

RESEARCH ARTICLE

Natural language access point to digital metal–organic polyhedra chemistry in *The World Avatar*

Simon D. Rihm¹ , Dan N. Tran² , Aleksandar Kondinski¹ , Laura Pascazio² , Fabio Saluz^{1,3}, Xinhong Deng², Sebastian Mosbach^{1,2,4} , Jethro Akroyd^{1,2,4}  and Markus Kraft^{1,2,4} 

¹Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK

²CARES, Cambridge Centre for Advanced Research and Education in Singapore, Singapore, Singapore

³Department of Mechanical and Process Engineering, ETH Zurich, Zurich, Switzerland

⁴CMCL, Cambridge, UK

Corresponding author: Markus Kraft; Email: mk306@cam.ac.uk

Received: 10 September 2024; **Revised:** 28 January 2025; **Accepted:** 11 February 2025

Keywords: dynamic knowledge graphs; metal–organic polyhedra; question-answering systems; retrieval-augmented generation

Abstract

Metal–organic polyhedra (MOPs) are discrete, porous metal–organic assemblies known for their wide-ranging applications in separation, drug delivery, and catalysis. As part of *The World Avatar* (TWA) project—a universal and interoperable knowledge model—we have previously systematized known MOPs and expanded the explorable MOP space with novel targets. Although these data are available via a complex query language, a more user-friendly interface is desirable to enhance accessibility. To address a similar challenge in other chemistry domains, the natural language question-answering system “Marie” has been developed; however, its scalability is limited due to its reliance on supervised fine-tuning, which hinders its adaptability to new knowledge domains. In this article, we introduce an enhanced database of MOPs and a first-of-its-kind question-answering system tailored for MOP chemistry. By augmenting TWA’s MOP database with geometry data, we enable the visualization of not just empirically verified MOP structures but also machine-predicted ones. In addition, we renovated Marie’s semantic parser to adopt in-context few-shot learning, allowing seamless interaction with TWA’s extensive MOP repository. These advancements significantly improve the accessibility and versatility of TWA, marking an important step toward accelerating and automating the development of reticular materials with the aid of digital assistants.

Impact statement

Molecular engineering based on the modular reuse of chemical building units is a powerful methodology for the rapid development of new advanced materials relevant to sustainability, energy transition, and life science. Metal–organic polyhedra (MOPs) are an emerging class of rationally designed advanced materials demanding increased digitalization efforts for automated exploration of their chemical space and allocation of candidates for tailor-made applications. Building on our previous effort in the digitalization of MOPs, we introduce a question-answering (QA) system for MOPs based on a knowledge graph-retrieval-augmented generation (KG-RAG) system. This system provides a user-friendly exploration of MOP chemistry and offers rapid adaptation to new data and domains. The developed methodology provides a solid platform supporting domain experts and shows strong potential as a blueprint for developing adaptable QA systems for specialized knowledge areas.

1. Introduction

Metal–organic polyhedra (MOPs) represent a class of materials characterized by their self-assembled, cage-like discrete nanomolecular architecture constructed from metal-based and organic building blocks (Gosselin et al., 2020; Lee et al., 2021). Considering their network-like discrete assembly topologies combining internal cavitation and a plethora of organic and inorganic cluster functionalities, MOPs are typically considered a subset of reticular materials with promising applications in catalysis, separation, and energy technologies (Perry Iv et al., 2009; Vardhan et al., 2016). However, in light of the chemical space that emerges from brute combinatorial derivation of new hypothetical reticular structures, past years in these domains have noted increased interest in the development of data-driven technologies for material discovery (Chong et al., 2020; Rosen et al., 2021; Kang and Kim, 2024), selection (Guan et al., 2022; Li et al., 2022), and synthesis (Luo et al., 2022), as well as the development of data infrastructures to support these tasks, including data cataloging (Moghadam et al., 2017), mining (Bai et al., 2024), and accessing (Kang and Kim, 2024).

Considering the relatively smaller sample size of MOPs compared to its extended metal–organic framework (MOF) analogs, developing data-driven digital tools for MOP discovery has remained challenging because big data-driven methods for MOFs cannot be easily extended to MOPs. In this regard, our group has developed new formal and semantic approaches to describe MOPs, including custom-designed inductive reasoning algorithms for the discovery of new structures. Thus, following a careful development of a knowledge model for MOP chemistry, we have instantiated 151 experimentally described MOPs, and based on them, our reasoning algorithm designed 1,418 new MOP instances that are rationally designed from existing building units, following expert-like patterns of molecular engineering (Kondinski et al., 2022). The overall research has been originally contextualized within our The World Avatar (TWA) digital infrastructure, which adopts Semantic Web principles to bridge the gap between digital and physical realms.

