

ORIGINAL RESEARCH

Open access

Knowledge Graphs for Materials Discovery: Data Structuring, Reasoning, and Applications

Maria Hernandez^{1*}, Carlos Vega¹

Abstract

Knowledge graphs (KGs) have emerged as a pivotal infrastructure in computational and data-driven materials engineering, enabling structured representation, reasoning, and integration of heterogeneous data for accelerated discovery. By organizing materials data into interconnected entities and relationships, KGs facilitate advanced querying, inference, and machine learning applications across domains such as materials informatics, high-throughput computation, and inverse design. This review synthesizes recent advancements in KG construction from multimodal datasets, including text corpora, biomolecular integrations, and crystalline structures. We examine how graph neural networks and representation learning enhance molecular contrastive learning and pre-training frameworks for improved molecular representations. In the landscape of computational materials ecosystems, KGs support semantic integration and terminology standardization, bridging simulation and experiment through active learning systems and uncertainty quantification. Applications in autonomous laboratories highlight closed-loop discovery, where KGs enable dynamic knowledge propagation and event-sourced provenance management. We provide an original synthesis framing KGs as unifying backbones for data-model-experiment cycles, emphasizing systems-level integration over isolated tools. Challenges in scalability and interoperability are noted, with future directions toward hybrid human-AI workflows. This narrative underscores KGs' role in transforming materials discovery from empirical to predictive paradigms, fostering interdisciplinary convergence in materials science.

Keywords Materials informatics, Graph neural networks, Representation learning, Autonomous laboratories, High-throughput computation, Knowledge graphs

*Correspondence:

Maria Hernandez
maria.hernandez@outlook.com

¹ Department of Computational Materials Science, Faculty of Engineering, Polytechnic University of Valencia, Valencia, Spain

Introduction

The field of materials engineering has undergone a profound transformation with the advent of computational and data-driven approaches, shifting from traditional trial-and-error methodologies to systematic, predictive frameworks. This evolution is driven by the exponential growth in materials data from high-throughput computations, experimental characterizations, and simulations, necessitating advanced tools for data structuring and utilization. Knowledge graphs (KGs) stand out as a versatile paradigm for organizing this vast, heterogeneous information into semantically rich networks of entities and relations, enabling sophisticated reasoning and discovery processes [1-4].

At the core of computational materials engineering lies materials informatics, which leverages machine learning (ML) and statistical methods to extract insights from data repositories. KGs enhance this by providing a structured ontology that captures not only atomic-level properties but also higher-order relationships, such as synthesis pathways, property correlations, and performance metrics [5-8]. For instance, autonomously generated KGs from text corpora have

demonstrated the ability to standardize materials terminology, facilitating cross-database interoperability [1, 5]. Similarly, in biomolecular contexts, KGs integrate diverse data types to support precision medicine analogies in materials design [6, 9].

Graph neural networks (GNNs) and representation learning further amplify KG utility by embedding complex graph structures into vector spaces suitable for ML tasks. Recent developments in knowledge graph-enhanced molecular contrastive learning have improved functional prompting for property prediction [10], while knowledge-guided pre-training frameworks address data scarcity in molecular representations [11]. These techniques are particularly potent in inverse materials design, where the goal is to engineer materials with targeted properties rather than forward prediction [12, 13].

High-throughput computation, a cornerstone of data-driven materials, generates vast datasets from density functional theory (DFT) and molecular dynamics simulations. KGs serve as integrators, linking computational outputs with experimental validations through multimodal datasets [14-16]. For example, infrastructures for managing materials data and services employ user-centered models to unify simulation and experiment [14], while ontology-based KGs handle RNA interactions with implications for bio-inspired materials [17].

The integration of simulation and experiment is exemplified in autonomous laboratories, where robotic systems conduct closed-loop experiments guided by ML models. KGs play a critical role in maintaining provenance and enabling reasoning over experimental workflows [2, 4, 18]. Event-sourced architectures track materials experiments, ensuring reproducibility and knowledge propagation [4, 18]. Moreover, dynamic KG approaches distribute self-driving lab operations, coordinating multiple agents for efficient discovery [2].

Active learning systems, incorporating uncertainty quantification, optimize exploration in vast materials spaces. By balancing exploitation of known regions with exploration of uncertainties, these systems accelerate identification of novel materials [12, 16, 19]. KGs enhance this by providing relational contexts for uncertainty propagation, as seen in noise-impacted inverse design for NMR spectra matching [13].

