Chinese Journal of Chemical Engineering 84 (2025) 179–189

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

journal homepage: www.elsevier.com/locate/CJChE

Review

Knowledge graphs in heterogeneous catalysis: Recent advances and future opportunities

Raúl Díaz, Hongliang Xin*

Virginia Polytechnic Institute and State University, Blacksburg, VA, United States

ARTICLE INFO

Article history:
Received 8 March 2025
Received in revised form
23 June 2025
Accepted 23 June 2025
Available online 30 June 2025

Keywords:
Heterogeneous catalysis
Knowledge graph
Ontology
Large language models
Deep learning

ABSTRACT

Knowledge graphs (KGs) offer a structured, machine-readable format for organizing complex information. In heterogeneous catalysis, where data on catalytic materials, reaction conditions, mechanisms, and synthesis routes are dispersed across diverse sources, KGs provide a semantic framework that supports data integration under the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. This review aims to survey recent developments in catalysis KGs, describe the main techniques for graph construction, and highlight how artificial intelligence, particularly large language models (LLMs), enhances graph generation and query. We conducted a systematic analysis of the literature, focusing on ontology-guided text mining pipelines, graph population methods, and maintenance strategies. Our review identifies key trends: ontology-based approaches enable the automated extraction of domain knowledge, LLM-driven retrieval-augmented generation supports natural-language queries, and scalable graph architectures range from a few thousand to over a million triples. We discuss state-of-the-art applications, such as catalyst recommendation systems and reaction mechanism discovery tools, and examine the major challenges, including data heterogeneity, ontology alignment, and long-term graph curation. We conclude that KGs, when combined with AI methods, hold significant promise for accelerating catalyst discovery and knowledge management, but progress depends on establishing community standards for ontology development and maintenance. This review provides a roadmap for researchers seeking to leverage KGs to advance heterogeneous catalysis research.

1. Introduction

Heterogeneous catalysis underpins a vast array of industrial processes and sustainable energy technologies. Its impact on chemical manufacturing has been profound, enabling the production of fuels, chemicals, and materials at scale. However, as the field is expanding, the sheer volume of data has grown exponentially, with hundreds of thousands of publications, patents, and proprietary reports in the past decade alone [1]. Critical insights into catalytic materials, reaction mechanisms, operating conditions, and catalyst synthesis, just to name a few, are often buried within this vast and fragmented literature. This dispersion of knowledge poses a significant challenge to innovation, as no individual can comprehensively review all relevant information. Consequently, there is a pressing need for knowledge management systems capable of extracting, organizing, and linking information to accelerate discovery [2].

While traditional databases capture various aspects of catalysis, they often lack a unified framework for representing conceptual knowledge. Adopting the FAIR principles—findability, accessibility, interoperability, and reusability—ensures that both experimental and computational data can be readily employed [3]. Knowledge graphs (KGs) address this gap by structuring key entities and their interconnections in a graph format, creating a semantic network that is both machine-readable and human-interpretable. By integrating diverse data sources, from databases to journal articles and patents, KGs enable a more comprehensive and queryable representation of catalysis knowledge [4]. This ability to interlink digitized information facilitates the discovery of hidden relationships, ultimately driving hypothesis generation and technology innovation.

This article is part of a special issue entitled: AI for Chemical Engineering published in Chinese Journal of Chemical Engineering.

* Corresponding author.
E-mail address: hxin@vt.edu (H. Xin).

https://doi.org/10.1016/j.cjche.2025.06.008

This Perspective explores the core concepts of ontologies, methodologies for building and maintaining knowledge graphs (KGs), and their integration with artificial intelligence to enhance data-driven discovery of catalytic materials. We highlight how KGs will serve as dynamic tools for structuring and leveraging scientific knowledge in heterogeneous catalysis, examining recent advances, key challenges, and future opportunities for accelerating innovation in the field.

2. Fundamentals of Knowledge Graphs and Ontologies

To ground the discussion, we begin by outlining the twin pillars of any knowledge graph—its graph structure and the high-level ontologies that assign scientific meaning to those nodes and edges.

A knowledge graph (KG) is a structured network of entities (nodes) and the relationships (edges) between them [5]. Each node represents a concept (such as a specific catalyst, a chemical substance, or a reaction), and each edge encodes a relationship linking two entities (for example, "Compound A is produced from Compound B"). These relationships are typically expressed as *triplets*, consisting of a head (subject), a predicate (relationship type), and a tail (object), for instance, the catalyst serves as the head, the reaction as the tail, and the predicate defines their relationship [5]. This triplet-based structure forms the foundation of the graph, offering a clear and adaptable way to represent relationships between entities. KGs can leverage this flexible schema to incorporate new types of nodes and relationships, making them particularly well-suited for the dynamic and interdisciplinary nature of catalysis research. Representing human knowledge in this graph-based format enables machine interpretability, allowing algorithms to traverse the network and infer new relationships [6].

Ontologies build the semantic foundation for constructing nodes and interrelationships. These structured frameworks form the backbone of our knowledge [7,8]. In essence, an ontology serves as a controlled vocabulary and a set of rules, specifying entity classes such as Catalyst, Reactant, Product, and Reaction and defining relationships like HAS_REACTANT, HAS_PRODUCT, and HAS_CATALYST that link these entities. Additionally, ontologies encode hierarchies (*e.g.*, a Zeolite is a subclass of Catalyst) and impose constraints (*e.g.*, a HAS_PRODUCT relationship must connect a Reaction to a ChemicalSpecies entity). By enforcing such a structure, ontologies ensure consistency within a KG and enable semantic reasoning, allowing queries to be answered based on entity types and inheritance [5]. For example, if an ontology defines MetalOxide as a subclass of Catalyst, a query for catalysts active in a given reaction would automatically include metal oxides, even if they are not explicitly labeled as catalysts. This hierarchical reasoning extends the graph's utility by incorporating logical inference. In catalysis, domain-specific ontologies have been developed to standardize terminology for catalytic materials, reaction mechanisms, and more [9].

3. Knowledge Extraction and Integration

With the structural framework sketched, we next consider how evidence from papers, databases, and lab notebooks can be distilled into the triples that populate a catalysis KG.

Building a catalysis knowledge graph involves extracting knowledge from diverse sources and integrating it according to the ontology schema. Knowledge extraction can be performed manually through expert curation or automated using natural language processing (NLP) and data mining techniques [10]. Recent efforts have focused on text mining the scientific literature [11,12]. NLP tools have been developed for named entity recognition (NER), which identifies key entities such as catalysts, reactants, products, or properties, as well as relation extraction, which detects relationships between those entities in text [1,13,14]. For instance, the CatalysisIE, a pre-trained NLP model fine-tuned specifically for catalysis literature, enables the automated identification of six categories of entities (Catalyst, Reaction, Reactant, Product, Characterization, Treatment) directly from text [15]. By applying such models to thousands of abstracts or full-text papers, researchers can systematically extract structured knowledge and populate KGs, creating a scalable and data-driven foundation for catalysis research.

A catalysis KG involves merging extracted knowledge from multiple sources into a unified KG, for example by incorporating information from (a) literature extracted facts, (b) databases, such as the Catalysis-Hub database for surface reaction energetics [16] or PubChem for chemical properties, (c) patents, which often describe new catalysts and processes, and (d) experimental data, including spectra, kinetic measurements, and other lab datasets. Integrating these diverse sources involves aligning different formats and terminologies. Ontologies facilitate this process by providing a common semantic framework—for instance, ensuring that *a* "Pt/C" from a publication is recognized as the same entity as a "platinum on carbon"entry in a database. Chemical identity can often be resolved using standardized identifiers such as InChI keys, SMILES strings for molecules, or unique database IDs. For materials, which often lack universal identifiers, integration may involve synonym matching or linking entries based on shared properties, such as composition or surface area. Domain ontologies, such as OntoSpecies for chemical species [17] or custom-built catalysis ontologies, further support standardization by providing reference identifiers for common entities, enhancing data interoperability across sources.

Another critical aspect of integration is managing heterogeneous data modalities. In principle, catalysis KGs can incorporate not only textual data but also numerical data (*e.g.*, electrode potentials, activation energies, and turnover frequencies) and even images or spectral data by linking to files or metadata nodes. While fully multimodal KGs remain an emerging area, frameworks exist to support such integration[18–20]. For example, an ontology can define a CharacterizationData class and relate it to a *Catalyst* node, enabling connections between a catalyst and its corresponding characterization records. By integrating diverse data types, KGs can evolve into comprehensive knowledge representations, supporting complex queries beyond simple entity relationships. Researchers can not only ask, "Which catalysts facilitate reaction X?" but also, "What is the underlying activity origin of the catalyst Y?" Achieving seamless integration remains challenging, but ongoing community efforts are developing standardized catalysis ontologies and metadata frameworks to facilitate data fusion across synthesis, characterization, kinetics, and process-scale information [9].