Access to chemical information for a long time came with the requirement of some forms of cheminformatics knowledge (Gasteiger, 2016). In a similar line, the acquisition of chemical information instantiated in the form of a knowledge graph (KG) typically requires the use of querying tools such as SPARQL (Quilitz and Leser, 2008; Pérez et al., 2009), which may appear unintuitive and even cumbersome to new users, thus limiting the accessibility of chemical information. Despite our initial success in describing MOP chemistry via a KG model and in developing agents for digital exploration of its chemical space, semantic query tools often appear as a barrier for experimental chemists who may want to rapidly leverage insights from our work toward developing new materials. Noticing similar experiences along different chemistry domains, we have been motivated to build tools that integrate semantic querying with natural language processing (NLP), enabling virtually any user with access to the Internet to be able to query verified and expert-derived knowledge models simply via prompting. In this regard, we have developed dedicated easy-to-use tools to navigate complex ideas and concepts that are either niche in nature or not fully in the public domain and, therefore, not accessible via traditional search engines or general-purpose large language models (LLMs). One such interface is Marie, a natural language question-answering (QA) system for chemistry. Previously designed to facilitate access to data in the domains of combustion kinetics and crystalline zeolitic materials, Marie has demonstrated the potential of NLP-driven tools to help human users navigate complex knowledge bases (Pascazio et al., 2024; Kondinski et al., 2024b). However, Marie's reliance on supervised fine-tuning in developing its semantic parser curtails its scalability. In TWA's dynamic environment, new knowledge domains are continually introduced and extended, making repeated retraining of Marie's semantic parser necessary, which is not only resource-costly but also risks catastrophic forgetting (Hadsell et al., 2020). Finally, specific to reticular chemistry is the problem of understanding complex information and structures, which calls for visualization.

The purpose of this article is to present an enriched knowledge base and an enhanced QA system tailored for digital engagement with MOP chemistry. TWA's MOP domain is restructured and augmented with geometry data for new MOP instances deduced in our previous work, allowing the visualization of not just empirically verified MOP structures but also those predicted by our “MOP Discovery” agent.

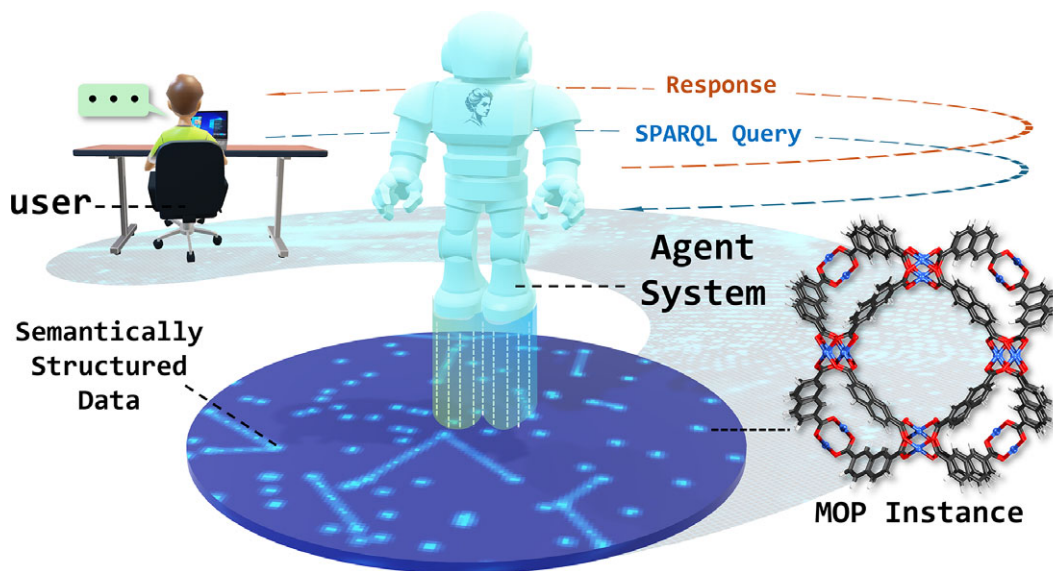


Figure 1. Illustration of TWA's digital infrastructure that enables the retrieval of structured and validated MOP data via natural language requests.