Multimodal datasets, encompassing text, images, and structured data, pose unique challenges for integration. KGs address this through semantic enrichment, such as extracting materials information from corpora [7, 20, 21] or building chemical reaction graphs [22]. In crystalline materials, KG representations capture zeolitic structures for property inference [23], while in broader chemistry, LLMs transform materials science via hackathon-inspired integrations [16].

The convergence of these elements underscores the need for a cohesive framework in materials discovery. KGs not only structure data but also enable inferential reasoning, such as reinforcement learning-based KG reasoning for aluminum alloys [24] or cell-cell communication inference with spatial transcriptomics [25]. This relational paradigm shifts materials engineering toward systems-level thinking, where data flows seamlessly across computational pipelines.

Despite these advances, the field lacks a unified synthesis of KG applications in computational ecosystems. Existing reviews often focus on narrow subdomains, such as ML in materials [12] or specific ontologies [3, 17], without integrative cross-analysis. This review positions itself as a comprehensive narrative on KGs for materials discovery, emphasizing original structuring around data-model-reasoning-application cycles. By synthesizing 29 key publications, we provide a fresh interpretive lens on how KGs bridge informatics, computation, and automation, fostering accelerated innovation in materials engineering. The functional positioning of knowledge graphs across discovery stages is synthesized in Table 1.

Table 1. Functional Roles of Knowledge Graphs across the Materials Discovery Lifecycle

Discovery Stage	KG Functional Role	Enabling Technologies	Representative Applications	System-Level Impact
Data Structuring	Entity extraction, ontology formation,	NLP pipelines, MatSciBERT,	Literature-derived materials databases,	Reduces fragmentation,

	semantic harmonization	automated corpus mining	terminology standardization	enables structured querying
Knowledge Integration	Linking simulation, experiment, and literature data	Multimodal fusion frameworks, semantic integration engines	Structure–property correlation mapping	Unifies heterogeneous materials knowledge
Graph Reasoning	Relational inference and uncertainty propagation	GNNs, probabilistic KG reasoning, reinforcement learning	Alloy optimization, property inference	Enhances predictive modeling capacity
Representation Learning	Embedding materials knowledge into latent spaces	Contrastive learning, KG-guided pre-training	Molecular property prediction, inverse design priors	Improves generalization under data scarcity
Discovery Design	Candidate screening and design navigation	Active learning, KG-query optimization	Targeted materials discovery	Accelerates exploration efficiency
Autonomous Experimentation	Closed-loop experimental steering	Robotics, self-driving labs, workflow orchestration	Robotic synthesis, adaptive testing	Enables continuous discovery cycles
Provenance & Governance	Tracking experimental lineage and decisions	Event-sourced KGs, blockchain provenance	Reproducibility auditing	Ensures transparency and trust
Infrastructure Scaling	Distributed knowledge management	Federated learning, interoperable schemas	Multi-institution collaboration	Supports global discovery ecosystems

Landscape of Computational & Data-Driven Materials Engineering

Foundations of materials informatics and data structuring

Materials informatics represents the intersection of data science and materials engineering, where structured data repositories enable predictive modeling and discovery. Knowledge graphs (KGs) have become essential for structuring heterogeneous materials data, transforming raw inputs into queryable, relational networks [1, 3, 5, 14]. Autonomously generated KGs, such as MatKG, compile vast materials knowledge from scientific literature, capturing entities like compounds, properties, and processes [1]. Similarly, terminology KGs constructed from text corpora standardize nomenclature, addressing inconsistencies across databases [5]. These efforts underscore KGs' role in semantic integration, where diverse data sources—ranging from mechanical strengthening mechanisms to biomolecular interactions—are unified under common ontologies [3, 6].

In practice, KG construction often leverages natural language processing (NLP) for information extraction. Models like MatSciBERT, specialized for materials domain text mining, facilitate entity recognition and relation extraction from publications [7]. This is complemented by automatically generated corpora for materials information, enabling scalable KG population [20]. For specialized domains, ontology-based KGs represent interactions in RNA molecules or zeolitic crystals, providing templates for broader materials applications [17, 23]. Such structuring not only organizes data but also enables reasoning, such as inferring property relationships absent in raw datasets [8, 24].

Machine learning and representation learning in materials ecosystems

Machine learning integration with KGs enhances representation learning, crucial for handling complex materials spaces. Graph neural networks (GNNs) embed KG structures into latent spaces, improving tasks like property prediction and molecular design [10, 11, 15]. Knowledge graph-enhanced molecular contrastive learning uses functional prompts to align representations, boosting generalization across chemical spaces [10]. Knowledge-guided pre-training frameworks further refine molecular embeddings by incorporating domain-specific priors, mitigating data limitations [11].