4. State-of-the-Art Implementations in Knowledge Graphs

The extraction and integration methods just discussed already power several large-scale catalysis KGs; examining these real-world graphs shows what they can achieve today and where gaps still lie.

Knowledge graphs (KGs) provide a machine-actionable fabric that links chemical composition, synthesis parameters, characterization data, and catalytic performance into a single, semantically consistent network. By rendering data FAIR (Findable, Accessible, Interoperable, Reusable) and queryable with standard graph and embedding techniques, KGs unlock capabilities that traditional databases cannot offer: large-scale hypothesis generation through link-prediction, on-the-fly reasoning across heterogeneous data modalities, and seamless integration with

autonomous experimentation workflows. Consequently, KGs are becoming a foundational infrastructure for accelerating catalyst discovery and process optimization. Here, we highlight key state-of-the-art implementations that demonstrate how KGs are transforming catalysis research.

Behr *et al.* [1] developed one of the first large-scale KGs for catalysis by applying AI-assisted text mining of the scientific literature. Their workflow leveraged CatalysisIE to extract key information from catalysis papers and structure it within an ontology-backed KG. The resulting graph comprised thousands of extracted triples, encoding relationships such as which catalysts were used for specific reactions under given conditions. By enabling queries via SPARQL, they demonstrated the ability to retrieve insights from the literature, such as identifying all reported instances of a given reaction along with corresponding catalysts and conditions. This work represents a significant development, illustrating how a largely automated pipeline can process the growing volume of literature and convert unstructured text into a structured, queryable knowledge resource for catalysis. Technically, Behr et al. fine-tuned the SciBERT-based CatalysisIE NER on a bespoke 10.4 M-word catalysis corpus, integrate ontology-learning and active-learning loops to auto-generate high-fidelity subject–predicate–object triples, and validate the end-to-end workflow on two test datasets via pre-formulated SPARQL queries, delivering a fully automated literature to KG pipeline with markedly improved precision and recall.

Zhang *et al.* [2] introduced CatKG, a multi-source semantic knowledge graph designed to recommend catalysts for organic reactions. Their approach integrated structured reaction data, curated reactant-catalyst-product triples, with knowledge extracted from approximately 220000 research article abstracts. The resulting KG employed a property graph schema that not only captured reactant-catalyst-product relationships but also incorporated additional semantic information, such as elemental composition and physical properties of substances. What distinguishes CatKG is its ability to learn from the graph itself. By applying word2vec and other machine learning techniques to the combined data, the system can be used to recommend potential catalysts for specific reactions, including those not previously explored, effectively predicting new links in graphs. CatKG's distinctive advance is its combination of word2vec-based analogical inference and a masked-language-model reasoning module: trained on curated reactant–catalyst–product records, the system achieves up to ~80% accuracy in recommending catalysts for large molecules when models are trained on smaller ones.

Gao et al. [13] developed a knowledge graph specifically for Cu-based electrocatalysts in CO_2 reduction, leveraging a SciBERT-based NLP framework to extract multiple entity types, including materials, synthesis methods, and performance metrics such as Faradaic efficiency from 757 publications. The resulting KG provided a structured representation of the field, enabling researchers to trace the historical development of Cu-based electrocatalysts and analyze how factors such as preparation methods influence performance. Beyond serving as a repository, the KG was integrated with predictive modeling. By generating both word embeddings from text and graph embeddings from the KG, a neural network was trained to predict catalyst performance. This approach allowed us to forecast the Faradaic efficiency of new catalyst formulations, demonstrating the power of combining knowledge-driven and data-driven methodologies. Technically, Gao et al. integrate graph embeddings into their regression framework (using a GraphSAGE-based model), reducing the mean absolute error on Faradaic-efficiency prediction from 0.148 (the best text-only MLP baseline) to 0.081, effectively halving the prediction error and underscoring the added value of KG structure.

Venugopal *et al.* [21] introduced MatKG, an autonomously generated materials knowledge graph encompassing over 2 million unique triples among approximately 80000 entities spanning the material–structure–property–processing paradigm. Their pipeline uses transformer-based language models to infer a pseudo-ontological schema via statistical co-occurrence mapping, then applies SHACL validations to enforce schema consistency before ingesting the graph into an RDF store with SPARQL endpoints. For downstream inference, MatKG learns both TransE embeddings [22] and ComplEx embeddings [23], achieving a mean reciprocal rank of 0.49 and hits@10 of 0.66 on link-prediction benchmarks. Finally, MatKG is published in CSV and RDF formats and exposed via interactive dashboards, enabling its seamless integration into materials-discovery recommendation systems and advanced analytics workflows.

On the ontology side, researchers have been systematically mapping available ontologies for catalysis data. The Ontologies4-Cat [24] initiative surveyed and cataloged ontologies spanning synthesis, characterization, reaction kinetics, and catalyst properties, developing workflows to interlink them into a cohesive framework. This effort effectively constructs a meta-knowledge graph at the ontology level, ensuring that initiatives such as CatKG, and future developments can align their schemas with broader standards. By fostering interoperability and machine-readability on a global scale, Ontologies4Cat represents a state-of-the-art approach to standardizing catalysis knowledge, facilitating seamless integration across diverse datasets and research efforts[6].

5. Development of Knowledge Graphs

Having seen what current catalysis KGs can achieve, we now look into their construction, tracing the journey from disparate raw data sources to a unified, queryable graph.

Building a catalysis knowledge graph involves multiple stages, including identifying and collecting relevant data sources, selecting or designing an ontology, and constructing the graph by systematically populating it with entities and relationships [23,25,26]. As illustrated in Fig. 1, we first read diverse data sources, structured, unstructured, and semi-structured, using appropriate parsing techniques. Once all sources are loaded, we apply knowledge-extraction methods to derive subject–predicate–object triples. These extracted relationships are then consolidated into a central knowledge graph, which can be interrogated via structured graph queries or retrieval-augmented LLM-based approaches. Finally, this unified graph supports a range of downstream applications—including catalyst discovery, reaction-mechanism analysis, and process optimization.

This process is inherently iterative, often blending automated methods with expert curation to ensure accuracy and completeness [10,27]. In this section, we outline the fundamental components and methodologies for developing a KG tailored to heterogeneous catalysis, highlighting best practices for data integration, ontology alignment, and graph construction.

A well-constructed catalysis KG integrates multiple data sources to comprehensively represent the field. Journal articles and conference papers serve as primary knowledge reservoirs, documenting the core elements and dynamics of catalysis research [1,28]. Text mining these publications can extract vast numbers of triples (subject-predicate-object assertions), forming the foundational relationships within the KG. Literature mining typically relies on NLP techniques such as named entity recognition (NER) and relationship extraction to identify and structure key information [12,28,29]. These methods extract entities, including catalyst materials, reactants, products, reaction conditions (*e.g.*, temperature,