Additionally, we update Marie's semantic parser to adopt the approach of few-shot in-context learning with demonstration retrieval, which enables more agile incorporation of new domains and acceleration of development cycles. The presented approach enables the fast and economical creation of reliable QA systems for specialized fields.

2. Background

In this section, we first introduce the TWA knowledge ecosystem and its application to the chemistry domain, particularly MOPs. We then give a short overview of current trends in QA systems in related domains.

2.1. TWA – A virtual hub for digital chemistry

TWA is a pioneering project that creates a universal digital twin of the real world, building on the early potential of the Semantic Web to enhance cheminformatics and broader chemical applications (Berners-Lee et al., 2001; Taylor et al., 2006; Murray-Rust, 2008). The Semantic Web, an evolution of the World Wide Web, addresses the gap between human-readable documents and machine-readable data by prioritizing semantics over presentation. It builds on layers of technologies standardized by the World Wide Web Consortium, starting with raw data expressed using Unicode and uniquely identified through Internationalized Resource Identifiers (IRIs). Information is represented through the Resource Description Framework (RDF) in triples of subject, predicate, and object. Ontologies formalize this knowledge by defining the structure of instances within a KG, enabling querying via SPARQL. The Semantic Web's core principle is Linked Data, which promotes accessibility and integration across different sources.

Initially conceptualized in 2010, TWA has evolved from the representation of a single chemical industry park on Jurong Island (Singapore) into an unrestricted world model capable of integrating a range of phenomena from the atom to multiscale features impacting environment, climate, and population health (Akroyd et al., 2021), including power and heat network optimizations for CO₂ savings, environmental monitoring, and cross-domain climate resilience planning through the Climate Resilience

Demonstrator (Mosbach et al., 2020; Akroyd et al., 2021; Akroyd et al., 2022). TWA operates on the principles of the Semantic Web and adheres to the FAIR guidelines, to ensure that all data are findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). It integrates software agents that manage information flows, interface with computational models, and continuously enhance TWA's KGs with new data (Zhou et al., 2019; Akroyd et al., 2021).

The digital chemistry in TWA is aligned and structured around foundational ontologies such as OntoSpecies, OntoKin, OntoCompChem, and OntoPESScan, facilitating a comprehensive mapping of chemical species, reaction mechanisms, and quantum chemistry calculations, respectively (Kondinski et al., 2024a, Kondinski et al., 2023). This framework supports detailed data relationships and enhances interoperability, enabling multifaceted data usage and reducing ambiguities (Farazi et al., 2020; Akroyd et al., 2021). Additionally, computational agents in TWA perform complex tasks such as calibrating kinetic mechanisms and automating discovery processes (Kondinski et al., 2024a), exemplified by the development of novel MOPs (Kondinski et al., 2022) which, among a variety of applications, can be used for photocatalytic CO₂ reduction (Ghosh et al., 2022; Adeola et al., 2024).

The OntoMOPs ontology is designed to provide and enrich semantic relationships between MOPs, chemical building units (CBUs), and assembly models (Kondinski et al., 2022). This ontology enables advanced query capabilities for professionals engaged in the modeling and preparation of MOPs, supporting informed decision-making with detailed information on the construction and functionalities of these materials. OntoMOPs links manually curated MOP instances to crucial metadata such as molecular mass, charge, formulas, and provenance information like DOIs and CCDC numbers for precise identification and cross-referencing with crystalline databases. Additionally, the assembly model concept details how different generic building units (GBUs) contribute to the formation of specific polyhedral shapes recognized in reticular chemistry, such as tetrahedra and octahedra, while the CBU concept models chemical functionalities and binding sites necessary for MOP formation.

2.2. Trends in knowledge-intensive chemistry QA systems

In recent years, the field of NLP has experienced a remarkable rise in popularity, primarily driven by the accessible deployment of LLMs. The advent of LLMs is marked by their ability to tackle diverse knowledge-intensive tasks that range from the humanities to the sciences, including chemistry (OpenAI et al., 2023). However, despite their impressive performance on standardized examinations, general-purpose LLMs like GPT-4 often struggle with more advanced and specialized requests, revealing their lack of in-depth understanding of the subject matter (Guo et al., 2024). While fine-tuning is a possible remedy (Zhang et al., 2024), a significant challenge remains: these models are inherently limited by the scope and recency of their training data, rendering them inadequate for querying up-to-date information or applying the latest research knowledge without undergoing further retraining.