Representation learning extends to multimodal contexts, where KGs fuse text, structural, and experimental data. For instance, chemical reaction KGs expand synthetic spaces by linking reactants, products, and conditions [22]. In materials tetrahedron reconstruction, NLP-driven extraction challenges highlight needs for integrated KG-NLP pipelines [21]. These advancements facilitate high-throughput screening, where KG-enriched ML models prioritize candidates for computation or synthesis [12, 16].

High-throughput computation and multimodal integration

High-throughput computation generates expansive datasets, demanding robust integration frameworks. KGs serve as backbones for linking DFT calculations, molecular simulations, and empirical measurements [8, 14, 15]. Propnet, a KG for materials science, interconnects properties via mathematical relations, enabling propagation and inference [8]. Infrastructures with user-centered data models manage services across computational workflows, ensuring accessibility and interoperability [14].

Multimodal datasets amplify this by incorporating diverse modalities. KGs for cell-cell communication infer spatial transcriptomics interactions, analogous to materials microstructure modeling [25]. In drug repurposing and natural products discovery, KGs integrate chemical, biological, and pharmacological data, offering paradigms for materials-property linkages [26, 27]. Sample-centric frameworks for natural products mirror materials discovery pipelines, where KG-driven computation accelerates hit identification [27].

Simulation-experiment integration benefits from KG-mediated bridges. Semantic data integration assesses phenomena like Orowan strengthening, validating computational predictions against experiments [3]. Petagraph unifies biomolecular and biomedical data, inspiring similar integrations in hybrid organic-inorganic materials [6]. These syntheses reveal KGs' capacity to harmonize disparate data streams, fostering comprehensive materials ecosystems.

Active learning and uncertainty quantification

Active learning systems optimize data acquisition in resource-constrained environments, with KGs providing relational contexts for decision-making [12, 13, 16, 19]. Uncertainty quantification guides exploration, as in noise-impacted inverse design for spectral matching [13]. KGs enhance this by encoding uncertainties in relations, enabling probabilistic reasoning over discovery loops [16].

In LLMs' application to materials, hackathon examples illustrate KG-augmented workflows for chemistry tasks, emphasizing active learning in data-scarce regimes [16]. Reinforcement learning on KGs reasons over alloy applications, balancing exploration and exploitation [24]. This subfield highlights KGs' evolution from static repositories to dynamic engines for adaptive discovery.

Systems-level perspectives and community needs

Broadening to systems integration, KGs address community needs for data management [28]. Laboratory of Babel analyses underscore requirements for unified platforms, where KGs mitigate fragmentation [28]. Integrating multiple projects via neural networks prefigures KG-centric unification [15]. Overall, this landscape synthesis frames KGs as

infrastructural enablers, synthesizing disparate threads into cohesive computational ecosystems for materials engineering.

Autonomous & closed-loop discovery systems

Autonomous laboratories represent the pinnacle of data-driven materials engineering, where closed-loop systems iteratively design, execute, and refine experiments guided by computational intelligence. Knowledge graphs (KGs) are instrumental in these setups, providing dynamic structures for knowledge propagation, decision-making, and provenance tracking [2, 4, 12, 18, 19].

Core architectures of self-driving laboratories

Self-driving labs distribute operations across agents, with KGs enabling real-time coordination [2]. Dynamic KG approaches integrate sensors, robots, and ML models, facilitating adaptive workflows [2]. Event-sourced architectures, like the Materials Experiment Knowledge Graph (MEKG) and ESAMP, maintain provenance by logging events as immutable facts, allowing reconstruction of discovery paths [4, 18]. This ensures reproducibility, critical for validating autonomous outcomes [18].

Integration of AI, high-performance computing, and robotics accelerates discovery cycles [12]. KGs unify these components, embedding experimental feedback into computational models for iterative refinement [12, 19]. For instance, autonomy integration into research platforms automates hypothesis testing, with KGs handling data flows [19].

Closed-loop mechanisms and active learning integration

Closed-loop discovery formalizes data-model-experiment cycles, where KGs support reasoning over iterations. A conceptual framework can be expressed as:

$$\begin{aligned} D_{\{t+1\}} \\ = f(M_t, E_t, KG) \end{aligned} \quad (1)$$

where $D_{\{t+1\}}$ is the next design proposal, M_t the current model, E_t experimental outcomes, and KG the knowledge graph updating relations [2, 12, 16]. This interpretive formula highlights KGs' role in fusing multimodal inputs for optimized loops.

Active learning enhances closures by selecting informative experiments, incorporating uncertainty [13, 16]. In KG-enriched systems, relational queries prioritize candidates, as in LLM-transformed workflows [16]. Inverse design under noise exemplifies KG-guided balancing of exploration and exploitation [13].