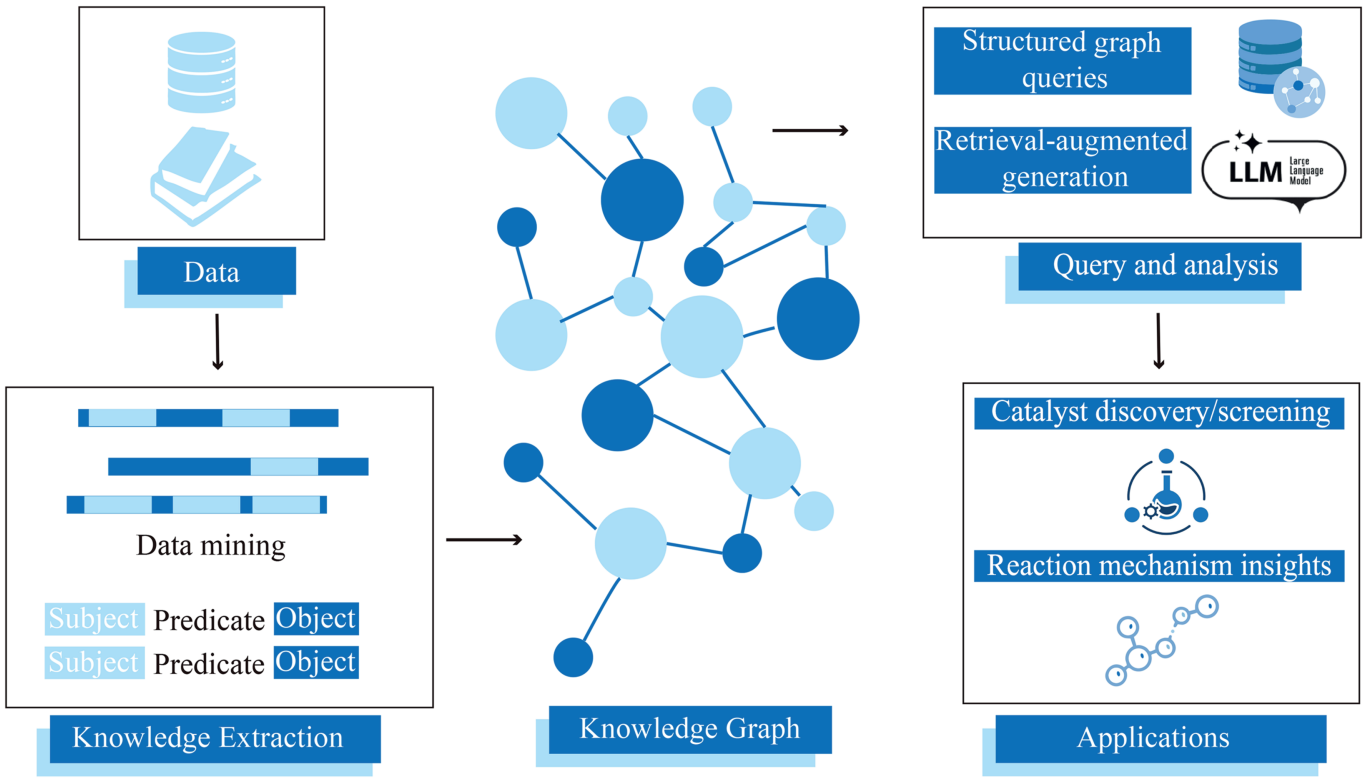

Fig. 1. Schematic knowledge-based workflow from data to final applications in heterogeneous catalysis.

pressure), and performance metrics (*e.g.*, yield, selectivity), while mapping their interconnections [15,27,30]. By systematically harvesting such knowledge, KGs transform unstructured textual data into a machine-readable framework, facilitating more efficient retrieval, synthesis, and discovery in catalysis research.

In heterogeneous catalysis, structured databases such as Catalysis-Hub [16] provide valuable foundational knowledge through DFT-calculated reaction energetics, reaction rate constants, and experimentally validated catalysts. These datasets underpin knowledge graphs (KGs), as demonstrated by CatKG [2]. Typically stored in tabular formats, these structured data are incorporated into KGs by mapping database fields to ontology classes, enabling more accurate and accessible representations of catalytic knowledge. Computational insights from simulations, including DFT calculations, molecular dynamics, and microkinetic modeling, further enrich KGs by detailing interactions such as adsorption energies from extensive datasets like those released by the Open Catalyst Project [31].

In parallel, patents represent an untapped reservoir of applied catalytic insights, encompassing industrial catalyst formulations, reaction conditions, and process optimizations that often remain undocumented in academic publications. While patent text mining requires NLP models to undergo domain-specific adaptation due to distinct terminologies, broader claims, and proprietary language, no publicly available catalysis patent KG has yet been reported. Nonetheless, integrating patent-derived data into existing catalysis KGs represents a promising next step, potentially bridging the existing gap between theoretical academic discoveries and industrial applications. Consequently, the synthesis of structured experimental data, computational modeling, and patent knowledge supports comprehensive catalyst design, enhances process optimization strategies, and fosters a robust and predictive research framework.

High-throughput experimentation, pilot plant runs, and characterization techniques generate valuable data for catalysis knowledge graphs, often stored in laboratory information management systems (LIMS), spreadsheets, or supplemental materials in publications. A major challenge in incorporating experimental data into KGs is its heterogeneity and scale. A single study may generate hundreds of data points under varying conditions, along with extensive metadata on catalyst preparation (*e.g.*, precursor selection, synthesis temperatures, aging times) and characterization parameters (*e.g.*, instrument settings). Without a consistent schema, merging such information into a KG risks inconsistencies and errors. To address this, community standards and ontologies for catalysis experiments are under development [32]. For instance, by systematically capturing synthesis details and linking them to catalyst entities, KGs enable complex queries such as “Find all catalysts prepared using sol-gel methods and their performance in reaction X.” Although labor intensive, metadata recording substantially enhances the value of KGs, making them more informative and actionable for catalysis research.

5.1. Ontology design

At the heart of any knowledge graph lies its ontology, a concise map of catalysts, reactions, conditions, and their relationships that we now unpack, outlining how these schemas are designed, extended, and used to drive seamless data integration and machine reasoning.

Designing the ontology (or schema) for the KG influences how well the graph can represent knowledge and answer questions. A good ontology for catalysis balances expressiveness (capturing all important concepts and relations) with simplicity (avoiding overly complex or unnecessary constructs).

Designing an ontology for KGs is a vital first step. It shapes how well the graph can store information and respond to questions. A strong catalysis ontology captures all key concepts and relationships while staying as simple as possible. In heterogeneous catalysis, common entities include Catalyst, Reaction, Reactant, Product, CatalystSupport, ActiveSite, Promoter, Condition, KineticData, and Characterization. These can be broken into smaller groups. For example, Catalyst can be split into MetalCatalyst, Zeolite, or Single-AtomCatalyst, and Reaction can be divided by types like hydrogenation, oxidation, or C—C coupling. Aligning these with existing reaction ontologies helps ensure consistency.

For instance, Fig. 2 illustrates an ontology-based representation clearly defining relationships among entities central to catalytic processes. Catalysts are associated with specific Active Sites and catalyze particular Reactions. These reactions require defined Conditions such as temperature, pH, and pressure, and involve specific Compounds that ultimately form distinct Products. This structured network clarifies how these elements interconnect, facilitating exploration, querying, and interpretation of catalytic knowledge.

Once the entities and relationships are established, the ontology serves as a plan for building and updating the KG. It guides how new information is added and how existing information is maintained over time, rather than creating an ontology from scratch, it is often beneficial to reuse existing ontologies or standards. In chemistry and materials science, various ontologies exist (*e.g.*, ChEBI for chemicals [33], OntoSpecies for chemical species properties [17], the RXNO reaction ontology [34], and domain ontologies for kinetics or process engineering). An ontology for catalysis can import or link to these, so that, for instance, chemicals in the KG are classified under OntoSpecies or ChEBI, inheriting those ontologies' detailed property definitions [1]. There have been efforts to survey and map the landscape of ontologies relevant to catalysis: for example, Behr *et al.* [24] compiled catalysis-related ontologies and even developed code to align classes between them. This kind of work helps avoid duplication and ensures that a catalysis KG can interoperate with external data (interoperability is a key part of FAIR). If an existing ontology is close to the needs, one can extend it by adding missing terms.

Ontologies can be implemented in formal languages like OWL (Web Ontology Language) [5,6] which allows one to use logical axioms and reasoners, or one can use a simpler schema definition if using a property graph database (*e.g.*, Neo4j). Using OWL/RDF (Resource Description Framework) has advantages for complex reasoning. For example, one can declare logical rules such as transitive properties or equivalence. For instance, an ontology can include the rule that if A *HAS_ACTIVE_SITE_B* and B *HAS_PROPERTY_X*, then A *HAS_PROPERTY_X* (transferring a property from site to catalyst). Such capabilities let the KG infer new knowledge (*e.g.*, infer that a catalyst is bifunctional if it has two distinct active sites with certain properties). On the other hand, property graphs allow attributes on nodes/edges which can be convenient for storing numerical data or text annotations directly on the graph elements. The choice may depend on the specific use-case; some projects even maintain dual representations (an RDF/OWL version for sharing and reasoning, and a property graph for fast querying and visualization).

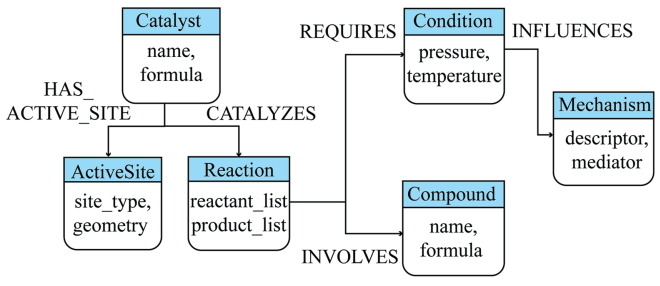

Fig. 2. Schematic ontology structure for heterogeneous catalysis.

Ontology development is iterative. Initial versions might be simplistic and then expanded. As the KG is populated, one might find new concepts that need to be added or relationships that were not anticipated. For example, while adding data, one may realize that an important concept like "Catalyst Deactivation" should be represented (perhaps as a property of the catalyst with a link to cause). The ontology should be updated to accommodate this. Maintaining alignment with evolving knowledge is itself a challenge but also a strength—an ontology can continuously grow to capture new aspects of catalysis [35]. Domain experts should validate the ontology by testing whether typical queries can be formulated and whether the classification makes sense (*e.g.*, ensure that all catalysts in the KG fall under the appropriate ontology classes, and that no important relationship type is missing).

