fundamentally reshape the landscape of scientific knowledge representation. In this section, we discuss four major challenges: data quality, interoperability, dynamism, and trustworthy reasoning, and highlight promising opportunities for advancing SciKGs. We further propose three complementary directions for next-generation SciKGs, aimed at enhancing their role as actionable knowledge infrastructures for scientific discovery.

6.1 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

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 ²⁶⁰ (2024)	Drug	Semantic Interface (Query Generation)	Grounding (Factual Verification)	M	Guide health information seeking
FactFinder ²⁶¹ (2024)	Drug	Semantic Interface (Query Generation)	Grounding (Factual Retrieval)	M	Life-science question answering
DDI-GPT ⁸⁴ (2024)	Drug	Reasoner (Prediction & Explanation)	Representation (Semantic Enhancement)	C	Explainable prediction of drug-drug interactions
Soman et al. ²⁴⁹ (2024)	Drug, Omics	Constructor, Interface (KG Construction, Text Generation)	Grounding (Knowledge Base & Traceability)	M, C	Drug repurposing and medical QA
BioLORD ²⁷⁵ (2024)	Drug, Omics	Reasoner (Semantic Representation Optimization)	Grounding (Knowledge Base & Semantic Support)	M	Enhance biomedical semantic similarity
HeCiX ²⁷⁰ (2024)	Drug, Omics	Semantic Interface (Format Conversion)	Grounding (Knowledge Base)	M	Enhance clinical trial research
KRAGEN ²⁷⁶ (2024)	Drug, Omics	Orchestrator (Plan Generation & Execution)	Grounding (Knowledge Base & Visualization)	M	Visualized biomedical QA system
MechGPT ²³⁰ (2024)	Material	Constructor, Reasoner, Orchestrator (KG Construction, Explanation, Multi-agent)	Grounding, Reasoning Constraints (Knowledge & Explainability)	C, S, I	Materials analysis and design
SciAgents ⁸³ (2024)	Material	Constructor, Reasoner, Generator (KG Construction, Analytical Reasoning, Hypothesis Generation)	Grounding (Knowledge Base)	M, I	Automated discovery in biomaterials science
MKG ¹⁴⁶ (2024)	Material	Constructor (KG Construction & Maintenance)	Grounding (Knowledge Base)	I	Multidisciplinary materials science discovery
OpenTCM ²⁸² (2025)	Drug	Interface, Reasoner, Constructor (Retrieval, Diagnosis, KG Construction)	Reasoning Constraints (Knowledge Retrieval Enhancement)	M	Traditional Chinese Medicine diagnosis
iKraph ²⁴ (2025)	Drug	Constructor (KG Construction)	Grounding (Knowledge Base)	S	Biomedical Research
KGT ⁴⁴ (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 ²⁷² (2025)	Drug, Omics	Generator, Orchestrator (Strategy & Code Generation)	Grounding (Knowledge Base)	S, I	Biomedical AI agent
Cat-KG ²⁶⁴ (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. ²⁵ (2025)	Chemistry	Constructor, Generator (KG Construction & Path Recommendation)	Grounding (Structured Knowledge Management)	S	Automated Retrosynthesis Planning of Macromolecules
KG-FM ²⁴³ (2025)	Material	Constructor, Reasoner (Multi-modal Extraction, QA & Reasoning)	Grounding (Knowledge Base & Visualization)	M	Improve LLM QA in framework materials
SciToolAgent ²⁸¹ (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.

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. 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. Moreover, barriers to data sharing, ranging from proprietary restrictions to inconsistent metadata standards, limit the reusability of valuable resources. Addressing interoperability will require the development of unified, extensible ontological frameworks, as well as technical infrastructures for federated knowledge integration. 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? 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. 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.

6.2 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 multi-modal, semantically unified knowledge backbones that integrate diverse data types and modalities. Foundational multi-modal 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 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 designing 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. For instance, in genomics, an agent could detect conflicting annotations about a gene's function, assess the

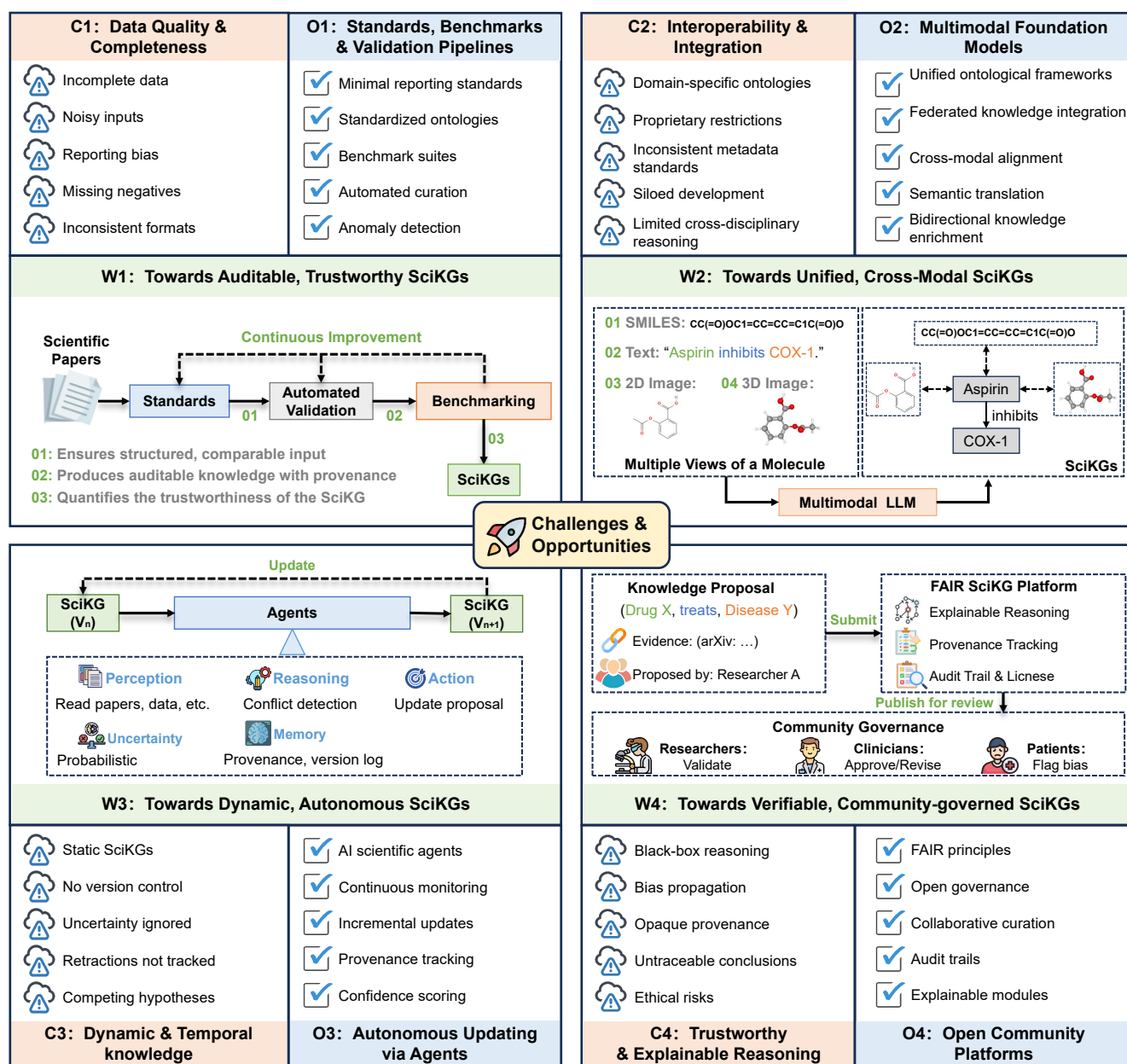


Figure 9. Challenges and Opportunities of SciKGs. This figure illustrates the major challenges (C1-C4) facing SciKGs, including data quality and completeness, interoperability and integration, dynamic and temporal knowledge, and trustworthy and explainable reasoning. Each challenge is paired with corresponding 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.

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 modeling, 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. 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.

6.3 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 annotations 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.

References

1. Shiflet, A. B. & Shiflet, G. W. *Introduction to computational science: Modeling and simulation for the sciences* (2014).
2. Leonelli, S. *Data-centric biology: A philosophical study* (2019).
3. Deshpande, D. *et al.* The evolution of computational research in a data-centric world. *Cell* **187**, 4449–4457 (2024).