Applications in materials synthesis and characterization

In synthesis, KGs drive robotic platforms for chemical reactions, expanding spaces via reaction graphs [22]. For crystalline materials, KG representations inform autonomous zeolite design [23]. Biomolecular analogies, like RNA interaction KGs, inspire bio-materials automation [17].

Characterization benefits from KG-integrated spatial inference, akin to cell communication graphs [25]. Drug repurposing KGs parallel materials optimization, where autonomous loops refine property targets [26]. Knowledge graphs function as integrative infrastructures linking multimodal data, reasoning engines, and autonomous experimentation into unified discovery systems (Figure 1).

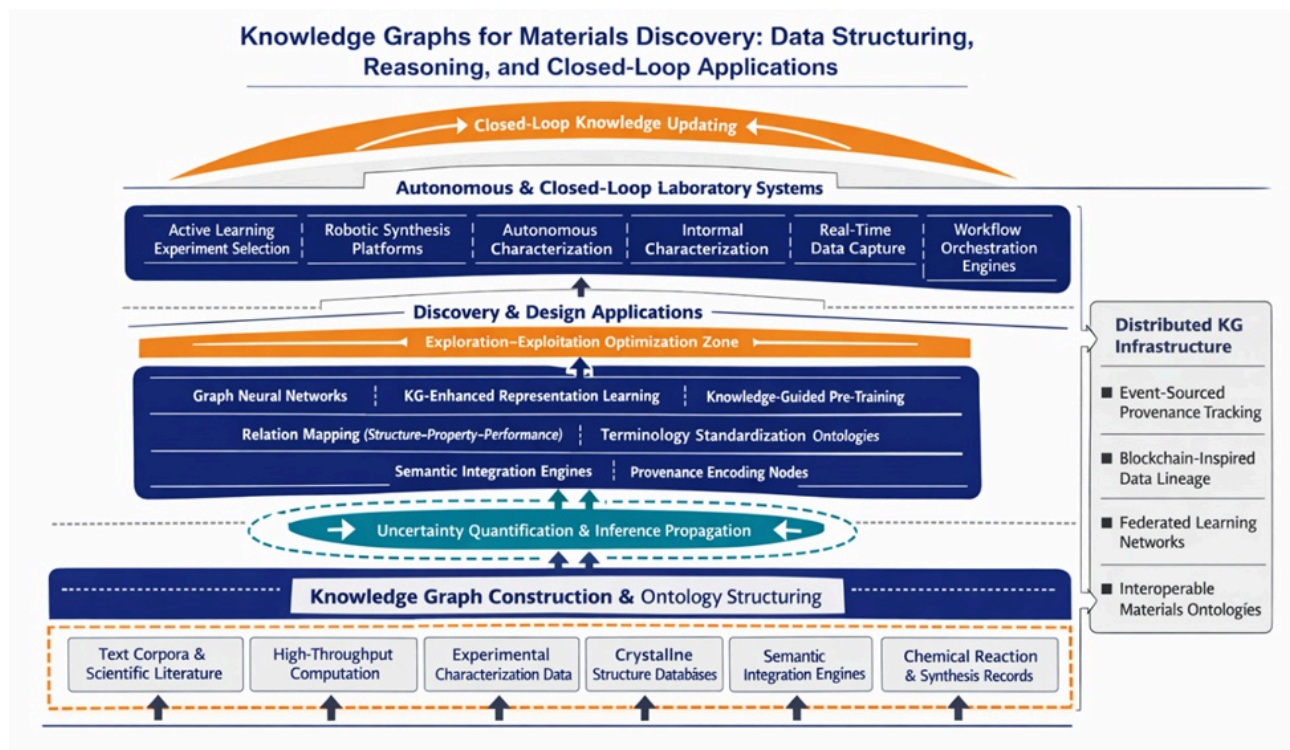


Figure 1. Knowledge Graphs for Materials Discovery: An Integrated Framework for Data Structuring, Reasoning, and Closed-Loop Learning

Figure 1 Conceptual systems architecture illustrating the role of knowledge graphs as integrative backbones in computational materials discovery. The framework depicts the progression from multimodal data ingestion through ontology structuring and graph-based reasoning to inverse design and autonomous laboratory execution. Closed-loop feedback mechanisms continuously update the knowledge graph, enabling adaptive learning across discovery cycles. Distributed infrastructure layers support provenance tracking, interoperability, and federated collaboration, positioning KGs as dynamic orchestrators of data–model–experiment ecosystems.