The ontology/schema is the blueprint of the knowledge graph. A thoughtful design, possibly leveraging existing ontologies and expert knowledge, will make the subsequent steps of graph construction and usage much more effective. In catalysis, ontology design is an active area of research, with initiatives dedicated to standardizing how we represent catalytic data and knowledge.

5.2. Graph construction approaches

With the ontology in place, we translate evidence into nodes and edges, combining automated NLP pipelines, scripted database ingests, and expert curation to populate the graph at scale. Various approaches can be used, often in combination, to achieve this.

One common strategy involves natural language processing (NLP) techniques that extract structured triples from text. Supervised machine learning models, trained on annotated corpora, are widely employed for this task. In catalysis, the CatalysisIE serves as a prime example: it was trained on a benchmark dataset of catalysis literature, where human-labeled entities and relationships provided a foundation for learning. Once trained, the model could extract new triples from previously unseen publications. Such models typically leverage transformer architectures [36] to process scientific language effectively [15]. The output of an NLP pipeline is usually a list of extracted entities and their relationships, which must then be mapped to the ontology and then used to construct or update the KG. Automated tools can thus generate substantial portions of a KG [37,38]. For instance, a text-mining study in materials science demonstrated how NLP techniques could extract hundreds of triples, which were subsequently integrated into an ontology-driven KG and made queryable *via* SPARQL [39]. As illustrated in Fig. 3, a text is parsed *via* named entity recognition (NER) to identify key entities—such as Ni/CeO_2, CO oxidation, 200 °C, notable activity, and support interactions (left). These extracted terms are then integrated into a node-based representation (right), illustrating how relationships among catalysts, reaction conditions, and performance descriptors can be systematically captured for deeper insights into catalytic behavior. These approaches significantly enhance the scalability and usability of knowledge graphs in scientific domains.

For structured sources such as databases or spreadsheets, constructing a KG often involves writing scripts to convert each record into graph entries. For example, a table of catalyst properties can be transformed by creating a Catalyst node for each row (if

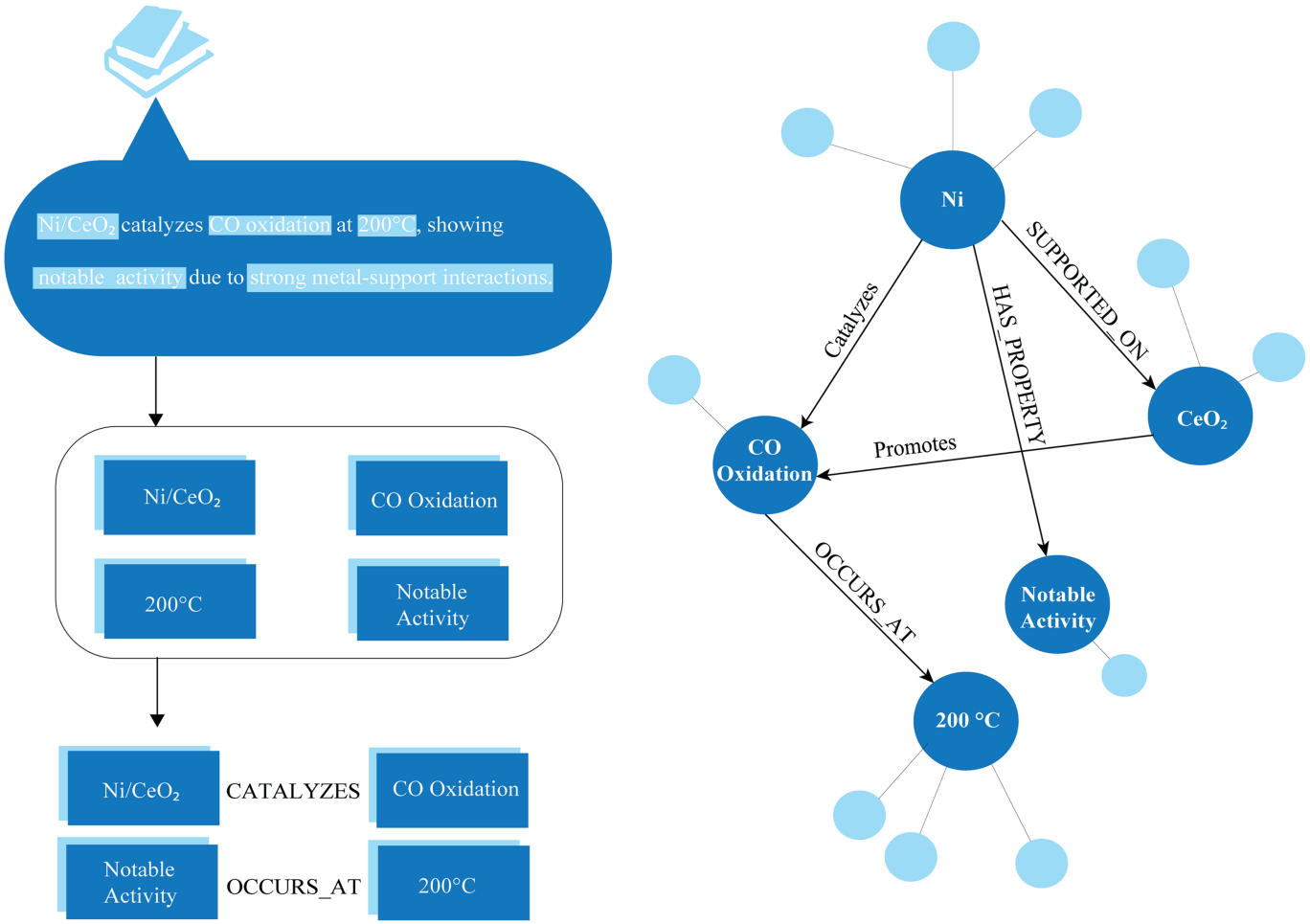

Fig. 3. Schematic of the knowledge-graph construction process in heterogeneous catalysis.

not already present), then attaching property values (surface area, particle size, *etc.*) as either literal attributes or as nodes connected by *HAS_PROPERTY* edges. Several tools facilitate this process. RDFizers exist for many common data formats, enabling structured data to be converted into RDF triples. Alternatively, custom scripts can be written in Python using libraries like RDFlib (for handling RDF data) or Neo4j connectors (for working with graph databases). Each material can then be represented as a node in the knowledge graph, linked to corresponding reaction nodes based on known catalytic reactions from other datasets. This approach allows structured data to be efficiently transformed into a rich, queryable graph format.

Despite advances in automation, expert curation remains essential for refining and validating the KG. In the early stages, domain experts may manually input foundational knowledge to establish foundational examples. Expert oversight is often required to verify critical relationships and prevent the propagation of errors. In the CatalysisIE pipeline, an active learning approach was employed [15], where human annotators helped resolve ambiguous cases, iteratively improving the model's accuracy. This continuous interplay between human expertise and machine learning enhances the quality and reliability of the KG over time.

Once an initial knowledge graph (KG) is constructed, AI techniques can predict missing links—a process known as link prediction or graph completion [40,41]. Knowledge graph embedding models encode entities and relationships as vector representations in a latent space, ensuring that known connections in the graph are preserved [4]. These embeddings can then suggest new plausible relationships; for example, if catalysts A and B have similar embedding vectors, and A is known to catalyze reaction R, the model might predict that B could also catalyze R. While link prediction has been widely used in domains such as social networks and general knowledge bases, its application in chemistry is emerging. For instance, a KG embedding approach could identify candidate catalysts for a reaction by drawing analogies to known catalysts [13]. Although such predictions require experimental validation, they demonstrate how a KG can evolve from a static repository into a hypothesis-generating tool, accelerating discovery in catalysis.

6. Querying Knowledge Graphs

With the KG fully assembled, we next explore how graph-oriented query languages and intuitive dashboards let researchers interrogate its structure, extracting catalysts, conditions, and performance relationships to drive data-driven hypotheses.

Knowledge graph value depends on how easily one can query it to retrieve relevant information and how reliably one can update it with new knowledge. This section describes m ethods for querying catalysis KGs and approaches to keeping them current and accurate over time. Knowledge graphs can be queried using specialized query languages that leverage the graph structure. For RDF-based KGs, the SPARQL query language—standardized by W3C—enables

pattern-matching queries using triple patterns [5,42]. For example, a SPARQL query to identify catalysts for a given reaction might search for triples of the form (Catalyst, catalyzes, Reaction X). SPARQL is highly expressive, supporting filters, optional patterns, aggregations, and other advanced features, making it analogous to SQL for relational databases. In the catalysis domain, prewritten SPARQL queries have been used to address specific competency questions within KGs [1]. For instance, Behr et al. designed SPARQL queries to extract all reactants, catalysts, supports, and products mentioned in a given publication or to identify all reactions that produce a specific product along with the catalysts involved. One of SPARQL's strengths is its ability to traverse paths of arbitrary length, making it particularly useful for complex queries. Many KG implementations provide SPARQL endpoints, allowing users and applications to submit queries and receive results in structured formats such as JSON or CSV.