4. Hey, T., Tansley, S., Tolle, K. M. *et al.* *The fourth paradigm: Data-intensive scientific discovery*, vol. 1 (Microsoft research Redmond, WA, 2009).
5. Szalay, A. & Gray, J. Science in an exponential world. *Nature* **440**, 413–414 (2006).
6. Hamid, J. S. *et al.* Data integration in genetics and genomics: Methods and challenges. *Human genomics and proteomics: HGP* **2009**, 869093 (2009).
7. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights* **14**, 1177932219899051 (2020).
8. Fillinger, S., de la Garza, L., Peltzer, A., Kohlbacher, O. & Nahnsen, S. Challenges of big data integration in the life sciences. *Analytical and bioanalytical chemistry* **411**, 6791–6800 (2019).
9. Qiu, Y. & Hu, Z. Data governance and open sharing in the fields of life sciences and medicine: A bibliometric analysis. *Digit. Heal.* **11**, 20552076251320302 (2025).
10. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *elife* **6**, e26726 (2017).
11. Fang, Y. *et al.* Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence* **5**, 542–553 (2023).
12. Hogan, A. *et al.* Knowledge graphs. *ACM Comput. Surv. (Csur)* **54**, 1–37 (2021).
13. Zou, X. A survey on application of knowledge graph. In *Journal of Physics: Conference Series*, vol. 1487, 012016 (IOP Publishing, 2020).
14. Suchanek, F. M., Kasneci, G. & Weikum, G. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, 697–706 (2007).
15. Liu, X., Mao, T., Shi, Y. & Ren, Y. Overview of knowledge reasoning for knowledge graph. *Neurocomputing* **585**, 127571 (2024).
16. Jin, X. *et al.* A survey on knowledge graph evolution: proliferation, dynamic embedding, and versioning. *Int. J. Web Grid Serv.* **21**, 88–111 (2025).
17. MacLean, F. Knowledge graphs and their applications in drug discovery. *Expert opinion on drug discovery* **16**, 1057–1069 (2021).
18. Zheng, X., Wang, B., Zhao, Y., Mao, S. & Tang, Y. A knowledge graph method for hazardous chemical management: Ontology design and entity identification. *Neurocomputing* **430**, 104–111 (2021).
19. Auer, S. *et al.* Towards a knowledge graph for science. In *Proceedings of the 8th international conference on web intelligence, mining and semantics*, 1–6 (2018).
20. Xu, J. *et al.* Pubmed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science. *Sci. Data* **12**, 1018 (2025).
21. Ye, Q. *et al.* A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nature communications* **12**, 6775 (2021).
22. Jeong, J., Lee, N., Shin, Y. & Shin, D. Intelligent generation of optimal synthetic pathways based on knowledge graph inference and retrosynthetic predictions using reaction big data. *Journal of the Taiwan Institute of Chemical Engineers* **130**, 103982 (2022).
23. Tu, R. *et al.* Drug repurposing using consilience of knowledge graph completion methods. *bioRxiv* 2023–05 (2024).
24. Zhang, Y. *et al.* A comprehensive large-scale biomedical knowledge graph for ai-powered data-driven biomedical research. *Nature Machine Intelligence* 1–13 (2025).
25. Ma, Q., Zhou, Y. & Li, J. Automated retrosynthesis planning of macromolecules using large language models and knowledge graphs. *Macromolecular Rapid Communications* 2500065 (2025).
26. Tripathy, R. K. *et al.* Effective integration of multi-omics with prior knowledge to identify biomarkers via explainable graph neural networks. *npj Syst. Biol. Appl.* **11**, 43 (2025).
27. Chen, J. *et al.* Knowledge graphs for the life sciences: Recent developments, challenges and opportunities. *arXiv:2309.17255* (2023).
28. Jarnac, L., Chabot, Y. & Couceiro, M. Uncertainty management in the construction of knowledge graphs: A survey. *arXiv:2405.16929* (2024).

29. Hofer, M., Obraczka, D., Saeedi, A., Köpcke, H. & Rahm, E. Construction of knowledge graphs: Current state and challenges. *Information* **15**, 509 (2024).
30. Zhao, Z., Luo, X., Chen, M. & Ma, L. A survey of knowledge graph construction using machine learning. *CMES-Computer Model. Eng. & Sci.* **139** (2024).
31. Bian, H. Llm-empowered knowledge graph construction: A survey. *arXiv:2510.20345* (2025).
32. Pan, S. *et al.* Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering* **36**, 3580–3599 (2024).
33. Ibrahim, N., Aboulela, S., Ibrahim, A. & Kashef, R. A survey on augmenting knowledge graphs (kgs) with large language models (llms): models, evaluation metrics, benchmarks, and challenges. *Discov. Artif. Intell.* **4**, 76 (2024).
34. Khorashadizadeh, H. *et al.* Research trends for the interplay between large language models and knowledge graphs. *arXiv:2406.08223* (2024).
35. Liu, B. *et al.* Large language models for knowledge graph embedding: A survey. *Mathematics* **13**, 2244 (2025).
36. Dai, X. *et al.* Large language models can better understand knowledge graphs than we thought. *Knowledge-Based Syst.* **312**, 113060 (2025).
37. Wang, X. *et al.* KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics* **9**, 176–194 (2021).
38. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
39. Yasunaga, M. *et al.* Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems* **35**, 37309–37323 (2022).
40. Kim, J., Kwon, Y., Jo, Y. & Choi, E. KG-GPT: A general framework for reasoning on knowledge graphs using large language models. *arXiv:2310.11220* (2023).
41. Luo, R. *et al.* BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics* **23**, bbac409 (2022).
42. Ouyang, S. *et al.* Structured chemistry reasoning with large language models. *arXiv:2311.09656* (2023).
43. Yang, R. *et al.* KG-Rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. *arXiv:2403.05881* (2024).
44. Feng, Y. *et al.* Knowledge graph-based thought: A knowledge graph-enhanced llm framework for pan-cancer question answering. *GigaScience* **14**, giae082 (2025).
45. Wu, J. *et al.* Medical Graph RAG: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv:2408.04187* (2024).
46. Jeong, M., Sohn, J., Sung, M. & Kang, J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics* **40**, i119–i129 (2024).
47. Meyer, L.-P. *et al.* Llm-assisted knowledge graph engineering: Experiments with chatgpt. In *Working conference on artificial intelligence development for a resilient and sustainable tomorrow*, 103–115 (Springer Fachmedien Wiesbaden Wiesbaden, 2023).
48. Ji, S., Pan, S., Cambria, E., Marttinen, P. & Yu, P. S. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems* **33**, 494–514 (2021).
49. Peng, C., Xia, F., Naseriparsa, M. & Osborne, F. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review* **56**, 13071–13102 (2023).
50. Zhong, L., Wu, J., Li, Q., Peng, H. & Wu, X. A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys* **56**, 1–62 (2023).
51. Liang, W., Meo, P. D., Tang, Y. & Zhu, J. A survey of multi-modal knowledge graphs: Technologies and trends. *ACM Computing Surveys* **56**, 1–41 (2024).
52. Zhao, H., Zi, C., Chen, A. & Li, J. A survey of cross-domain graph learning: Progress and future directions. *arXiv preprint arXiv:2503.11086* (2025).
53. Liang, X., Wang, Z. & Liu, J. A survey of large language model-augmented knowledge graphs for advanced complex product design. *J. Manuf. Syst.* **80**, 883–901 (2025).

54. Luo, H. *et al.* Drug-drug interactions prediction based on deep learning and knowledge graph: A review. *Iscience* **27** (2024).
55. Lu, Y., Goi, S. Y., Zhao, X. & Wang, J. Biomedical knowledge graph: A survey of domains, tasks, and real-world applications. *arXiv:2501.11632* (2025).
56. Wang, C., Yang, Y., Song, J. & Nan, X. Research progresses and applications of knowledge graph embedding technique in chemistry. *Journal of Chemical Information and Modeling* **64**, 7189–7213 (2024).
57. Cui, H. *et al.* A review on knowledge graphs for healthcare: Resources, applications, and promises. *arXiv preprint arXiv:2306.04802* (2023).
58. Sun, Z. *et al.* Knowledge graph in astronomical research with large language models: Quantifying driving forces in interdisciplinary scientific discovery. *arXiv preprint arXiv:2406.01391* (2024).
59. Díaz, R. & Xin, H. Knowledge graphs in heterogeneous catalysis: Recent advances and future opportunities. *Chin. J. Chem. Eng.* (2025).
60. Yao, L., Mao, C. & Luo, Y. KG-BERT: Bert for knowledge graph completion. *arXiv:1909.03193* (2019).
61. Venugopal, V., Pai, S. & Olivetti, E. MatKG: The largest knowledge graph in materials science—entities, relations, and link prediction through graph representation learning. *arXiv:2210.17340* (2022).
62. Wang, X. *et al.* KDGene: knowledge graph completion for disease gene prediction using interactional tensor decomposition. *Briefings Bioinforma.* **25**, bbae161 (2024).
63. Zhang, C., Zang, T. & Zhao, T. Kge-unit: toward the unification of molecular interactions prediction based on knowledge graph and multi-task learning on drug discovery. *Briefings in Bioinformatics* **25**, bbae043 (2024).
64. Cao, J., Fang, J., Meng, Z. & Liang, S. Knowledge graph embedding: A survey from the perspective of representation spaces. *ACM Comput. Surv.* **56**, 1–42 (2024).
65. Achiam, J. *et al.* Gpt-4 technical report. *arXiv:2303.08774* (2023).
66. Rydholm, E. *et al.* Expanding the chemical space using a chemical reaction knowledge graph. *Digit. Discov.* **3**, 1378–1388 (2024).
67. Yang, J., Lu, G., He, S., Cao, Q. & Liu, Y. A novel model for relation prediction in knowledge graphs exploiting semantic and structural feature integration. *Sci. Reports* **14**, 12962 (2024).
68. Nisar, U. *et al.* Graph neural networks for drug–drug interaction prediction—predicting safe drug pairings with ai. *Eng. Proc.* **107**, 42 (2025).
69. Kulkarni, A. *et al.* Scientific hypothesis generation and validation: Methods, datasets, and future directions. *arXiv:2505.04651* (2025).
70. Vogt, L., Kuhn, T. & Hoehndorf, R. Semantic units: organizing knowledge graphs into semantically meaningful units of representation. *J. biomedical semantics* **15**, 7 (2024).
71. Wehner, C., Iliopoulou, C. & Besold, T. R. From latent to lucid: Transforming knowledge graph embeddings into interpretable structures. *arXiv:2406.01759* (2024).
72. Dibowski, H. Full traceability and provenance for knowledge graphs. In *Formal Ontology in Information Systems*, 223–237 (IOS Press, 2024).
73. Zhou, H., Bamler, R., Wu, C. M. & Tejero-Cantero, Á. Predictive, scalable and interpretable knowledge tracing on structured domains. *arXiv:2403.13179* (2024).
74. Garcia Trelles, E., Schweizer, C., Thomas, A., von Hartrott, P. & Janka-Ramm, M. Digitalizing material knowledge: A practical framework for ontology-driven knowledge graphs in process chains. *Appl. Sci.* **14**, 11683 (2024).
75. Benson, D. A. *et al.* Genbank. *Nucleic acids research* **41**, D36–D42 (2012).
76. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).
77. Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. & Morissette, J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. biomedical informatics* **41**, 706–716 (2008).
78. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. biotechnology* **25**, 1251–1255 (2007).
79. Hamilton, W. L., Ying, R. & Leskovec, J. Representation learning on graphs: Methods and applications. *arXiv:1709.05584* (2017).

80. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. & Yakhnenko, O. Translating embeddings for modeling multi-relational data. *Adv. neural information processing systems* **26** (2013).
81. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. *Adv. neural information processing systems* **30** (2017).
82. Chen, B. & Bertozzi, A. L. AutoKG: Efficient automated knowledge graph generation for language models. In *2023 IEEE International Conference on Big Data (BigData)*, 3117–3126 (IEEE, 2023).
83. Ghafarollahi, A. & Buehler, M. J. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials* **24**13523 (2024).
84. Xu, C., Bulusu, K. C., Pan, H. & Elemento, O. DDI-GPT: Explainable prediction of drug-drug interactions using large language models enhanced with knowledge graphs. *BioRxiv* 2024–12 (2024).
85. Kim, S. *et al.* Pubchem 2023 update. *Nucleic acids research* **51**, D1373–D1380 (2023).
86. Consortium, U. UniProt: A worldwide hub of protein knowledge. *Nucleic acids research* **47**, D506–D515 (2019).
87. Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials* **1** (2013).
88. Ohta, T., Tateisi, Y., Kim, J.-D., Mima, H. & Tsujii, J. The genia corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the human language technology conference*, 73–77 (2002).
89. Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. Overview of biocreative: Critical assessment of information extraction for biology. *BMC bioinformatics* **6**, S1 (2005).
90. Degtyarenko, K. *et al.* ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic acids research* **36**, D344–D350 (2007).
91. Ekbal, A., Saha, S. & Sikdar, U. K. Biomedical named entity extraction: Some issues of corpus compatibilities. *SpringerPlus* **2**, 601 (2013).
92. Ravikumar, K., Rastegar-Mojarad, M. & Liu, H. Belminer: Adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences. *Database* **2017**, baw156 (2017).
93. Neumann, M., King, D., Beltagy, I. & Ammar, W. ScispaCy: Fast and robust models for biomedical natural language processing. *arXiv:1902.07669* (2019).
94. Fabian, B. *et al.* Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv:2011.13230* (2020).
95. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **3**, 1–23 (2021).
96. Sung, M. *et al.* BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* **38**, 4837–4839 (2022).
97. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
98. Song, Y., Miret, S. & Liu, B. MatSci-NLP: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. *arXiv preprint arXiv:2305.08264* (2023).
99. Lo, K., Wang, L. L., Neumann, M., Kinney, R. & Weld, D. S. S2ORC: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782* (2019).
100. Beltagy, I., Lo, K. & Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv:1903.10676* (2019).
101. Shamsabadi, M., D’Souza, J. & Auer, S. Large language models for scientific information extraction: An empirical study for virology. *arXiv:2401.10040* (2024).
102. Li, H. *et al.* A survey on large language model acceleration based on kv cache management. *arXiv:2412.19442* (2024).
103. Feng, K. *et al.* SciKnowEval: A comprehensive dataset for evaluating scientific knowledge of large language models. In *NeurIPS 2025 AI for Science Workshop*.
104. Yu, J. *et al.* Scicueval: A comprehensive dataset for evaluating scientific context understanding in large language models. *arXiv:2505.15094* (2025).

105. Zhu, Z., Tang, Y., Zhang, Q. & Ding, K. Synergizing large language models and knowledge graphs in science: A survey. In *NeurIPS 2025 AI for Science Workshop*.
106. Dagdelen, J. *et al.* Structured information extraction from scientific text with large language models. *Nat. communications* **15**, 1418 (2024).
107. Zhang, J. *et al.* A study of biomedical relation extraction using gpt models. *AMIA Summits on Transl. Sci. Proc.* **2024**, 391 (2024).
108. Choi, S. & Jung, Y. Knowledge graph construction: Extraction, learning, and evaluation. *Appl. Sci.* **15**, 3727 (2025).
109. Yang, Y. *et al.* Pseudo-knowledge graph: Meta-path guided retrieval and in-graph text for rag-equipped llm. *arXiv:2503.00309* (2025).
110. Noy, N. F. *et al.* BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucleic acids research* **37**, W170–W173 (2009).
111. Silva, M. C., Faria, D. & Pesquita, C. Matching multiple ontologies to build a knowledge graph for personalized medicine. In *European Semantic Web Conference*, 461–477 (Springer, 2022).
112. Chen, J., Zhang, R., Guo, J., Fan, Y. & Cheng, X. Fedmatch: Federated learning over heterogeneous question answering data. In *Proceedings of the 30th ACM international conference on information & knowledge management*, 181–190 (2021).
113. Fecho, K. *et al.* An approach for collaborative development of a federated biomedical knowledge graph-based question-answering system: Question-of-the-month challenges. *J. clinical translational science* **7**, e214 (2023).
114. Fareedi, A. A., Ismail, M., Ghazawneh, A., Ahmed, S. & Bukhari, S. A. C. Empowering health data fusion: A federated virtual knowledge graph approach leveraging the ontop platform. *Mediterr. Conf. on Inf. Syst.* (2024).
115. Kokash, N. *et al.* Ontology-and llm-based data harmonization for federated learning in healthcare. *arXiv:2505.20020* (2025).
116. Osman, I., Pileggi, S. F. & Yahia, S. B. Uncertainty in automated ontology matching: Lessons from an empirical evaluation. *Appl. Sci.* **14**, 4679 (2024).
117. Wu, N., Lai, X., Chen, M. & Pan, Y. Ontology matching and repair based on semantic association and probabilistic logic. *IEICE TRANSACTIONS on Inf. Syst.* **107**, 1433–1443 (2024).
118. Liu, Y., Pena, E., Santos, A., Wu, E. & Freire, J. Magneto: Combining small and large language models for schema matching. *arXiv:2412.08194* (2024).
119. Santos Jr, E., Jurmain, J. & Ragazzi, A. Bayesian-knowledge driven ontologies: A framework for fusion of semantic knowledge under uncertainty and incompleteness. *Plos one* **19**, e0296864 (2024).
120. Dai, Y., Wang, S., Xiong, N. N. & Guo, W. A survey on knowledge graph embedding: Approaches, applications and benchmarks. *Electronics* **9**, 750 (2020).
121. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).
122. Zhang, F., Sun, B., Diao, X., Zhao, W. & Shu, T. Prediction of adverse drug reactions based on knowledge graph embedding. *BMC Med. Informatics Decis. Mak.* **21**, 38 (2021).
123. Joshi, P., Masilamani, V. & Mukherjee, A. A knowledge graph embedding based approach to predict the adverse drug reactions using a deep neural network. *J. biomedical informatics* **132**, 104122 (2022).
124. Chen, S. *et al.* An effective framework for predicting drug–drug interactions based on molecular substructures and knowledge graph neural network. *Comput. Biol. Medicine* **169**, 107900 (2024).
125. Alberts, M., Schilter, O., Zipoli, F., Hartrampf, N. & Laino, T. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *Adv. Neural Inf. Process. Syst.* **37**, 125780–125808 (2024).
126. Huang, Q. *et al.* Multi-dimensional perceptual quality assessment for magnetic resonance images. *Heal. Inf. Sci. Syst.* **13**, 65 (2025).
127. Burgess, J. *et al.* Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19552–19564 (2025).
128. Chen, Z. *et al.* Knowledge graphs meet multi-modal learning: A comprehensive survey. *arXiv:2402.05391* (2024).

129. Lee, J., Wang, Y., Li, J. & Zhang, M. Multimodal reasoning with multimodal knowledge graph. *arXiv:2406.02030* (2024).
130. Lopez, V. *et al.* Enhancing foundation models for scientific discovery via multimodal knowledge graph representations. *J. Web Semant.* **84**, 100845 (2025).
131. Buehler, M. J. Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Mach. Learn. Sci. Technol.* **5**, 035083 (2024).
132. Hua, M. *et al.* KG-IBL: Knowledge graph driven incremental broad learning for few-shot specific emitter identification. *IEEE Transactions on Information Forensics and Security* (2024).
133. Cao, X. *et al.* Knowledge graph enhanced generative multi-modal models for class-incremental learning. *arXiv:2503.18403* (2025).
134. Liu, J. *et al.* Towards continual knowledge graph embedding via incremental distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 8759–8768 (2024).
135. Tian, Z., Zhang, D. & Dai, H.-N. Continual learning on graphs: A survey. *arXiv preprint arXiv:2402.06330* (2024).
136. Liu, J. *et al.* Fast and continual knowledge graph embedding via incremental lora. *arXiv preprint arXiv:2407.05705* (2024).
137. Kleinstaubert, E., Al Mustafa, T., Zander, F., König-Ries, B. & Babalou, S. Managing provenance data in knowledge graph management platforms. *Datenbank-Spektrum* **24**, 43–52 (2024).
138. Zhao, X. *et al.* Agentigraph: An interactive knowledge graph platform for llm-based chatbots utilizing private data. *arXiv:2410.11531* (2024).
139. Lu, Y. & Wang, J. Karma: Leveraging multi-agent llms for automated knowledge graph enrichment. *arXiv:2502.06472* (2025).
140. Terdalkar, H., Bonifati, A. & Mauri, A. Graph repairs with large language models: An empirical study. In *Proceedings of the 8th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, 1–10 (2025).
141. Ren, S. *et al.* Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv:2503.24047* (2025).
142. Bang, D., Lim, S., Lee, S. & Kim, S. Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. *Nature Communications* **14**, 3570 (2023).
143. Sudhahar, S. *et al.* An experimentally validated approach to automated biological evidence generation in drug discovery using knowledge graphs. *Nature Communications* **15**, 5703 (2024).
144. Huang, K. *et al.* A foundation model for clinician-centered drug repurposing. *Nature Medicine* **30**, 3601–3613 (2024).
145. Lin, X., Quan, Z., Wang, Z.-J., Ma, T. & Zeng, X. KGNN: Knowledge graph neural network for drug-drug interaction prediction. In *IJCAI*, vol. 380, 2739–2745 (2020).
146. Wu, D., Sun, W., He, Y., Chen, Z. & Luo, X. 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*, vol. 38, 10216–10224 (2024).
147. Wang, Y., Yang, Z. & Yao, Q. Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning. *Communications Medicine* **4**, 59 (2024).
148. Luo, Y. *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications* **8**, 573 (2017).
149. Hoang, T. L. *et al.* Knowledge enhanced representation learning for drug discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 10544–10552 (2024).
150. Shen, X. *et al.* TarIKGC: A target identification tool using semantics-enhanced knowledge graph completion with application to cdk2 inhibitor discovery. *Journal of Medicinal Chemistry* (2025).
151. Shang, J., Xiao, C., Ma, T., Li, H. & Sun, J. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 1126–1133 (2019).
152. Evangelista, J. E. *et al.* Toxicology knowledge graph for structural birth defects. *Communications Medicine* **3**, 98 (2023).
153. Bhoi, S., Lee, M. L., Hsu, W., Fang, H. S. A. & Tan, N. C. Personalizing medication recommendation with a graph-based approach. *ACM Transactions on Information Systems (TOIS)* **40**, 1–23 (2021).

154. Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. *Nature communications* **12**, 1796 (2021).
155. Kanehisa, M. The KEGG database. In *'In silico' simulation of biological processes: Novartis Foundation Symposium 247*, vol. 247, 91–103 (Wiley Online Library, 2002).
156. Madeddu, F. *et al.* VitaGraph: Building a knowledge graph for biologically relevant learning tasks. *arXiv:2505.11185* (2025).
157. Xu, T. *et al.* NAFLDKB: A knowledge base and platform for drug development against nonalcoholic fatty liver disease. *Journal of Chemical Information and Modeling* **64**, 2817–2828 (2023).
158. Islam, M. K. *et al.* Molecular-evaluated and explainable drug repurposing for covid-19 using ensemble knowledge graph embedding. *Sci. Reports* **13**, 3643 (2023).
159. Knox, C. *et al.* DrugBank 6.0: The DrugBank knowledgebase for 2024. *Nucleic acids research* **52**, D1265–D1275 (2024).
160. Tatonetti, N. P., Ye, P. P., Daneshjou, R. & Altman, R. B. Data-driven prediction of drug effects and interactions. *Science translational medicine* **4**, 125ra31–125ra31 (2012).
161. Su, X., Hu, L., You, Z., Hu, P. & Zhao, B. Attention-based knowledge graph representation learning for predicting drug-drug interactions. *Briefings in bioinformatics* **23**, bbac140 (2022).
162. Yu, Y. *et al.* SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics* **37**, 2988–2995 (2021).
163. Cheng, J., Zhang, Y., Zhang, H., Ji, S. & Lu, M. Transfol: A logical query model for complex relational reasoning in drug-drug interaction. *IEEE J. Biomed. Heal. Informatics* **28**, 4975–4985 (2024).
164. Lin, X., Yin, Z., Zhang, X. & Hu, J. KGRLFF: detecting drug-drug interactions based on knowledge graph representation learning and feature fusion. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* (2024).
165. Karim, M. R. *et al.* Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-lstm network. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 113–123 (2019).
166. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. & Kanehisa, M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **24**, i232–i240 (2008).
167. Walsh, B., Mohamed, S. K. & Nováček, V. BioKG: A knowledge graph for relational learning on biological data. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3173–3180 (2020).
168. Geleta, D. *et al.* Biological insights knowledge graph: an integrated knowledge graph to support drug development. *Biorxiv* 2021–10 (2021).
169. Mohamed, S. K., Nováček, V. & Nounu, A. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* **36**, 603–610 (2020).
170. Quan, Y. *et al.* Evolution-strengthened knowledge graph enables predicting the targetability and druggability of genes. *PNAS nexus* **2**, pgad147 (2023).
171. Zeng, X. *et al.* Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics* **36**, 2805–2812 (2020).
172. Ma, T., Lin, X., Song, B., Yu, P. S. & Zeng, X. KG-MTL: Knowledge graph enhanced multi-task learning for molecular interaction. *IEEE Transactions on Knowledge and Data Engineering* **35**, 7068–7081 (2022).
173. Feng, Z., Shen, Z., Li, H. & Li, S. E-TSN: An interactive visual exploration platform for target–disease knowledge mapping from literature. *Briefings in Bioinformatics* **23**, bbac465 (2022).
174. Talukder, A. K. *et al.* Drugomics: Knowledge graph & ai to construct physicians' brain digital twin to prevent drug side-effects and patient harm. In *International Conference on Big Data Analytics*, 149–158 (Springer, 2022).
175. Bresso, E. *et al.* Investigating adr mechanisms with explainable AI: A feasibility study with knowledge graph mining. *BMC medical informatics decision making* **21**, 171 (2021).
176. Romano, J. D., Hao, Y., Moore, J. H. & Penning, T. M. Automating predictive toxicology using comptoxai. *Chem. research toxicology* **35**, 1370–1382 (2022).
177. Muñoz, E., Nováček, V. & Vandenbussche, P.-Y. Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models. *Briefings bioinformatics* **20**, 190–202 (2019).

178. Feng, F. *et al.* GenomicKB: A knowledge graph for the human genome. *Nucleic Acids Research* **51**, D950–D956 (2023).
179. Jha, A. *et al.* GenomicsKG: A knowledge graph to visualize poly-omics data. *J Adv Heal.* **1**, 70–84 (2019).
180. Mulero-Hernández, J., Mironov, V., Miñarro-Giménez, J. A., Kuiper, M. & Fernández-Breis, J. T. Integration of chromosome locations and functional aspects of enhancers and topologically associating domains in knowledge graphs enables versatile queries about gene regulation. *Nucleic Acids Research* **52**, e69–e69 (2024).
181. Zaripova, K., Özsoy, E., Navab, N. & Farshad, A. PhenoKG: Knowledge graph-driven gene discovery and patient insights from phenotypes alone. *arXiv:2506.13119* (2025).
182. Youn, J., Rai, N. & Tagkopoulos, I. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. *Nat. communications* **13**, 2360 (2022).
183. Stear, B. J. *et al.* PetaGraph: A large-scale unifying knowledge graph framework for integrating biomolecular and biomedical data. *Sci. Data* **11**, 1338 (2024).
184. Yang, K. *et al.* Heterogeneous network embedding for identifying symptom candidate genes. *J. Am. Med. Informatics Assoc.* **25**, 1452–1459 (2018).
185. Vlietstra, W. J., Vos, R., van Mulligen, E. M., Jenster, G. W. & Kors, J. A. Identifying genes targeted by disease-associated non-coding snps with a protein knowledge graph. *Plos one* **17**, e0271395 (2022).
186. Zhu, Q. *et al.* An integrative knowledge graph for rare diseases, derived from the genetic and rare diseases information center (gard). *J. biomedical semantics* **11**, 13 (2020).
187. Dimitrakopoulos, C. *et al.* Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **34**, 2441–2448 (2018).
188. Wang, X., Gong, Y., Yi, J. & Zhang, W. Predicting gene-disease associations from the heterogeneous network using graph embedding. In *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 504–511 (2019).
189. Zhou, H. & Skolnick, J. A knowledge-based approach for predicting gene–disease associations. *Bioinformatics* **32**, 2831–2838 (2016).
190. Santos, A. *et al.* A knowledge graph to interpret clinical proteomics data. *Nature biotechnology* **40**, 692–702 (2022).
191. Binder, J. *et al.* Machine learning prediction and tau-based screening identifies potential alzheimer’s disease genes relevant to immunity. *Communications Biology* **5**, 125 (2022).
192. Zhang, N. *et al.* OntoProtein: Protein pretraining with gene ontology embedding. *arXiv:2201.11147* (2022).
193. Cheng, S., Liang, X., Bi, Z., Zhang, N. & Chen, H. ProteinKG65: A knowledge graph for protein science. *arXiv:2207.10080* (2022).
194. Gray, A. J., Papadopoulos, P., Asif, I., Mičetić, I. & Hatos, A. Creating and exploiting the intrinsically disordered protein knowledge graph (idp-kg). In *CEUR Workshop Proceedings*, vol. 3127, 1–10 (CEUR-WS, 2022).
195. Nováček, V. *et al.* Accurate prediction of kinase-substrate networks using knowledge graphs. *PLoS computational biology* **16**, e1007578 (2020).
196. Chen, X.-H. *et al.* Transporter proteins knowledge graph construction and its application in drug development. *Comput. Struct. Biotechnol. J.* **21**, 2973–2984 (2023).
197. Wang, X. *et al.* A novel approach for target deconvolution from phenotype-based screening using knowledge graph. *Sci. Reports* **15**, 2414 (2025).
198. Chen, M. *et al.* Unified knowledge-guided molecular graph encoder with multimodal fusion and multi-task learning. *Neural Networks* **184**, 107068 (2025).
199. Ni, S. *et al.* Identifying compound-protein interactions with knowledge graph embedding of perturbation transcriptomics. *Cell Genomics* **4** (2024).
200. Shao, X. *et al.* Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with spatalk. *Nature communications* **13**, 4429 (2022).
201. Cavalleri, E. *et al.* An ontology-based knowledge graph for representing interactions involving rna molecules. *Scientific Data* **11**, 906 (2024).
202. Delmas, M. *et al.* Building a knowledge graph from public databases and scientific literature to extract associations between chemicals and diseases. *Bioinformatics* **37**, 3896–3904 (2021).