Broader implications for systems integration

Community frameworks, like OREGANO for repurposing, extend to materials via KG ecosystems [9, 26]. Life sciences KGs provide open-source templates for materials [29]. Challenges in scaling are mitigated by distributed designs [2, 28].

This synthesis positions KGs as orchestrators in autonomous systems, enabling seamless transitions from computation to realization in materials discovery.

Challenges & limitations

The integration of knowledge graphs (KGs) into computational and data-driven materials engineering, while promising, encounters several challenges that impede widespread adoption and scalability. These limitations span data quality, interoperability, computational demands, and methodological gaps, requiring targeted advancements to realize KGs' full potential [1–4, 7, 16, 28].

Data heterogeneity and quality issues

A primary challenge lies in the heterogeneity of materials data sources, which include unstructured text, structured databases, simulations, and experimental outputs. Extracting reliable entities and relations from text corpora is fraught

with ambiguities, such as varying terminologies for the same material property [5, 7, 20, 21]. MatSciBERT and similar domain-specific language models mitigate this through fine-tuned extraction, but noise in corpora can propagate errors into KGs, affecting downstream reasoning [7, 20, 21]. For instance, reconstructing the materials tetrahedron reveals inconsistencies in property linkages, underscoring the need for robust validation mechanisms [21].

In multimodal datasets, integrating diverse modalities exacerbates quality issues. Biomolecular KGs like Petagraph handle large-scale data but struggle with incomplete or conflicting entries from disparate biomedical sources [6]. Similarly, in crystalline materials, KG representations for zeolites must contend with structural variabilities that are not fully captured in existing ontologies [23]. Uncertainty quantification becomes critical here, as noisy inputs in inverse design workflows can lead to suboptimal material proposals [13].

Interoperability and standardization barriers

Interoperability across KG ecosystems remains a significant limitation, with fragmented standards hindering data exchange. While ontologies like OREGANO provide frameworks for drug repurposing, adapting them to materials requires custom extensions, often leading to siloed systems [17, 26, 29]. Community needs assessments highlight the "laboratory of Babel" problem, where disparate data management practices prevent seamless integration [28]. User-centered infrastructures attempt to address this, but lack of universal schemas limits their efficacy [14].

Semantic integration efforts, such as assessing Orowan strengthening, demonstrate partial successes but reveal gaps in linking computational predictions to experimental realities [3]. In life sciences KGs, open-source ecosystems offer interoperability templates, yet materials-specific adaptations face challenges in aligning with existing databases like propnet [8, 29].

Computational scalability and resource demands

Scalability poses another key challenge, particularly for large-scale KGs in high-throughput contexts. Autonomously generating KGs from vast corpora demands significant computational resources, with processing times scaling nonlinearly with data volume [1, 5]. Graph neural networks for representation learning, while effective, require extensive training on specialized hardware, limiting accessibility [10, 11, 15].

In autonomous systems, dynamic KGs for self-driving labs must handle real-time updates, but event-sourced architectures like ESAMP can become bottlenecked by provenance overhead [4, 18]. Distributed approaches alleviate this, yet coordinating multiple agents introduces latency and synchronization issues [2]. LLM integrations in materials science further amplify demands, as hackathon examples show resource-intensive fine-tuning for domain tasks [16].

Methodological and reasoning limitations

Methodological gaps in KG reasoning affect applications like reinforcement learning for alloys, where incomplete graphs lead to biased inferences [24]. Knowledge-guided pre-training helps, but over-reliance on functional prompts can overlook rare relations [10, 11]. In closed-loop systems, active learning struggles with high-dimensional spaces, where uncertainty quantification fails to capture epistemic uncertainties fully [12, 13, 19].

Broader limitations include the integration of spatial and temporal data, as seen in cell-cell communication KGs, which require extensions for materials microstructure dynamics [25]. Chemical reaction KGs expand spaces but often neglect kinetic aspects, limiting predictive accuracy [22]. Sample-centric frameworks for natural products discovery illustrate similar issues in handling complex interactions [27].

Addressing these challenges necessitates hybrid approaches, combining KGs with advanced ML to enhance robustness and efficiency. This synthesis reveals that while KGs provide powerful structuring, their limitations in handling real-world complexities demand ongoing refinement.

Results and Discussion

Advancing knowledge graphs (KGs) in computational materials engineering requires a shift from viewing them as passive data repositories toward recognizing them as active epistemic infrastructures embedded within discovery systems. Future work must therefore integrate advances in human–AI collaboration, reasoning formalisms, multimodal learning, distributed infrastructures, and ethical governance. These trajectories collectively position KGs as adaptive, reflexive systems capable of steering materials innovation under conditions of uncertainty, scale, and societal responsibility [2, 4, 6, 9, 12, 16, 28, 29].