If the KG is implemented as a property graph (*e.g.*, in Neo4j), Cypher is a widely used query language [43]. Cypher uses a SQL-like syntax tailored for graph structures, using pattern matching with an intuitive ASCII-art notation [44]. For example, a Cypher query to retrieve catalysts for a reaction might be:

```
MATCH (c:Catalyst)-[:catalyzes]->(r:Reaction {name: "Reaction X"})
RETURN c.
```

Cypher supports conditions on node/edge properties, making it convenient for filtering results—such as selecting catalysts with a specific surface area or those discovered after a certain year. Its readability has made it popular in various knowledge graph projects[45–47]. Other alternatives for querying graphs include Gremlin [48], an API-based traversal language, and emerging standards like GraphQL, which can be integrated with graph databases. The choice of query language depends on the underlying database technology and the intended use case. For scientists and other end-users, writing SPARQL or Cypher queries directly can present a steep learning curve. To enhance accessibility, user interfaces or query builders are often developed on top of these languages. For example, a web interface might offer drop-down menus for selecting reactants and products, with the backend automatically generating and executing a SPARQL query to retrieve matching catalysts. Some systems also support natural language queries, where users can phrase questions in plain language, and the system translates them into structured graph queries[49–51]. This remains an active research area, integrating natural language processing (NLP) with knowledge graph querying to enable more intuitive interaction with complex scientific data.

7. Updating Knowledge Graphs

Having shown how to effectively query the graph, we now turn to maintaining the trustworthiness of its answers. This section outlines lightweight update loops that allow the KG to evolve reliably alongside the rapidly changing catalysis landscape.

Since catalysis knowledge is dynamic, the KG should be updated to remain relevant and avoid becoming a stale snapshot. Updating a knowledge graph involves adding new nodes/relationships, and occasionally modifying or removing existing ones in light of new evidence.

A robust catalysis KG should implement periodic updates to integrate new findings. For literature-based updates, this can involve automatically retrieving new publications *via* APIs such as CrossRef or Scopus, based on predefined search criteria (*e.g.*, keywords like catalysis or more targeted queries for specific reactions and materials). These new documents can then be processed through an NLP extraction pipeline to generate additional triples, which are subsequently inserted into the KG. Behr *et al.* [1] demonstrated this approach by using query results to identify and incorporate newly relevant publications, allowing the KG to expand continuously as the literature grows. Similarly, if a database source has updates (*e.g.*, an updated version of a surface reactions database with new DFT entries), those can be fetched and merged. Keeping a log of what data/version has been ingested is important to avoid duplication. For RDF-based KGs, SPARQL Update queries (a W3C standard) allow efficient insertion and deletion of triples, enabling seamless integration of new knowledge. In property graphs, updates can be performed using the database's API or query language to create, modify, or remove nodes and edges as needed.

Sometimes new data doesn't fit neatly into the existing schema. For instance, researchers might start reporting a new type of descriptor (*e.g.*, catalyst surface defect density) that wasn't in the ontology. In such cases, the ontology can be updated by adding new properties or classes to accommodate the evolving knowledge. Ontologies support versioning, keeping backward compatibility (*e.g.*, add new subclasses or relationships without altering the existing structure). This flexibility allows the KG to evolve as the field evolves [28]. However, modifications must be approached carefully; adding a new relationship type can cause existing queries to overlook relevant results unless they are updated accordingly. For this reason, ontology changes are often batched into major version releases of the KG to ensure consistency and minimize disruptions.

When new data is added, especially through automated methods, validation steps are essential to catch errors. NLP-based extraction is not perfect, as it may misidentify compounds or relationships. A common strategy is to assign confidence scores to extracted triples, only incorporating those above a certain threshold while flagging low-confidence ones for manual review. To further ensure data integrity, constraint-based validation tools, such as SHACL (shapes constraint language), can enforce rules on the graph [52]. If a new entry violates a predefined rule, the system should either flag or reject it to maintain high precision and reliability ensuring user trust in the KG, as erroneous or inconsistent data can undermine its utility.

Every fact in a catalysis KG ideally links to its source. Provenance tracking plays a crucial role in maintaining data integrity, preventing duplicate entries, and resolving conflicts. For example, if an earlier publication reported that "Catalyst A works for Reaction R″ and a new publication reports it did not work, the KG might end up with conflicting information. Rather than deleting the old fact, one could annotate the graph with context (the new finding might be a specific condition under which it didn't work). However, if the earlier fact is determined to be a clear error (*e.g.*, due to a misreading or mislabeling), curators may decide to correct or remove it. Provenance trails ensure that each assertion can be checked, and if needed, corrected by referring to the original source [1]. This transparency allows users to evaluate the reliability of the data and make informed decisions, reinforcing trust in the KG as a dynamic and credible knowledge repository.

Knowledge drift occurs when a knowledge graph gradually becomes outdated or misaligned with the latest understanding [53]. In catalysis, this can happen if a long-accepted reaction mechanism is later revised and the KG should be updated accordingly to reflect the new consensus. Preventing drift requires regular maintenance, which can be achieved through scheduled updates, such as monthly literature ingestion, and by actively monitoring critical topics. Large language models (LLMs) may offer a potential solution. One proposed approach involves using an LLM agent that periodically scans new publications to identify findings

that contradict or extend existing KG entries, then suggests updates—a form of automated literature surveillance [14]. While still experimental, such methods can significantly reduce the effort required to keep the KG current.

Maintaining a catalysis KG is an ongoing effort, much like managing a living database. In industrial settings, some organizations integrate KGs into their knowledge management infrastructure, employing dedicated teams to update them with new lab results and published data. In academia, open KGs often rely on a combination of automated scripts for data ingestion and community curation for validation and refinement. The primary goal is to ensure that the KG remains an accurate and dynamic reflection of state-of-the-art knowledge in catalysis. When well-maintained, a KG can mitigate knowledge drift for researchers by continuously integrating the latest findings. Rather than relying on scattered sources, scientists can use the KG as an up-to-date reference that evolves alongside the field, helping them stay informed and make data-driven discoveries more efficiently.

8. Enhancing Queries with Retrieval-Augmented Generation

With an up-to-date KG in place, we now demonstrate how retrieval-augmented generation (RAG) combines knowledge graphs with large language models to transform free-form questions into context-rich answers. This section outlines the pipeline that enables this integration.

The intersection of knowledge graphs with advanced AI, particularly large language models (LLMs), has opened new frontiers for how we interact with and augment the knowledge in catalysis [13,54,55]. Retrieval-augmented generation (RAG) refers to the approach of improving LLM outputs by retrieving relevant information from a knowledge source and providing it as context to the model. In catalysis, integrating KGs with LLMs can significantly enhance tasks such as question-answering, hypothesis generation, and automated literature analysis [56]. LLMs are powerful at understanding and generating human-like text, but they have limitations: they can produce hallucinations (incorrect statements) if they rely solely on their internal training [57]. RAG addresses this by coupling the LLM with an external knowledge retrieval step [46,58,59]. For instance, consider a researcher asking an AI assistant: "*What catalysts are known for oxidative coupling of methane and what yields do they achieve?*" A standalone LLM might not recall specific data or might fabricate an answer. In a RAG setup, the question would first be used to query the catalysis KG (via SPARQL/Cypher or a semantic search) to fetch relevant facts, *e.g.*, a list of catalysts studied for oxidative coupling of methane (OCM) and associated yield data from the KG. These facts (with references) are then provided to the LLM, which uses them to formulate a coherent answer. The result is an answer that is both fluent and grounded in the KG's factual data, reducing the chance of error and providing traceability to sources.

GraphRAG is a specific RAG implementation leveraging graph databases[56,60–62]. A pipeline might work as follows: the user's query is converted (possibly via a trained model or templates) into a graph query to the catalysis KG; the relevant subgraph (set of triples) is retrieved; the LLM then encodes this information and generates an answer or report. The KGs might store not only direct answers but also contextual information (*e.g.*, reaction conditions, publication references), which the LLM can integrate into a nuanced response. This approach can essentially create an AI research assistant that can be asked high-level questions ("*Suggest some catalysts for biomass conversion to lactic acid*") and it will utilize the KG to deliver evidence-based suggestions, complete with literature references. For instance, Fig. 4 illustrates a retrieval-augmented generation (RAG) pipeline where a user prompt is enhanced by combining information from both a vector database and a knowledge graph. The system issues a query to retrieve semantically relevant content (vector query) and structured domain knowledge (graph query), which are then integrated into an enriched prompt. By feeding this enriched prompt to a large language model (LLM), the pipeline produces more accurate and context-aware responses than would be possible with the prompt alone.