203. Xiao, F., Huang, C., Chen, A., Xiao, W. & Li, Z. Identification of metabolite-disease associations based on knowledge graph. *Metabolomics* **21**, 32 (2025).
204. Qin, G. *et al.* Generating biomedical knowledge graphs from knowledge bases, registries, and multiomic data. *bioRxiv* (2024).
205. Sun, H. *et al.* MMiKG: A knowledge graph-based platform for path mining of microbiota–mental diseases interactions. *Briefings in bioinformatics* **24**, bbad340 (2023).
206. Liu, T. *et al.* Predicting the relationships between gut microbiota and mental disorders with knowledge graphs. *Heal. information science systems* **9**, 3 (2020).
207. Fu, C., Huang, Z., van Harmelen, F., He, T. & Jiang, X. Food4healthKG: Knowledge graphs for food recommendations based on gut microbiota and mental health. *Artif. Intell. Medicine* **145**, 102677 (2023).
208. Goetz, S. L., Glen, A. K. & Glusman, G. MicrobiomeKG: Bridging microbiome research and host health through knowledge graphs. *bioRxiv* (2024).
209. Jha, A., Khan, Y., Sahay, R. & d’Aquin, M. Metastatic site prediction in breast cancer using omics knowledge graph and pattern mining with kirchhoff’s law traversal. *bioRxiv* 2020–07 (2020).
210. Di Maria, A. *et al.* BioTAGME: A comprehensive platform for biological knowledge network analysis. *Frontiers in Genetics* **13**, 855739 (2022).
211. Struck, A. *et al.* Exploring integrative analysis using the biomedical evidence graph. *JCO Clin. Cancer Informatics* **4**, 147–159 (2020).
212. Fernández-Torras, A., Duran-Frigola, M., Bertoni, M., Locatelli, M. & Aloy, P. Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque. *Nat. communications* **13**, 5304 (2022).
213. Segler, M. H. & Waller, M. P. Modelling chemical reasoning to predict and invent reactions. *Chemistry—A European Journal* **23**, 6118–6128 (2017).
214. McDermott, M. J., Dwaraknath, S. S. & Persson, K. A. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nature communications* **12**, 3097 (2021).
215. Xie, J., Wang, Y., Rao, J., Zheng, S. & Yang, Y. Self-supervised contrastive molecular representation learning with a chemical synthesis knowledge graph. *Journal of Chemical Information and Modeling* **64**, 1945–1954 (2024).
216. O’Ryan, C. *et al.* An automated approach for domain-specific knowledge graph generation— graph measures and characterization. *Journal of Chemical Information and Modeling* (2025).
217. Farazi, F. *et al.* Knowledge graph approach to combustion chemistry and interoperability. *ACS omega* **5**, 18342–18348 (2020).
218. Zhou, X. *et al.* Marie and bert— a knowledge graph embedding based question answering system for chemistry. *ACS omega* **8**, 33039–33057 (2023).
219. Li, B. & Chen, H. Prediction of compound synthesis accessibility based on reaction knowledge graph. *Molecules* **27**, 1039 (2022).
220. Deagen, M. E. *et al.* Fair and interactive data graphics from a scientific knowledge graph. *Sci. Data* **9**, 239 (2022).
221. Fang, Y. *et al.* Molecular contrastive learning with chemical element knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 3968–3976 (2022).
222. Zhang, Y., Li, Z., Duan, B., Qin, L. & Peng, J. MKGE: Knowledge graph embedding with molecular structure information. *Computational Biology and Chemistry* **100**, 107730 (2022).
223. Jiang, P. *et al.* Bi-level contrastive learning for knowledge-enhanced molecule representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 352–360 (2025).
224. Onishi, T., Kadohira, T. & Watanabe, I. Relation extraction with weakly supervised learning based on process-structure-property-performance reciprocity. *Sci. technology advanced materials* **19**, 649–659 (2018).
225. McCusker, J. P. *et al.* Nanomine: A knowledge graph for nanocomposite materials science. In *International semantic web conference*, 144–159 (2020).
226. Nie, Z. *et al.* Automating materials exploration with a semantic knowledge graph for li-ion battery cathodes. *Advanced Functional Materials* **32**, 2201437 (2022).

227. Gao, Y., Wang, L., Chen, X., Du, Y. & Wang, B. Revisiting electrocatalyst design by a knowledge graph of cu-based catalysts for co2 reduction. *ACS Catalysis* **13**, 8525–8534 (2023).
228. Venugopal, V. & Olivetti, E. MatKG: An autonomously generated knowledge graph in material science. *Sci. Data* **11**, 217 (2024).
229. Ye, Y. *et al.* Construction and application of materials knowledge graph in multidisciplinary materials science via large language model. *Advances in Neural Information Processing Systems* **37**, 56878–56897 (2024).
230. Buehler, M. J. Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design. *ACS Eng. Au* **4**, 241–277 (2024).
231. Mrdjenovich, D. *et al.* PropNet: A knowledge graph for materials science. *Matter* **2**, 464–480 (2020).
232. Liu, J. & Qian, Q. Reinforcement learning-based knowledge graph reasoning for aluminum alloy applications. *Computational Materials Science* **221**, 112075 (2023).
233. Song, G. *et al.* Bridging the semantic-numerical gap: A numerical reasoning method of cross-modal knowledge graph for material property prediction. *arXiv:2312.09744* (2023).
234. Statt, M. J. *et al.* The materials experiment knowledge graph. *Digit. Discov.* **2**, 909–914 (2023).
235. Durmaz, A. R., Thomas, A., Mishra, L., Murthy, R. N. & Straub, T. An ontology-based text mining dataset for extraction of process-structure-property entities. *Sci. data* **11**, 1112 (2024).
236. Huang, C., Chen, C., Shi, L. & Chen, C. Material property prediction with element attribute knowledge graphs and multimodal representation learning. *arXiv:2411.08414* (2024).
237. Shu, C., Xin, Z. & Xie, C. EBSD grain knowledge graph representation learning for material structure-property prediction. In *China Conference on Knowledge Graph and Semantic Computing*, 3–15 (2021).
238. Shu, C., He, J., Xue, G. & Xie, C. Grain knowledge graph representation learning: A new paradigm for microstructure-property prediction. *Crystals* **12**, 280 (2022).
239. Zhang, Y. *et al.* A materials terminology knowledge graph automatically constructed from text corpus. *Scientific Data* **11**, 600 (2024).
240. Guo, P., Meng, W. & Bao, Y. Knowledge graph-guided data-driven design of ultra-high-performance concrete (uhpc) with interpretability and physicochemical reaction discovery capability. *Construction and Building Materials* **430**, 136502 (2024).
241. Anand, A., Kumari, P. & Kalyani, A. K. High throughput screening of new piezoelectric materials using graph machine learning and knowledge graph approach. *Computational Materials Science* **246**, 113445 (2025).
242. Zheng, X. *et al.* High-throughput computing assisted by knowledge graph to study the correlation between microstructure and mechanical properties of 6xxx aluminum alloy. *Materials* **15**, 5296 (2022).
243. Bai, X. *et al.* Construction of a knowledge graph for framework material enabled by large language models and its application. *npj Computational Materials* **11**, 51 (2025).
244. Desheng, C., Jian, S., Mingxin, L. & Sensen, X. Digital twin-based fault diagnosis platform for final rolling temperature in hot strip production. *Materials* **16**, 7021 (2023).
245. Liu, T. *et al.* Exploring the microbiota-gut-brain axis for mental disorders with knowledge graphs. *Journal of Artificial Intelligence for Medical Sciences* **1**, 30–42 (2021).
246. To, V.-T. *et al.* KGG: Knowledge-guided graph self-supervised learning to enhance molecular property predictions. *J. Chem. Inf. Model.* **65**, 9443–9458 (2025).
247. Qin, C. *et al.* Inverse design of semiconductor materials with deep generative models. *J. Mater. Chem. A* **12**, 22689–22702 (2024).
248. Hu, Z. & Yan, W. Data-driven modeling of process-structure-property relationships in metal additive manufacturing. *Npj Adv. Manuf.* **1**, 3 (2024).
249. Soman, K. *et al.* Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics* **40**, btae560 (2024).
250. Liu, H., Wang, S., Zhu, Y., Dong, Y. & Li, J. Knowledge graph-enhanced large language models via path selection. *arXiv preprint arXiv:2406.13862* (2024).