Hybrid human–AI collaborative frameworks

A central frontier lies in the development of hybrid intelligence systems in which human expertise and KG-driven automation operate in reciprocal feedback loops rather than sequential workflows. Although KG-enabled pipelines have demonstrated capacity in literature extraction, materials screening, and hypothesis linkage, they remain constrained by the epistemic limits of codified knowledge. Human experts continue to provide tacit reasoning, contextual interpretation, and mechanistic skepticism that resist formal graph encoding.

Future architectures should therefore move toward adaptive ontological systems capable of incorporating real-time expert feedback. Such systems would allow specialists to refine edge semantics, recalibrate confidence scores, and annotate contested relationships. Over time, this would produce living ontologies that evolve alongside disciplinary knowledge rather than lag behind it. Experimental insights, including negative or null results often absent from published corpora, could be embedded into graph layers, mitigating publication bias.

Early signals of this paradigm are visible in LLM-enabled collaborative environments where AI systems assist in materials reasoning tasks. However, these environments remain episodic. A more mature trajectory would embed persistent conversational agents within KG interfaces, enabling continuous dialogue between researchers and graph infrastructures [16]. Lessons from open-source life sciences ecosystems demonstrate that community-curated ontologies can accelerate schema evolution and semantic harmonization [29]. Translating this model into materials engineering could foster federated expert stewardship across subdomains such as alloys, polymers, and quantum materials.

In autonomous laboratories, hybrid oversight becomes a governance necessity. Fully automated closed-loop experimentation risks propagating KG-derived biases into physical experimentation. Embedding human arbitration layers—particularly at high-uncertainty decision nodes—would enhance interpretive robustness and experimental reliability [2, 12, 19].

Advanced reasoning and multimodal fusion

A second research trajectory concerns the expansion of KG reasoning capabilities beyond deterministic link traversal. Most current materials KGs encode static relational knowledge, limiting their capacity to reason under uncertainty or infer causal mechanisms. Future systems should integrate probabilistic reasoning frameworks, including Bayesian graphical models layered atop deterministic graphs, enabling uncertainty-aware inference across compositional and processing spaces [13, 24].

Causal reasoning represents an even deeper frontier. Embedding cause–effect structures into KGs would allow systems to distinguish correlation from mechanistic dependency, particularly in processing–structure–property relationships. Such causal augmentation could support counterfactual simulations, enabling researchers to query how modifications in synthesis parameters might propagate through microstructural evolution pathways.

Multimodal fusion further expands KG reasoning horizons. Existing biomedical KGs integrating cell–cell and RNA interactions illustrate how heterogeneous data modalities can be harmonized within unified graph structures [17, 25]. Translating this paradigm to materials science would involve integrating time-resolved simulations, in situ characterization

streams, spectroscopy outputs, and textual knowledge into synchronized graph layers. The temporalization of KGs—encoding process trajectories rather than static states—would significantly enhance predictive modeling of materials evolution.

Knowledge-guided pre-training offers an additional integrative pathway. By embedding graph constraints into foundation models, cross-modal representations could be learned that fuse crystallographic structures, processing text, and spectroscopic signatures. Such models would not merely retrieve knowledge from graphs but internalize graph logic within latent representations, enhancing interpretability and transferability across materials classes [7, 10, 11].

Within inverse design, exploration–exploitation dynamics may be formalized through KG-mediated policies expressed conceptually as:

$$\pi = \operatorname{argmax} (U(Q(KG)) + E(V(KG)))^{(2)}$$

Here, π represents the discovery policy, U denotes utility derived from uncertainty quantification Q , and $E(V)$ reflects expected value extracted from graph-informed predictions. This framing situates KGs not only as knowledge repositories but as active navigational maps guiding active learning toward underexplored materials regions [12, 13, 16].

Scalable and distributed KG infrastructures

As materials data ecosystems expand, scalability emerges as a structural constraint. Future KG infrastructures must therefore transition from monolithic graph databases toward distributed, event-driven architectures capable of real-time updating. Dynamic graph systems already proposed for self-driving laboratories provide early prototypes, enabling experimental results to be ingested and semantically linked as they are generated [2].

Event-sourced architectures could further enhance provenance tracking by recording every graph modification as an immutable transaction. Blockchain-inspired provenance layers offer one conceptual pathway, ensuring tamper-resistant traceability of data lineage, model updates, and experimental decisions [4, 18]. Such infrastructures would be particularly valuable in collaborative environments where data integrity and reproducibility are paramount.