Using a knowledge graph as the retrieval source has particular benefits. KGs can store structured relationships that might be hard for an LLM to infer from raw text alone [59,63,64]. For example, the KG can explicitly connect catalyst compositions to performance metrics, so a retrieval step can pull not just documents but exact data points. The LLM then can present, compare, or reason about these points. This structured retrieval leads to more accurate and transparent AI responses. In the medical domain, RAG has shown improved accuracy over unassisted LLMs [65], and we expect similar advantages in technical domains, where precision is critical.

9. Key Challenges and Future Directions for Catalysis Knowledge Graphs

To advance the impact of knowledge graphs in catalysis, it is essential to confront persistent challenges, including inconsistent data formats, fragmented ontologies, limited scalability, and barriers to user adoption. This section explores the strategies, both technical and collaborative, that are shaping the path forward.

Catalysis research spans a broad range of data modalities—textual descriptions, numerical measurements, images (*e.g.*, microscopy, spectroscopy), and computational outputs—making it challenging to integrate everything into a single knowledge graph (KG). Linking characterization data, including spectra or

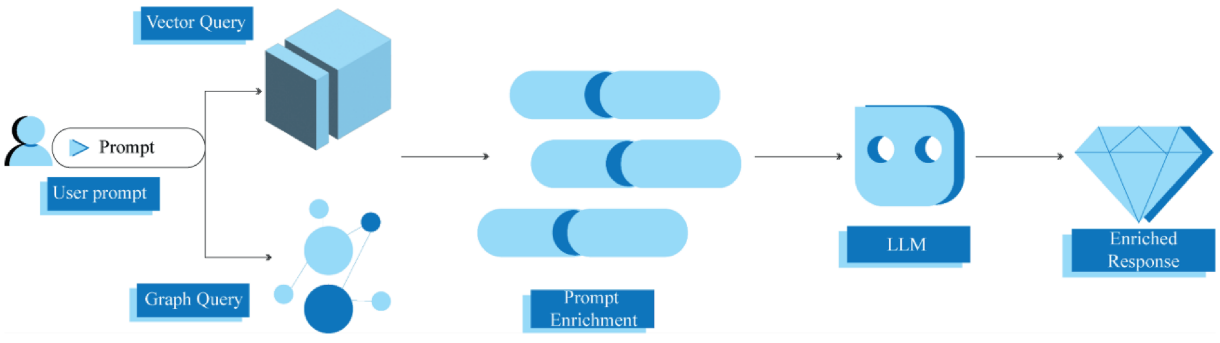

Fig. 4. Example retrieval-augmented workflow that integrates vector database and knowledge graph queries to enrich user prompts for improved LLM responses.

microscopy images with performance metrics demands consistent metadata and ontology terms. However, many experimental details (*e.g.*, protocols, equipment specifications, or chemical batches) remain unstated or poorly standardized, which risks losing critical context. Automated text extraction further complicates curation, as domain-specific jargon and subtle distinctions between correlation and causation can lead to misinterpretations. A related issue is publication bias: negative results are often omitted, skewing the KG's representation of what has been tried and tested.

A major technical hurdle is the absence of a universally accepted ontology for catalysis. Different research groups and databases may define core concepts or more specialized terms using unique schemas. Merging data across these sources requires careful ontology alignment to ensure that, for example, *SupportMaterial* in one ontology is recognized as *CatalystSupport* in another. Tools for automated alignment are improving, but they often need domain-expert intervention. Compounding this challenge is the field's rapid evolution; concepts such as *Single-atom Catalysts* were less relevant a decade ago but are increasingly important now. Ontologies must be flexible to accommodate such shifts without breaking older systems and queries. This might involve modular ontology design (with domain-specific extensions) and version control so that new terms can be introduced while preserving backward compatibility. Standards bodies like NFDI4Cat and IUPAC can facilitate convergence, but achieving broad adoption remains a social and technical endeavor [66–68].

Scalability and performance issues emerge as KGs grow to include millions of triples. Graph databases must be optimized to handle complex queries without long delays. Indexing strategies, caching results, or focusing on domain-specific subgraphs can help ensure that users get near-instantaneous answers. Beyond technical optimization, community adoption poses another significant barrier. Many chemists and engineers are unfamiliar with semantic technologies, and if using a KG requires learning complex query languages or intricate data-wrangling procedures, potential users may disengage. Demonstrating tangible benefits, *e.g.*, faster literature triage or automated hypothesis generation, and providing intuitive interfaces are key to broader acceptance. Moreover, trust is essential: researchers need confidence in both the accuracy of KG data and the transparency of its provenance. Community-driven projects risk inconsistent contributions, so editorial boards or similar curation models may be necessary to maintain quality and uniformity.

Despite the conceptual strengths of catalysis KGs, real-world deployment faces two key practical constraints. First, computational cost can be prohibitive: embedding millions of triples and performing large-scale graph reasoning often requires GPU-accelerated clusters with tens to hundreds of gigabytes of RAM, driving up both infrastructure and energy expenses [69]. Second, user accessibility remains limited, most experimentalists and process engineers lack familiarity with SPARQL, Cypher, or graph-API paradigms, making intuitive GUIs or notebook-style interfaces essential to lower the technical barrier to entry.

A second set of challenges revolves around ongoing maintenance and integration in the field. Catalysis knowledge can change, sometimes dramatically, KGs must also guard against knowledge drift. A reaction mechanism accepted as fact today can be disproven tomorrow, and the graph must be updated accordingly. Outdated or conflicting information can linger, potentially misleading new research. Automated alerts triggered by newly published findings, combined with expert review, can mitigate this inertia. Proprietary data in industry adds another layer of complexity: companies may maintain private KGs or share only non-sensitive data publicly, limiting the completeness of open resources. Balancing commercial interests with the scientific community's need for shared knowledge remains an ongoing challenge.

Over the longer term, we may see KGs function as a "brain" for self-driving labs, guiding autonomous experimentation. By identifying gaps in the graph, such as missing links, untested catalysts, or underexplored reaction conditions, a machine could design and execute new experiments, feeding the results back into the KG in a closed-loop cycle. This approach promises to accelerate catalyst discovery and reduce time-to-insight. Moreover, community-curated KGs, akin to the Protein Data Bank for protein structures, can become a widely recognized repository of catalyst information.

Finally, catalysis KGs will likely link to knowledge in adjacent domains, from process engineering to environmental impact. Cross-KG queries such as *"Given catalyst performance data and process sustainability metrics, which catalyst best balances yield and minimal waste?*", could transform how researchers and industries approach green chemistry. Ontology alignment remains pivotal here: developing robust, extensible schemas is critical for seamless inter-domain data exchange. These future scenarios also promise a new level of explainable AI, where predictions about catalyst performance can be grounded in explicit, well-curated knowledge of reaction pathways, material properties, and mechanistic details. While practical and conceptual challenges persist, catalysis knowledge graphs are poised to become a core driver of innovation, enabling faster discovery, deeper insights, and more transparent research across the field.

10. Conclusions

Knowledge graphs are poised to fundamentally enhance catalysis research by transforming how knowledge is collected, connected, and utilized. As we have reviewed, KGs enable researchers to navigate the complex web of relationships in heterogeneous catalysis, linking catalysts to reactions, conditions, mechanisms, and outcomes in a structured and machine-readable way. This provides a foundation for data-driven insights that complement traditional theory and experimentation. Early implementations have shown that KGs can accelerate literature review, suggest new catalyst candidates, and even aid in predicting catalytic performance. The incorporation of domain ontologies ensures that these graphs carry rich semantic meaning, allowing sophisticated queries and logical inferences that were previously impractical.

The integration of KGs with AI techniques, especially LLMs and graph algorithms, stands out as a game-changer. By using retrieval-augmented generation, we can harness the generative power of LLMs while keeping their outputs grounded in factual, up-to-date catalysis knowledge. Conversely, using AI to help build and maintain the graphs reduces the barrier of manual effort and keeps the KG in sync with the fast-paced advances in the field. This symbiosis of KGs and AI is expected to yield intelligent systems that assist researchers in making sense of data and even autonomously generate new hypotheses. We anticipate that future catalysis labs will routinely consult knowledge graphs through natural language interfaces as part of the research workflow, much like having a domain expert always on hand to provide relevant information.

While challenges remain in terms of data integration, standardization, and ensuring broad adoption, the progress to date and ongoing efforts suggest these can be overcome through community collaboration and technological innovation. Initiatives to develop shared ontologies and data platforms indicate a trend toward collective knowledge management in catalysis. As more data is contributed and more organizations invest in this knowledge infrastructure, the KGs will become increasingly comprehensive and valuable.