251. Pons, G., Bilalli, B. & Queral, A. Knowledge graphs for enhancing large language models in entity disambiguation. In *International Semantic Web Conference*, 162–179 (Springer, 2024).
252. Luo, S. *et al.* Generating multiple choice questions from scientific literature via large language models. In *2024 IEEE International Conference on Knowledge Graph*, 219–226 (2024).
253. Luo, L., Yang, C., Kharlamov, E. & Pan, S. Integrating large language models and knowledge graphs for next-level agi. In *Companion Proceedings of the ACM on Web Conference 2025*, 33–36 (2025).
254. Fang, J. *et al.* LightMem: Lightweight and efficient memory-augmented generation. *arXiv:2510.18866* (2025).
255. Sriramanan, G. *et al.* LLM-Check: Investigating detection of hallucinations in large language models. *Adv. Neural Inf. Process. Syst.* **37**, 34188–34216 (2024).
256. Zhang, Z. *et al.* LLM hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *Proc. ACM on Softw. Eng.* **2**, 481–503 (2025).
257. Ke, Y. H. *et al.* Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digit. Medicine* **8**, 187 (2025).
258. Zhou, H. *et al.* A collaborative large language model for drug analysis. *Nat. Biomed. Eng.* 1–12 (2025).
259. Dang, H. A., Tran, V. & Nguyen, L.-M. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Front. Artif. Intell.* **8**, 1622292 (2025).
260. Yan, Y., Hou, Y., Xiao, Y., Zhang, R. & Wang, Q. Knownet: Guided health information seeking from llms via knowledge graph integration. *IEEE Transactions on Visualization and Computer Graphics* (2024).
261. Steinigen, D. *et al.* Fact finder—enhancing domain expertise of large language models by incorporating knowledge graphs. *arXiv:2408.03010* (2024).
262. Sui, Y. *et al.* Fidelis: Faithful reasoning in large language model for knowledge graph question answering. *arXiv preprint arXiv:2405.13873* (2024).
263. Luo, L. *et al.* Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. *arXiv preprint arXiv:2410.13080* (2024).
264. Fu, F. *et al.* Synergizing knowledge graph and large language model for relay catalysis pathway recommendation. *National Science Review* **12**, nwaf271 (2025).
265. Zeng, X., Tu, X., Liu, Y., Fu, X. & Su, Y. Toward better drug discovery with knowledge graph. *Curr. opinion structural biology* **72**, 114–126 (2022).
266. Cai, B. *et al.* Temporal knowledge graph completion: A survey. *arXiv preprint arXiv:2201.08236* (2022).
267. Zhang, Y. *et al.* A survey on temporal knowledge graph embedding: Models and applications. *Knowledge-Based Syst.* **304**, 112454 (2024).
268. Callahan, T. J. *et al.* An open source knowledge graph ecosystem for the life sciences. *Sci. Data* **11**, 363 (2024).
269. Lin, J., Wang, S., Guo, X., Shun, J. & Zhu, Y. Temporal reasoning with large language models augmented by evolving knowledge graphs. *arXiv preprint arXiv:2509.15464* (2025).
270. Kulkarni, P. S., Jain, M., Sheshanarayana, D. & Parthiban, S. HeCiX: Integrating knowledge graphs and large language models for biomedical research. *arXiv:2407.14030* (2024).
271. Hatem, S., Khoriba, G., Gad-Elrab, M. H. & ElHelw, M. Up to date: automatic updating knowledge graphs using llms. *Procedia Comput. Sci.* **244**, 327–334 (2024).
272. Matsumoto, N. *et al.* ESCARGOT: An AI agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for enhanced reasoning. *Bioinformatics* **41**, btaf031 (2025).
273. Yoshitake, M. & Nagata, T. A method for llm-based construction of a materials property knowledge graph: A case study. *Appl. Sci.* **15**, 10511 (2025).
274. Dreger, M., Malek, K. & Eikerling, M. Large language models for knowledge graph extraction from tables in materials science. *Digit. Discov.* (2025).
275. Remy, F., Demuynck, K. & Demeester, T. BioLORD-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association* **31**, 1844–1855 (2024).

276. Matsumoto, N. *et al.* KRAGEN: A knowledge graph-enhanced rag framework for biomedical problem solving using large language models. *Bioinformatics* **40**, btae353 (2024).
277. Wu, H., Shi, W. & Wang, M. D. Developing a novel causal inference algorithm for personalized biomedical causal graph learning using meta machine learning. *BMC Med. Informatics Decis. Mak.* **24**, 137 (2024).
278. Bai, J. *et al.* A dynamic knowledge graph approach to distributed self-driving laboratories. *Nat. Commun.* **15**, 462 (2024).
279. Ran, N., Wang, Y., Zhang, X. & Allmendinger, R. MOLLM: Multi-objective large language model for molecular design—optimizing with experts. *arXiv preprint arXiv:2502.12845* (2025).
280. Bai, X. *et al.* An integrated ai system for multi-objective screening of mof materials. *Sep. Purif. Technol.* 133939 (2025).
281. Ding, K. *et al.* SciToolAgent: A knowledge graph-driven scientific agent for multi-tool integration. *Nat. Comput. Sci.* (2025).
282. He, J. *et al.* OpenTCM: A graphrag-empowered LLM-based system for traditional chinese medicine knowledge retrieval and diagnosis. *arXiv:2504.20118* (2025).
283. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* **44**, D1045–D1053 (2016).
284. Davis, A. P. *et al.* Comparative toxicogenomics database (ctd): update 2021. *Nucleic acids research* **49**, D1138–D1143 (2021).
285. Piñero, J. *et al.* DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research* gkw943 (2016).
286. Ursu, O. *et al.* DrugCentral: Online drug compendium. *Nucleic acids research* gkw993 (2016).
287. Barbarino, J. M., Whirl-Carrillo, M., Altman, R. B. & Klein, T. E. PharmGKB: A worldwide resource for pharmacogenomic information. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **10**, e1417 (2018).
288. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The sider database of drugs and side effects. *Nucleic acids research* **44**, D1075–D1079 (2016).
289. Cunningham, F. *et al.* Ensembl 2022. *Nucleic acids research* **50**, D988–D995 (2022).
290. Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic acids research* **46**, D649–D655 (2018).
291. Paysan-Lafosse, T. *et al.* Interpro in 2022. *Nucleic acids research* **51**, D418–D427 (2023).
292. Rnacentral 2021: Secondary structure integration, improved sequence search and new member databases. *Nucleic acids research* **49**, D212–D220 (2021).
293. Szklarczyk, D. *et al.* String v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research* **43**, D447–D452 (2015).
294. Vasilevsky, N. A. *et al.* Mondo: Unifying diseases for the world, by the world. *MedRxiv* 2022–04 (2022).
295. Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic acids research* **32**, D267–D270 (2004).
296. Gaulton, A. *et al.* The chembl database in 2017. *Nucleic acids research* **45**, D945–D954 (2017).
297. Goodman, J. Computer software review: Reaxys (2009).
298. Tingle, B. I. *et al.* Zinc-22— a free multi-billion-scale database of tangible compounds for ligand discovery. *Journal of chemical information and modeling* **63**, 1166–1176 (2023).
299. Kirklin, S. *et al.* The open quantum materials database (oqmd): Assessing the accuracy of dft formation energies. *npj Computational Materials* **1**, 1–15 (2015).
300. Zhang, N. *et al.* DeepKE: A deep learning based knowledge extraction toolkit for knowledge base population. *arXiv:2201.03335* (2022).
301. Luo, Y. *et al.* OneKE: A dockerized schema-guided LLM Agent-based knowledge extraction system. In *Companion Proceedings of the ACM on Web Conference 2025*, 2871–2874 (2025).
302. Zhu, Y. *et al.* LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web* **27**, 58 (2024).
303. Neo4j. Neo4j official website (2025). Software available at <https://neo4j.com>.

304. Janusgraph. Janusgraph official website (2024). Software available at <https://github.com/JanusGraph/janusgraph>.
305. ArangoDB. ArangoDB official website (2025). Software available at <https://arangodb.com>.
306. Virtuoso. Virtuoso official website (2025). Software available at <https://vos.openlinksw.com/owiki/wiki/VOS>.
307. Deutsch, A., Xu, Y., Wu, M. & Lee, V. Tigergraph: A native mpp graph database. *arXiv:1901.08248* (2019).
308. Han, X. *et al.* OpenKE: An open toolkit for knowledge embedding. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, 139–144 (2018).
309. Zheng, D. *et al.* DGL-KE: Training knowledge graph embeddings at scale. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 739–748 (2020).
310. Ali, M. *et al.* PyKEEN 1.0: A python library for training and evaluating knowledge graph embeddings. *J. Mach. Learn. Res.* **22**, 1–6 (2021).
311. Costabello, L. *et al.* AmpliGraph: A library for representation learning on knowledge graphs. *Retrieved Oct 10*, 2019 (2019).
312. Broscheit, S., Ruffinelli, D., Kochsiek, A., Betz, P. & Gemulla, R. LibKGE-A knowledge graph embedding library for reproducible research. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 165–174 (2020).
313. Yu, S.-Y., Chhetri, S. R., Canedo, A., Goyal, P. & Al Faruque, M. A. Pykg2vec: A python library for knowledge graph embedding. *J. Mach. Learn. Res.* **22**, 1–6 (2021).
314. Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y. & Liang, X. Doccano: Text annotation tool for human (2018). Software available from <https://github.com/doccano/doccano>.
315. Tkachenko, M., Malyuk, M., Holmanyuk, A. & Liubimov, N. Label Studio: Data labeling software (2020). Open source software available from <https://github.com/HumanSignal/label-studio>.
316. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, vol. 3, 361–362 (2009).
317. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504 (2003).
318. Shenoy, V. GraphGPT — build complex directed graphs from natural language (2025).
319. Liu, J. LlamaIndex (2022). Software available at https://github.com/jerryjliu/llama_index.
320. Pelletier, A. R. *et al.* A knowledge graph approach to elucidate the role of organellar pathways in disease via biomedical reports. *J. Vis. Exp.* (2023).
321. Aggour, K. S. *et al.* Compound knowledge graph-enabled ai assistant for accelerated materials discovery. *Integrating Mater. Manuf. Innov.* **11**, 467–478 (2022).