Interoperability remains another critical frontier. The persistence of schema fragmentation—the so-called “materials Babel” problem—continues to hinder cross-database integration. Community-driven standards must prioritize harmonized ontologies, shared identifiers, and interoperable metadata structures spanning simulation repositories, experimental databases, and literature corpora [8, 14, 28].

Scaling knowledge extraction pipelines will also require advances in scientific natural language processing. Integrating domain-specialized models such as MatSciBERT into distributed text-mining infrastructures could enable continuous ingestion of the rapidly expanding materials literature, transforming static KGs into perpetually updating knowledge substrates [7, 20, 21].

Lessons from biomolecular and precision medicine KGs provide valuable infrastructural blueprints. Federated learning paradigms, already explored in biomedical data governance, could be adapted to materials consortia, allowing institutions to collaboratively train KG-enhanced models while preserving proprietary or sensitive datasets [6, 9]. Similarly, reaction chemistry and natural products graphs could expand to encode sustainability metrics, lifecycle impacts, and environmental costs, embedding ecological intelligence into materials discovery pipelines [22, 27].

Ethical and societal implications

As KGs become infrastructural to materials innovation, ethical considerations must be foregrounded rather than appended. One core concern lies in epistemic bias embedded within source corpora. Literature-driven graph construction

risks amplifying geographic, institutional, and disciplinary inequities, privileging well-funded research ecosystems while marginalizing underrepresented contributions [1, 3, 5].

Future research must therefore develop bias-auditing methodologies capable of diagnosing representational imbalances within KGs. Algorithmic transparency in edge weighting, entity inclusion, and inference ranking will be essential to maintain epistemic fairness. Equally important is ensuring equitable access to KG infrastructures, particularly for researchers in resource-constrained environments.

Societal impact integration represents another frontier. Embedding sustainability indicators, ethical risk metrics, and supply-chain considerations into KG reasoning layers could align materials discovery with global development goals. Rather than optimizing solely for performance metrics, KG-guided design could incorporate ecological cost, recyclability, and geopolitical material sourcing risks into discovery heuristics [12, 16].

Conclusion

Knowledge graphs have emerged as foundational infrastructures within computational and data-driven materials engineering, transforming fragmented data landscapes into interconnected reasoning ecosystems. Their capacity to integrate heterogeneous datasets, encode mechanistic relationships, and guide machine learning workflows has positioned them at the core of contemporary materials informatics and autonomous discovery systems [1–29].

This review advances an interpretive synthesis that reframes KGs not merely as databases but as dynamic backbones orchestrating data–model–experiment cycles. Through structured representations, they enable advanced reasoning, uncertainty-aware inference, and closed-loop discovery acceleration. Their integration into autonomous laboratories further signals a paradigm shift from linear experimentation toward continuously learning discovery ecosystems.

Persistent challenges remain, particularly in data quality, interoperability, scalability, and epistemic bias. However, emerging trajectories in hybrid human–AI governance, probabilistic reasoning, multimodal fusion, and distributed infrastructures indicate that these limitations are transitional rather than structural.

Ultimately, the maturation of KG ecosystems will catalyze a broader transformation in materials science—from empirically bounded exploration toward predictive, knowledge-steered innovation. By embedding intelligence not only in models but within the relational fabric of scientific knowledge itself, KGs stand poised to drive the next era of interdisciplinary materials discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 10 Mar 2024

Revised: 10 Apr 2024

Accepted: 19 Jun 2024

Published online: 18 September 2024

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Venugopal V, Olivetti E. MatKG: An autonomously generated knowledge graph in Material Science. *Sci Data*. 2024;11.
- Bai J, Mosbach S, Taylor CJ, Karan D, Lee KF, Rihm SD, et al. A dynamic knowledge graph approach to distributed self-driving laboratories. *Nat Commun*. 2024;15(642).
- Bayerlein B, Schilling M, von Hartrott P, Waitelonis J. Semantic integration of diverse data in materials science: Assessing Orowan strengthening. *Sci Data*. 2024;11(434).
- Statt MJ, Rohr BA, Guevarra D, Breeden J, Suram SK, Gregoire JM. The materials experiment knowledge graph. *Digit Discov*. 2023;2(4):909-14.
- Zhang Y, Chen F, Liu Z, Ju Y, Cui D, Zhu J, et al. A materials terminology knowledge graph automatically constructed from text corpus. *Sci Data*. 2024;11(600).
- Stear BJ, Mohseni Ahooyi T, Simmons JA, Kollar C, Hartman L, Beigel K, et al. Petagraph: A large-scale unifying knowledge graph framework for integrating biomolecular and biomedical data. *Sci Data*. 2024;11(1338).
- Gupta T, Zaki M, Krishnan NMA, Mausam. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Comput Mater*. 2022;8(102).
- Mrdjenovich D, Horton MK, Montoya JH, Legaspi CM, Dwaraknath S, Tshitoyan V, et al. Propnet: A knowledge graph for materials science. *Matter*. 2020;2(2).
- Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. *Sci Data*. 2023;10(67).
- Fang Y, Zhang Q, Zhang N, Chen Z, Zhuang X, Shao X, et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nat Mach Intell*. 2023;5:542-53.
- Li H, Zhang R, Min Y, Ma D, Zhao D, Zeng J. A knowledge-guided pre-training framework for improving molecular representation learning. *Nat Commun*. 2023;14(7568).
- Pyzer-Knapp EO, Pitera JW, Staar PWJ, Takeda S, Laino T, Sanders DP, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput Mater*. 2022;8(84).