In conclusion, knowledge graphs represent an important step toward a more intelligent and interconnected catalysis research ecosystem. By capturing what we know in a form that machines can reason over, they enable the kind of cross-cutting insights and rapid information access that is essential for modern scientific discovery. Catalysis, being a data- and knowledge-intensive discipline, stands to benefit immensely from this paradigm. In the coming years, we expect knowledge graphs, enhanced by large language models and continuous data streams, to help researchers discover novel catalysts faster, optimize processes with fewer experiments, and unify insights across different subfields of catalysis. This convergence of human knowledge and artificial intelligence, embodied in knowledge graphs, promises to accelerate innovation and drive the field of heterogeneous catalysis into a new era of discovery.

CRediT Authorship Contribution Statement

Raúl Díaz: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. Hongliang Xin: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

R.Diaz acknowledges support from the Full Bridge Fellowship for enabling the research stay at Virginia Tech. H.Xin acknowledge the financial support from the US Department of Energy, Office of Basic Energy Sciences under contract no. DE-SC0023323 and from the National Science Foundation through the grant 2245402 from CBET Catalysis and CDS&E programs.

References

[1] A.S. Behr, D. Chernenko, D. Koßmann, A. Neyyathala, S. Hanf, S.A. Schunk, N. Kockmann, Generating knowledge graphs through text mining of catalysis research related literature, *Catal. Sci. Technol.* 14 (19) (2024) 5699–5713.
[2] Z.Y. Zhang, S.M. Ma, S.S. Zheng, Z.W. Nie, B.X. Wang, K. Lei, S.N. Li, F. Pan, Semantic knowledge graph as a companion for catalyst recommendation, *Natl. Sci. Open* 3 (2) (2024) 20230040.
[3] A.L. Lamprecht, L. Garcia, M. Kuzak, C. Martinez, R. Arcila, E. Martin Del Pico, V. Dominguez Del Angel, S. van de Sandt, J. Ison, P.A. Martinez, P. McQuilton, A. Valencia, J. Harrow, F. Psomopoulos, J.L. Gelpi, N. Chue Hong, C. Goble, S. Capella-Gutierrez, Towards FAIR principles for research software, *Data Sci.* 3 (1) (2020) 37–59.
[4] C.H. Wang, Y.Q. Yang, J.S. Song, X.F. Nan, Research progresses and applications of knowledge graph embedding technique in chemistry, *J. Chem. Inf. Model.* 64 (19) (2024) 7189–7213.
[5] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier, A.-C.N. Ngomo, A. Polleres, S.M. Rashid, A. Rula, L. Schmelzeisen, Knowledge graphs, 2020, https://doi.org/10.1145/3447772.
[6] C.Y. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: opportunities and challenges, *Artif. Intell. Rev.* 56 (2023) 13071–13102.
[7] S. Auer, V. Kovtun, M. Prinz, A. Kasprzik, M. Stocker, M.E. Vidal, Towards a knowledge graph for science, in: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. Novi Sad Serbia., ACM, 2018.
[8] L. Ehrlinger, W. Wöß, Towards a definition of knowledge graphs, in: Proceedings of International Conference on Semantic Systems, Leipzig, Germany, 2016.
[9] Ontologies. [cited 18 Feb 2025]. Available: https://nfdi4cat.org/en/services/ontology-collection/.
[10] M. Hofer, D. Obraczka, A. Saeedi, H. Köpcke, E. Rahm, Construction of knowledge graphs: current state and challenges, *Information* 15 (8) (2024) 509.
[11] N. Kertkeidkachorn, R. Ichise, T2KG: an end-to-end system for creating knowledge graph from unstructured text. https://cdn.aaai.org/ocs/ws/ws0328/15129-68427-1-PB.pdf.
[12] D. Buscaldi, D. Dessì, E. Motta, F. Osborne, D. Reforgiato Recupero, Mining Scholarly Publications for Scientific Knowledge Graph Construction. In: the Semantic Web: ESWC 2019 Satellite Events 12, Springer International Publishing, 2019.
[13] Y. Gao, L.D. Wang, X.Q. Chen, Y. Du, B. Wang, Revisiting electrocatalyst design by a knowledge graph of Cu-based catalysts for CO_2 reduction, *ACS Catal.* 13 (13) (2023) 8525–8534.
[14] A. Oarga, M. Hart, A.M. Bran, M. Lederbauer, P. Schwaller, Scientific knowledge graph and ontology generation using open large language models. Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges, 2024. https://openreview.net/pdf?id=kHsfqBhZjC.
[15] Y. Zhang, C. Wang, M. Soukaseum, D.G. Vlachos, H. Fang, Unleashing the power of knowledge extraction from scientific literature in catalysis, *J. Chem. Inf. Model.* 62 (14) (2022) 3316–3330.
[16] K.T. Winther, M.J. Hoffmann, J.R. Boes, O. Mamun, M. Bajdich, T. Bligaard, Catalysis-Hub.org, an open electronic structure database for surface reactions, *Sci. Data* 6 (1) (2019) 75.
[17] L. Pascazio, S. Rihm, A. Naseri, S. Mosbach, J. Akroyd, M. Kraft, Chemical species ontology for data integration and knowledge discovery, *J. Chem. Inf. Model.* 63 (21) (2023) 6569–6586.
[18] V. Lopez, L. Hoang, M. Martinez-Galindo, R. Fernández-Díaz, M.L. Sbodio, R. Ordonez-Hurtado, M. Zayats, N. Mulligan, J. Bettencourt-Silva, Enhancing foundation models for scientific discovery *via* multimodal knowledge graph representations, *J. Web Semant.* 84 (2025) 100845.
[19] M.J. Buehler, Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning, *Machine Learning: Science and Technology* 5 (3) (2024) 035083.
[20] X. Wang, B.Y. Meng, H. Chen, Y. Meng, K. Lv, W.W. Zhu, TIVA-KG: a multimodal knowledge graph with text, image, video and audio, in: Proceedings of the 31st ACM International Conference on Multimedia, ACM, Ottawa ON Canada, 2023.
[21] V. Venugopal, E. Olivetti, MatKG: an autonomously generated knowledge graph in material science, *Sci. Data* 11 (1) (2024) 217.
[22] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2013. https://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf.
[23] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in: Proceedings of the 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research 48, PMLR, 2016, pp. 2071–2080, arXiv:1606.06357 [cs.AI].
[24] A.S. Behr, H. Borgelt, N. Kockmann, Ontologies4Cat: investigating the landscape of ontologies for catalysis research data management, *J. Cheminf.* 16 (1) (2024) 16.
[25] A.S. Behr, H. Borgelt, N. Kockmann, Reac4Cat-ontology: harnessing the power of ontological description logic in catalysis research as aPractical approach to knowledge inferences, *Datenbank-Spektrum* 24 (2) (2024) 139–150.
[26] N. Huskova, Y. Dikova, T. Petrenko, T. Bönisch, Improvement of data and metadata quality in catalysis research: a use case-driven methodology, *Catal. Today* 446 (2025) 115111.
[27] X. Wang, V. Hu, X.C. Song, S. Garg, J.F. Xiao, J.W. Han, ChemNER: fine-grained chemistry named entity recognition with ontology-guided distant supervision, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta cana, Dominican Republic, USAACL, Stroudsburg, PA, 2021.
[28] A.S. Behr, M. Völkenrath, N. Kockmann, Ontology extension with NLP-based concept extraction for domain experts in catalytic sciences, *Knowl. Inf. Syst.* 65 (12) (2023) 5503–5522.
[29] W.J. Liu, P. Zhou, Z. Zhao, Z.R. Wang, Q. Ju, H.T. Deng, P. Wang, K-BERT: enabling language representation with knowledge graph, *Proc. AAAI Conf. Artif. Intell.* 34 (3) (2020) 2901–2908.
[30] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K.A. Persson, G. Ceder, A. Jain, Named entity recognition and normalization applied to large-scale information extraction from the materials science literature, *J. Chem. Inf. Model.* 59 (9) (2019) 3692–3702.
[31] R. Tran, J. Lan, M. Shuaibi, B.M. Wood, S. Goyal, A. Das, J. Heras-Domingo, A. Kolluru, A. Rizvi, N. Shoghi, A. Sriram, F. Therrien, J. Abed, O. Voznyy, E.H. Sargent, Z. Ulissi, C.L. Zitnick, The open catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts, *ACS Catal.* 13 (5) (2023) 3066–3084.
[32] S.L. Scott, T.B. Gunnoe, P. Fornasiero, C.M. Crudden, To err is human; to reproduce takes time, *ACS Catal.* 12 (6) (2022) 3644–3650.
[33] ChEBI. [cited 22 Feb 2025]. https://www.ebi.ac.uk/chebi/.
[34] RXNO - ontology lookup service. [cited 22 Feb 2025]. Available: https://www.ebi.ac.uk/ols4/ontologies/rxno.
[35] L. Takahashi, K. Takahashi, Visualizing scientists' cognitive representation of materials data through the application of ontology, *J. Phys. Chem. Lett.* 10 (23) (2019) 7482–7491.
[36] I. Beltagy, K. Lo, A. Cohan, SciBERT, A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural

Language Processing and the 9th International Joint Conference on Natural Language Processing (Volume 1, pp. 3615 – 3620), Association for Computational Linguistics, 2019, https://doi.org/10.18653/v1/D19-1371 arXiv: 1903.10676 [cs.CL], http://arxiv.org/abs/1903.10676.
[37] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, SCICERO: a deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain, *Knowl. Base Syst.* 258 (2022) 109945.
[38] M.D.L. Tosi, J.C. dos Reis, SciKGraph: a knowledge graph approach to structure a scientific field, *J. Informetr.* 15 (1) (2021) 101109.
[39] Y.W. Zhang, F.Y. Chen, Z.Y. Liu, Y.Z. Ju, D.L. Cui, J.Y. Zhu, X. Jiang, X. Guo, J. He, L. Zhang, X.T. Zhang, Y.J. Su, A materials terminology knowledge graph automatically constructed from text corpus, *Sci. Data* 11 (1) (2024) 600.
[40] L. Wang, J. Ren, B. Xu, J. Li, W. Luo, F. Xia, MODEL: Motif-based deep feature learning for link prediction, 2020, https://doi.org/10.1109/TCSS.2019.2962819.
[41] L. Yao, C. Mao, Y. Luo, KG-BERT: BERT for knowledge graph completion, arXiv [cs.CL], 2019. Available: http://arxiv.org/abs/1909.03193.
[42] A. Khan, Knowledge graphs querying, *SIGMOD Rec* 52 (2) (2023) 18–29.
[43] B.Z. Liu, X. Wang, P.K. Liu, S.Z. Li, Q. Fu, Y.P. Chai, UniKG: a unified interoperable knowledge graph database System, in: 2021 IEEE 37th International Conference on Data Engineering (ICDE), IEEE, Chania, Greece, 2021.
[44] Comparing cypher with SQL. In: Neo4j Graph Data Platform [Internet]. [cited 22 Feb 2025], https://neo4j.com/docs/getting-started/cypher-intro/cypher-sql/.
[45] Text2Cypher: Bridging natural language and graph databases. [cited 22 Feb 2025], https://arxiv.org/html/2412.10064v1.
[46] Z. Zhong, L.Q. Zhong, Z.Z. Sun, Q.Y. Jin, Z.C. Qin, X.F. Zhang, SyntheT2C: generating synthetic data for fine-tuning large language models on the Text2Cypher task, arXiv [cs.AI], 2024. Available: http://arxiv.org/abs/2406.10710.
[47] B.B. Kömeçoğlu, B. Yılmaz, Knowledge-augmented large language model prompting for cypher query construction, in: 2024 9th International Conference on Computer Science and Engineering (UBMK), Antalya, Turkiye, IEEE, 2024.
[48] M.A. Rodriguez, The Gremlin graph traversal machine and language, arXiv [cs.DB], 2015. https://doi.org/10.1145/2815072.2815073.
[49] Y. Liang, T. Xie, G. Peng, Z. Huang, Y. Lan, W. Qian, NAT-NL2GQL: a novel multi-agent framework for translating natural language to graph query language, arXiv [cs.CL], 2024. Available: http://arxiv.org/abs/2412.10434.
[50] G. Vargas-Solar, K. Dao, J.-A. Espinosa-Oviedo, Translating data science queries from natural language into graph analytics queries using NLDS-QL, in: CEUR Workshop Proceedings, EDBT-ICDT, Ioannina, Greece, 2023.
[51] L. Nie, S. Cao, J.X. Shi, J. Sun, Q.W. Tian, L. Hou, J.Z. Li, J.D. Zhai, GraphQ IR: unifying the semantic parsing of graph query languages with one intermediate representation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, UAE, 2022.
[52] P. Ghiasnezhad Omran, K. Taylor, S. Rodríguez Méndez, A. Haller, Learning SHACL shapes from knowledge graphs, *Semant. Web* 14 (1)101–121.
[53] J.Y. Chen, F. Lécué, J.Z. Pan, S.M. Deng, H.J. Chen, Knowledge graph embeddings for dealing with concept drift in machine learning, *J. Web Semant.* 67 (2021) 100625.
[54] D. Garay-Ruiz, C. Bo, Chemical reaction network knowledge graphs: the OntoRXN ontology, *J. Cheminf.* 14 (1) (2022) 29.
[55] A. Kondinski, P. Rutkevych, L. Pascazio, D.N. Tran, F. Farazi, S. Ganguly, M. Kraft, Knowledge graph representation of zeolitic crystalline materials, *Digit. Discov.* 3 (10) (2024) 2070–2084.
[56] H. Han, Y. Wang, H. Shomer, K. Guo, J. Ding, Y. Lei, M. Halappanavar, R.A. Rossi, S. Mukherjee, X. Tang, Q. He, Z. Hua, B. Long, T. Zhao, N. Shah, A. Javari, Y. Xia, J. Tang, Retrieval-augmented generation with graphs (GraphRAG), arXiv [cs.IR], 2024. Available: http://arxiv.org/abs/2501.00309.
[57] G. Perković, A. Drobnjak, I. Botički, Hallucinations in LLMs: understanding and addressing challenges, in: 2024 47th MIPRO ICT and Electronics Convention (MIPRO). Opatija, Croatia, IEEE, 2024.
[58] G. Agrawal, T. Kumarage, Z. Alghamdi, H. Liu, Can knowledge graphs reduce hallucinations in LLMs? : a survey, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024), Mexico City, Mexico, Association for Computational Linguistics, 2023.
[59] H. Abu-Rasheed, C. Weber, M. Fathi, Knowledge graphs as context sources for LLM-based explanations of learning recommendations, in: 2024 IEEE Global Engineering Education Conference (EDUCON). Kos Island, Greece, IEEE, 2024.
[60] Y.X. Dong, S. Wang, H.Y. Zheng, J.J. Chen, Z.H. Zhang, C.H. Wang, Advanced RAG models with graph structures: optimizing complex knowledge reasoning and text generation, in: 2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC), IEEE, Wuhan, China, 2024.
[61] Β. Τσακαλάκης, Augmentation of large language model capabilities with knowledge graphs, University of Western Attica, 2024, https://doi.org/10.26265/POLYNOE-5908.
[62] A. Saleh, G. Tur, Y. Saygín, SG-RAG: multi-hop question answering with large language models through knowledge graphs. Proceedings of the 7th International Conference on Natural Language and Speech Processing, Association for Computational Linguistics. ICNLSP, Trento, Italy, 2024.
[63] A. Kau, X. He, A. Nambissan, A. Astudillo, H. Yin, A. Aryani, Combining knowledge graphs and large language models, arXiv [cs.CL], 2024. Available: http://arxiv.org/abs/2407.06564.
[64] C. Feng, X. Zhang, Z. Fei, Knowledge solver: teaching LLMs to search for domain knowledge from knowledge graphs., arXiv [cs.CL], 2023. Available: http://arxiv.org/abs/2309.03118.
[65] K.K.Y. Ng, I. Matsuba, P.C. Zhang, RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations, *Nejm Ai* 2 (1) (2025) 2400380.
[66] S. Schimmler, T. Bönisch, M.T. Horsch, T. Petrenko, B. Schembera, V. Kushnarenko, NFDI4Cat: local and overarching data infrastructures, in: E-Science-Tage 2021: Share Your Research Data, 2022, https://doi.org/10.11588/heibooks.979.c137.
[67] C. Wulf, P. Matthias Beller, D.I. Thomas Boenisch, P. Olaf Deutschmann, D. Schirin Hanf, P. Norbert Kockmann, D.I. Ralph Kraehnert, P. Mehtap Oezaslan, D. Stefan Palkovits, D. Sonja Schimmler, D.S.A. Schunk, P. Kurt Wagemann, D. David Linke, A unified research data infrastructure for catalysis research—challenges and concepts, *ChemCatChem* 13 (14) (2021) 3223–3236.
[68] K.M. Jablonka, L. Patiny, B. Smit, Making the collective knowledge of chemistry open and machine actionable, *Nat. Chem.* 14 (4) (2022) 365–376.
[69] S. Ji, S. Pan, E. Cambria, P. Marttinen, P.S. Yu, A survey on knowledge graphs: representation, acquisition and applications, *IEEE Transactions on Neural Networks and Learning Systems* 33 (2) 494–514.