A Supplementary Information

Table S1. Commonly Used Databases for Scientific Knowledge Graph Construction

Domain	Database	Short description	Statistics	Update Frequency
Drug Databases	BindingDB ²⁸³	Publicly accessible collection of measured drug-target binding affinities.	3.1M binding data for 1.3M compounds & 9.6K targets.	Weekly
	DrugBank ¹⁵⁹	Richly annotated resource combining drug data with target, pathway & pharmacogenomic info.	18K approved & investigational drugs, 23K drug-target links, 3.6K drug-transporter links, 6K drug-enzyme links.	Monthly
	CTD ²⁸⁴	The comparative toxicogenomics database links chemicals, genes, phenotypes and diseases.	101M toxicogenomic interactions, 19K chemicals, 57K genes, 7K diseases.	–
	DisGeNET ²⁸⁵	Comprehensive platform integrating genes, variants, and human diseases, combining curated data and text-mined evidence.	2.0M gene–disease associations, 4.4M variant–disease associations, and 28M disease–disease associations	–
	DrugCentral ²⁸⁶	Authoritative, open-access compendium of active pharmaceutical ingredients approved worldwide.	5K drugs, 152K pharmaceutical products.	–
	PharmGKB ²⁸⁷	Provide PGx data from literature annotations to genotype-based treatment guidelines.	209 clinical guideline annotations, 1.2K drug label annotations, 483 FDA drug label annotations	–
	SIDER ²⁸⁸	Database of marketed drugs and their recorded adverse drug reactions (ADRs).	1.4K drugs, 6K side effects, 140K drug–side effect pairs.	Static
Omics Databases	Uniprot ⁸⁶	Comprehensive, high-quality protein sequence & functional annotation database.	573K reviewed entries, 253K unreviewed entries.	4 Weeks
	Ensembl ²⁸⁹	Genome browser & annotation resource for vertebrates and selected eukaryotes.	300+ species, 40K coding genes (human), 1M variants.	3 Months
	KEGG ¹⁵⁵	Database integrating pathways, genes, compounds, drugs and diseases for system analysis.	75K pathways, 54 M genes, 12K drugs, 11K diseases.	Daily
	Reactome ²⁹⁰	Curated, peer-reviewed pathway database emphasizing human biology.	2.8K human pathways covering 11.6K proteins, 16K reactions.	Monthly
	InterPro ²⁹¹	Comprehensive resource integrating multiple protein signature databases.	13 member databases covering millions of protein sequences.	Quarterly
	RNAcentral ²⁹²	Comprehensive ncRNA sequence collection representing all ncRNA types across diverse organisms.	44.5M non-coding RNA sequences, covering 1.1K species from 54 databases.	Twice a year
	STRING ²⁹³	Database of known and predicted protein–protein interactions across multiple organisms.	59.3M proteins, 20B PPIs, 12.5K organisms	–
	MONDO ^{†294}	Ontology harmonizing disease concepts with standardized identifiers, mappings, and classifications for clinical use.	17 disease resources integrated into 22K unified disease concepts.	Monthly
Chemical Database	UMLS ²⁹⁵	Comprehensive biomedical ontology integrating multiple vocabularies to unify concepts, names, and relationships.	17M names, 3.4M concepts, 8.7M codes, 190 vocabularies, 29 languages	Twice a year
	ChEBI ⁹⁰	Chemical entities of biological interest, a dictionary and ontology of small molecular entities.	62K compounds.	Monthly
	ChEMBL ²⁹⁶	A curated database of drug-like bioactive molecules that integrates chemical, bioactivity and genomic data to support drug discovery.	2.5M compounds, 1.7M assays, 15.5K drugs, 48.8K drug indications.	–
	Reaxys ²⁹⁷	Elsevier-curated chemical reactions, substances, properties & literature.	283M chemical substances, 73M reactions, 500M physicochemical data points.	–
	PubChem ⁸⁵	NIH repository of chemical substances, bioactivities & patents.	122M compounds, 338M substances, 297 M bioactivities.	Daily
Materials Databases	ZINC ²⁹⁸	Free database of commercially available compounds for virtual screening.	980M purchasable compounds.	–
	OQMD ²⁹⁹	Open-access database of DFT-calculated properties for inorganic and hybrid materials.	1.2M materials.	–
	Materials Project ⁸⁷	High-throughput DFT database of materials properties & crystal structures.	144K inorganic compounds, 76K bandstructures, 64K molecules, 530K nanoporous materials, and diverse tensors and electrodes.	–

Table S2. Commonly Used Software and Tools for Scientific Knowledge Graph Construction and Management

Category	Software Name	Short Description	Supported Tasks	License
Automated Construction	KG DeepKE ³⁰⁰	A knowledge extraction toolkit for knowledge graph construction supporting cnSchema, low-resource, document-level and multimodal scenarios.	Named Entity Recognition, Relation Extraction, Attribute Extraction, etc.	MIT License
	OneKE ³⁰¹	A flexible dockerized system for schema-guided knowledge extraction, capable of extracting information from web and PDF books.	Named Entity Recognition, Web News Extraction, Book Knowledge Extraction, etc.	MIT License
	AutoKG ³⁰²	An LLM-powered multi-agent framework for automated KG construction and reasoning, integrating external knowledge sources.	Entity/Relation Extraction, KG Construction, KG Reasoning, etc.	MIT License
Graph Databases and Storage	Neo4j ³⁰³	A widely used native graph database with ACID transactions and Cypher query language, suitable for highly connected data analysis.	Graph Storage, Graph Querying, Graph Algorithms, etc.	GPLv3
	JanusGraph ³⁰⁴	A highly scalable graph database optimized for storing and querying large graphs with billions of vertices and edges.	Graph Storage, Gremlin Query, etc.	CC-BY-4.0
	ArangoDB ³⁰⁵	A scalable graph database system with native graphs, integrated search engine, and JSON support via single query language.	Multi-Model Storage, Graph Traversal, Path Querying, etc.	BSL 1.1
	Virtuoso ³⁰⁶	A hybrid relational-RDF database supporting both SPARQL and SQL, widely used for Linked Data publishing.	RDF Storage, SPARQL Query, Ontology Reasoning, etc.	GPL v2
	TigerGraph ³⁰⁷	A commercial distributed parallel graph database optimized for real-time graph analytics with GSQL.	Graph Storage, Parallel Graph Computation, Real-time Querying, etc.	Proprietary
Representation Learning & Reasoning	OpenKE ³⁰⁸	A sub-project of OpenSKL, providing an Open-source Knowledge Embedding toolkit for knowledge representation learning.	KG Embedding, Link Prediction, Triple Classification, etc.	MIT License
	DGL-KE ³⁰⁹	A high performance, easy-to-use, and scalable package for learning large-scale knowledge graph embeddings.	KG Embedding, Large-scale Link Prediction, etc.	Apache 2.0
	PyKEEN ³¹⁰	A Python library for KG embeddings with modular design, automated hyperparameter tuning, and reproducibility guarantees.	KG Embedding, Model Evaluation, Hyperparameter Optimization, etc.	MIT License
	AmpliGraph ³¹¹	A suite of neural machine learning models for relational Learning on knowledge graphs with supervised learning.	Generate KG embeddings, Link Prediction, Anomaly Detection, etc.	Apache 2.0
	LibKGE ³¹²	A PyTorch-based library for efficient training, evaluation, and hyperparameter optimization of knowledge graph embeddings.	Link Prediction, Training, Evaluation of KGE Models, etc.	MIT License
	Pykg2vec ³¹³	A library for learning the representation of entities and relations in Knowledge Graph with various embedding models.	KGE Model Implementations, Hyperparameters Discovery, Learned Embedding Inspecting, etc.	MIT License
Auxiliary Tools	Doccano ³¹⁴	An open-source text annotation tool with a web interface for humans to label text data efficiently.	Annotation for Text Classification, Sequence Labeling, Sequence to Sequence tasks, etc.	MIT License
	Label Studio ³¹⁵	An open source data labeling tool supporting multi-modal data including text, images, audio, video, time series.	Multi-modal Data Annotation, Quality Assurance, etc.	Apache 2.0
	Gephi ³¹⁶	An award-winning open-source platform for visualizing and manipulating large graphs with interactive exploration.	Graph Visualization, Network Analysis, Community Detection, etc.	CDDL 1.0
	Cytoscape ³¹⁷	A network visualization platform originally designed for bioinformatics, now supporting general-purpose network analysis.	Graph Visualization, Attribute Integration, Topology Analysis, etc.	LGPL
	GraphGPT ³¹⁸	An experimental tool using GPT models to extract entities and relations from text and generate interactive KG visualizations.	Triple Extraction, KG Construction, Visualization, etc.	MIT License
	LlamaIndex ³¹⁹	A component for building KG indices from unstructured text, integrating triples into LLM-based retrieval pipelines.	Triple Extraction, KG Indexing, KG-based QA, etc.	MIT License

Table S3. Representative SciKGs for Drug Development and Optimization

Application	Year	Publication	KG Used	Entity Types	Relation Types	Entity/Relation Counts
Drug Repurposing	2023	Bang et al. ¹⁴²	RepoMultiKG ¹⁴²	drug, disease, gene, GO_MF, etc.	drug-disease, drug-target, drug-gene, etc.	–
	2024	Huang et al. ¹⁴⁴	Medical KG ¹⁴⁴	drug, protein, phenotype, etc.	drug-drug, disease-disease, drug-protein, etc.	124K/8.06M
	2024	Huang et al. ¹⁴⁴	Healx KG ¹⁴⁴	compound, disease, gene, etc.	compound-treats-disease, gene-associates-disease, pathway-targets-gene, etc.	–
	2025	Zhang et al. ²⁴	iKgraph ²⁴	disease, gene/protein, chemical compound, etc.	chemical-gene, chemical-disease, disease-gene, etc.	10.7M/30.8M
	2025	Madeddu et al. ¹⁵⁶	VITAGRAPH ¹⁵⁶	compound, disease, gene, side effect, etc.	compound-compound, gene-gene, compound-disease, etc.	48K/4.00M
DDI Prediction	2020	Lin et al. ¹⁴⁵	Drug KG ¹⁴⁵	drug	drug-drug interaction	24K/669K
	2024	Wang et al. ¹⁴⁷	KnowDDI ¹⁴⁷	drug, protein, disease, etc.	drug-drug, protein-protein, drug-target, etc.	–
	2024	Xu et al. ⁸⁴	iBKH ⁸⁴	drug, protein, ATC/category, etc.	drug-target-gene, drug-transporter-gene, drug-association-pathway, etc.	129K/4.03M
	2024	Wu et al. ¹⁴⁶	MMDDI-KG ¹⁴⁶	drug, chemical, transporter, etc.	drug-chemical, drug-substructure, drug-drug interaction, etc.	–
DTI Prediction	2017	Luo et al. ¹⁴⁸	LTN ¹⁴⁸	drug, protein, disease, etc.	drug-protein, drug-drug, drug-disease, etc.	12K/1.90M
	2021	Ye et al. ²¹	UniDTI-KG ²¹	drug, protein, disease, etc.	drug-protein, drug-drug, drug-disease, etc., treatment, regulates, etc.	–
	2022	Ma et al. ¹⁷²	DRKG ¹⁷²	drug, protein, disease, etc.	target-disease, etc.	97K/5.87M
	2024	Feng et al. ¹⁷³	e-TSN KG ¹⁷³	drug, disease, target, etc.	target_of, interacts_with, protein-protein, etc.	10.4M/315.8M
Virtual Screening	2019	Shang et al. ¹⁵¹	Gamenet ¹⁵¹	patient, clinical event, diagnosis, etc.	medication combination, drug-drug interaction, etc.	–
	2021	Bhoi et al. ¹⁵³	PerMedRec KG ¹⁵³	drug, diagnose, procedure, etc.	drug co-occurrence, drug-drug interaction, etc.	–
Drug Toxicity	2018	Zimik et al. ¹²¹	Decagon KG ¹²¹	drug, protein, etc.	–	20K/5.39M
	2023	Evangelista et al. ¹⁵²	ReproTox-KG ¹⁵²	birth defect, gene, drug, etc.	birth defect-gene association, birth defect-drug association, drug-gene association, etc.	24K/580K