Lemm D, von Rudorff GF, von Lilienfeld OA. Impact of noise on inverse design: the case of NMR spectra matching. *Digit Discov.* 2024;3.

Liu S, Su Y, Yin H, Zhang D, He J, Huang H, et al. An infrastructure with user-centered presentation data model for integrated management of materials data and services. *npj Comput Mater.* 2021;7(88).

Hatakeyama-Sato K, Oyaizu K. Integrating multiple materials science projects in a single neural network. *Commun Mater.* 2020;1(49).

Jablonka KM, Ai Q, Al-Feghali A, Badhwar S, Bocarsly JD, Bran AM, et al. 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. *Digit Discov.* 2023;2.

Cavalleri E, Cabri A, Soto-Gomez M, Bonfitto S, Perlasca P, Gliozzo J, et al. An ontology-based knowledge graph for representing interactions involving RNA molecules. *Sci Data.* 2024;11(906).

Statt MJ, Rohr BA, Brown K, Guevarra D, Hummelshøj J, Hung L, et al. ESAMP: Event-sourced architecture for materials provenance management and application to accelerated materials discovery. *Digit Discov.* 2023;2(4):1078-88.

Canty RB, Koscher BA, McDonald MA, Jensen KF. Integrating autonomy into automated research platforms. *Digit Discov.* 2023;2.
<https://doi.org/10.1039/D3DD00135K>.

Yan R, Jiang X, Wang W, Dang D, Su Y. Materials information extraction via automatically generated corpus. *Sci Data.* 2022;9(401).

Hira K, Zaki M, Sheth D, Mausam, Krishnan NMA. Reconstructing the materials tetrahedron: Challenges in materials information extraction. *Digit Discov.* 2024;3(15):1021-37.

Rydholm E, Bastys T, Svensson E, Kannas C, Engkvist O, Kogej T. Expanding the chemical space using a chemical reaction knowledge graph. *Digit Discov.* 2024;3.
<https://doi.org/10.1039/D3DD00230F>.

Kondinski A, Rutkevych P, Pascasio L, Tran DN, Farazi F, Ganguly S, et al. Knowledge graph representation of zeolitic crystalline materials. *Digit Discov.* 2024;3(10):2070-84.

Liu J, Qian Q. Reinforcement learning-based knowledge graph reasoning for aluminum alloy applications. *Comput Mater Sci.* 2023;221.
<https://doi.org/10.1016/j.commatsci.2023.112075>.

Shao X, Li C, Yang H, Lu X, Liao J, Qian J, et al. Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with SpaTalk. *Nat Commun.* 2022;13(4429).

Boudin M, Diallo G, Drancé M, Mougin F. The OREGANO knowledge graph for computational drug repurposing. *Scientific Data.* 2023;10(871).

Gaudry A, Pagni M, Mehl F, Moretti S, Quiros-Guerrero LM, Cappelletti L, et al. A sample-centric and knowledge-driven computational framework for natural products drug discovery. *ACS Cent Sci.* 2024;10(3).

Pelkie BG, Pozzo LD. The laboratory of Babel: Highlighting community needs for integrated materials data management. *Digit Discovery.* 2023;2.
<https://pubs.rsc.org/en/content/articlelanding/2023/dd/d3dd00022b#:~:text=Article%20information-,DOI,https%3A//,-Article%20type>
<https://doi.org/10.1039/D3DD00022B>.

Cavalleri E, Cabri A, Soto-Gómez M, Bonfitto S, Perlasca P, Gliozzo J, et al. An open source knowledge graph ecosystem for the life sciences. *Sci Data.* 2024;11(363).