Table S4. Representative SciKGs for Omics Interpretation and Analysis

Application	Year	Publication	KG Used	Entity Types	Relation Types	Entity/Relation Counts
Genomics	2022	Feng et al. ¹⁷⁸	GenomicKB ¹⁷⁸	chromosome chain, coding element, non-coding element, etc.	position, regulation, expression, etc.	347M/1.36B
	2024	Mulero et al. ¹⁸⁰	GenoRegKG ¹⁸⁰	enhancer, transcription factor, disease/phenotype, TAD, gene	enhancer-gene, enhancer-transcription factor, etc.	–
	2025	Zaripova et al. ¹⁸¹	PhenoKG ¹⁸¹	phenotype, molecular function, cellular component, etc.	protein-protein, disease-phenotype, phenotype-phenotype, etc.	105K/1.10M
Proteomics	2022	Santos et al. ¹⁹⁰	CKG ¹⁹⁰	protein, disease, metabolite, etc.	HAS-PARENT, HAS_QUANTIFIED_PROTEIN, etc.	20M/220M
	2022	Binder et al. ¹⁹¹	PKG ¹⁹¹	phenotype, endogenous ligand-drug, gene, etc.	expression, association, gene signatures, etc.	–
	2022	Zhang et al. ¹⁹²	ProteinKG25 ¹⁹²	gene Ontology term (molecular function, cellular component, etc.), protein sequence	protein-GO, GO-GO	612K/5.0M
	2023	Chen et al. ¹⁹⁶	TransPKG ¹⁹⁶	transporter, gene, drug, etc.	transporter-gene, transporter-drug, drug-gene, etc.	20K/528K
	2023	Pelletier et al. ³²⁰	CaseOLAP ³²⁰	protein, disease, pathway, etc.	protein-disease association, protein-protein interaction, protein-pathway association, etc.	27.2K/540.6K
	2024	Cheng et al. ¹⁹³	ProteinKG65 ¹⁹³	protein, GO term (molecular function, cellular component, biological process)	Protein-GO	570K/5.5M
Transcriptomics	2024	Ni et al. ¹⁹⁹	Biomedical KG ¹⁹⁹	protein, RBP, TF, etc.	PPI, RBP_regulates, miRNA_regulates, etc.	79K/6.79M
	2022	Shao et al. ²⁰⁰	LRT-KG ²⁰⁰	receptor, pathway, target gene, etc.	ligand-receptor interactions (LRIs), receptor activation of transcription factors (TFs), transcription factor regulation of target genes, etc.	–
	2024	Cavalleri et al. ²⁰¹	RNA-KG ²⁰¹	RNA molecule, cellular component, Ontology term, etc.	molecular interaction, regulation, association with disease/phenotype, etc.	674K/12.7M
Metabolomics	2021	Delmas et al. ²⁰²	FORUM ²⁰²	chemical compound & class, MeSH terms (disease, anatomy, etc.), publication	–	–
Microbiomics	2023	Sun et al. ²⁰⁵	MMiKG ²⁰⁵	microbiota, intermediate, disease, etc.	promote, inhibit, associated, etc.	770/1.26K
Multi-omics	2020	Jha et al. ²⁰⁹	Kirchhoff's KG ²⁰⁹	target (gene/protein), disease, drug, etc.	gene-gene interaction, gene-drug interaction, gene-disease association, etc.	~459.8M
	2022	Di et al. ²¹⁰	BioTAGME KG ²¹⁰	gene, protein, disease, etc.	literature, STRING, BioTAGME, etc.	161K/40M

Table S5. Representative SciKGs for Chemical Reaction and Synthesis

Application	Year	Publication	KG Used	Entity Types	Relation Types	Entity/Relation Counts
Chemical Reaction Prediction	2016	Segler et al. ²¹³	MolReactKG ²¹³	molecule, reaction, etc.	reactant, reagent, catalyst, etc.	22.6M/-
	2021	McDermott et al. ²¹⁴	ChemReactKG ²¹⁴	chemical phase, combinations of phases	chemical reaction	Varies by system: C-Cl-Li-Mn-O-Y: 5.9K/121K, C-Cl-Mn-Na-O-Y: 4.4K/46K, Fe-S-Si: 0.5K/12K, Ba-Cu-O-Y: 3.0K/34K
	2023	Zhou et al. ²¹⁸	ChemQAKG ²¹⁸	chemical species, chemical reaction, etc.	chemical reaction, chemical property, etc.	-
	2024	Xie et al. ²¹⁵	ChemSynKG ²¹⁵	reactant, product	reaction templates	820K/587K
	2025	O’Ryan et al. ²¹⁶	OrgSynKG ²¹⁶	entity, chemical, temperature, etc.	other compound of, work up, moiety of, etc.	19.7M/22.1M
Chemical Synthesis Path Optimization	2021	Jeong et al. ²²	Reaction KG ²²	compound, reaction, etc.	compound-reaction, reactant-reaction, catalyst/solvents-reaction, etc.	1.67M/-
	2022	Li et al. ²¹⁹	SA Reaction KG ²¹⁹	reactant, product, etc.	chemical reaction	2.19M/9.04M
	2022	Deagen et al. ²²⁰	FAIRIntKG ²²⁰	keyword, country/institution, research topic/cluster, etc.	co-citation relationships, co-occurrence relationships, etc.	8.9K/25.8K
	2025	Ma et al. ²⁵	MacroRetroSynKG ²⁵	reactant, product, intermediate, etc.	reaction condition, numerical identifier, yield, etc.	3.1K/292
Molecular Property Prediction	2022	Fang et al. ²²¹	ChemElemKG ²²¹	element, attribute, etc.	element-attribute	225/1.64K
	2022	Zhang et al. ²²²	MolStruct-KG ²²²	compound, drug, etc.	compound reaction, drug-drug interaction, etc.	-
	2023	Fang et al. ¹¹	ElementKG ¹¹	chemical element, functional group, etc.	inRadiusGroup1, has StateGas isPartof, etc.	201/52.1K
	2025	Jiang et al. ²²³	MolKG ²²³	molecule, gene/protein, disease, etc.	rotatable_bond_count, covalent_unit_count, hydrogen_bond_donor_count, etc.	185K/2.52M

Table S6. Representative SciKGs for Materials Design and Discovery

Application	Year	Publication	KG Used	Entity Types	Relation Types	Entity/Relation Counts
New Material Design	2018	Onishi et al. ²²⁴	PSPP KG ²²⁴	process, structure, property, etc.	binary relationship (positive/negative) between factors	2K/104
	2020	Mecusker et al. ²²⁵	NanoMine ²²⁵	property, material, processing method, etc.	in relation to, has value, has attribute, etc.	–
	2022	Nie et al. ²²⁶	LiB Cathodes KG ²²⁶	material	material-material	–
	2022	Aggour et al. ³²¹	CKG ³²¹	material, property, processing, microstructure, etc.	processing-microstructure, property-material, processing-material, etc.	–
	2023	Gao et al. ²²⁷	Cu-Based Catalysts KG ²²⁷	material, regulation method, product, etc.	–	–
	2024	Venugopal et al. ²²⁸	MatKG ²²⁸	material, symmetry phase label, synthesis etc.	–	70K/5.4M
	2024	Ye et al. ²²⁹	MKG ²²⁹	material, formula, acronym, etc.	–	163K/732K
	2024	Ghafarollahi et al. ⁸³	OntoBioMatKG ⁸³	biomaterial, mechanical property, etc.	provide, possess, exhibited by, etc.	33K/49K
Material Performance Prediction	2020	Mrdjenovich et al. ²³¹	Propnet KG ²³¹	material property (band gap, density, elastic moduli), property relationship/ model, etc.	property as input to model, property as output from model, etc.	184/–
	2022	Shu et al. ²³⁸	Grain KG ²³⁸	grain, size attribute, orientation attribute.	grain-grain, grain-size attribute, grain-orientation.	77K/745.8K
	2023	Liu et al. ²³²	Aluminum Alloy Domain KG ²³²	alloy, series, material, etc.	ClassOf, subClassOf, PropertyOf, etc.	–/1.15K
	2023	Song et al. ²³³	NR-KG ²³³	HEA, element, processing techniques, etc.	is_structured_by, is_contained_by, is_processed, etc.	–
	2023	Statt et al. ²³⁴	MekG ²³⁴	analysis, analysisDetails, collection, etc.	analysis_details, contains, collection-sample, etc.	52.3M/111.4M
	2024	Durmaz et al. ²³⁵	MaterioMiner ²³⁵	physical quantities, materials, mechanisms, etc.	associatedwith, causeof, correlatedwith, etc.	2.19K/–
	2024	Huang et al. ²³⁶	Element KG ²³⁶	element, attribute, etc.	isperiodof, ismetallicityof, isstateof, etc.	–
	2024	Zhang et al. ²³⁹	MGED-KG ²³⁹	term, category, etc.	subclassof, isakindof, isrelatedto, etc.	8.9K/–
Material Screening & Optimization	2025	Bai et al. ²⁴³	KG-FM ²⁴³	property, structure, application, etc.	derived from, published in (journal), published at (date), etc.	2.53M/4.01M