In-context few-shot learning is a technique where LLMs are provided with a few input–output examples during testing to align their behavior with user expectations without updating their weights. In-context means the model processes and utilizes examples provided within the same input context, rather than relying on weight updates or external training. First observed in GPT-3 during model scaling experiments (Brown et al., 2020), in-context learning enables models to perform tasks based on demonstrations alone, and recent research suggests that smaller models can also be trained for this capability (Min et al., 2022). Few-shot means the model requires only a small number of examples to generalize and perform a task effectively. This helps reduce cost by eliminating the need for extensive fine-tuning or retraining, leveraging existing model capabilities to adapt to new tasks efficiently.

In the realm of chemistry, LLMs are increasingly utilized for a variety of tasks, including data processing, engineering, inference, and augmentation, in conjunction with various computational tools (Guo et al., 2023; Jablonka et al., 2024; M. Bran A et al., 2024). Despite these advancements, concerns about the explainability of these technologies continue to persist (Gallegos et al., 2024), prompting further research into integrating LLMs with semantic technologies. QA systems have historically leveraged external knowledge bases, particularly through KG-based QA systems. These are designed to retrieve and

reason over structured data from KGs to deliver precise and fact-based answers (Zhang et al., 2018; Kim et al., 2023).

The emergence of retrieval-augmented generation (RAG) systems has taken this a step further by combining the reasoning capabilities of LLMs with the retrieval of up-to-date information from external sources (Lewis et al., 2020). This allows RAG systems to generate more contextually relevant and accurate responses (Zheng et al., 2023; Kang and Kim, 2024). Using KGs as the foundation for information retrieval, recent studies have shown the great promise of KG-RAG to reliably handle knowledge-intensive and cognitive tasks (Sanmartin, 2024).

Another challenge for knowledge-intensive QA systems is handling of private, niche, or proprietary data, which is encountered in both industrial contexts and academic research. This necessitates a flexible QA system capable of integrating various data sources and domains while also allowing for the dynamic inclusion of new information. LLMs' strong generalization ability and versatility are key to addressing these dual goals. For example, ChemCrow (M. Bran A et al., 2024) is a tool-calling agent capable of incorporating information pulled from a mixture of public and private data sources and computational tools, including the PubChem database and the RoboRXN platform by IBM Research (IBM, 2021). It does so by employing an LLM pretrained for the tool-calling task to orchestrate when to use which external tool and how to process and combine the results to form a coherent response (Schick et al., 2023).

Similarly, Marie is capable of querying across various domains and accessing information from distributed data sources within the fields of combustion kinetics and crystalline zeolitic materials (Pascasio et al., 2024; Kondinski et al., 2024b). However, the previous version of Marie relies on supervised fine-tuning for its semantic parser, which necessitates retraining whenever it needs to integrate with a new knowledge domain in TWA. In contrast, the in-context learning capability of LLMs (Brown et al., 2020) offers a promising approach to expanding Marie's coverage across TWA's domains without the need for retraining. This capability allows LLMs to perform tasks based solely on task demonstrations provided at test time, without updating model weights—particularly, if coupled with advanced entity linking algorithms (Nie et al., 2024).

3. Methodology and Implementation

In this section, we detail the methods developed for our natural language access point for MOP chemistry. We begin by outlining the refinements and extensions made to the existing knowledge model within TWA. Following this, we describe the integration of Marie into the MOP chemistry domain and the substantial improvements to its architecture.

3.1. Updates to *OntoMOPs*

To include MOP knowledge in our chemistry QA system Marie and extend it for better user interactivity, the MOP knowledge base needed to be restructured and extended first. As a first step, we made adjustments to the original *OntoMOPs* ontology to improve robustness and ease of querying. The changes concern two main aspects: the storage of geometry data and the elimination of potential data redundancy. In a second step, the MOP KG was enriched with 370 new geometries of machine-predicted MOPs in addition to the 151 existing geometries of previously synthesized MOPs. These molecular geometries were deduced from information represented in the KG and will help researchers to better visualize these structures and screen possible synthesis candidates.

The updated ontology is shown in Figure 2. Its core concepts form a rectangle: MOPs can be classified by their geometric assembly models made up of distinct GBUs as which a variety of CBUs can function (Kondinski et al., 2022). These four core concepts now provide access to a range of geometric and molecular properties alike.

In the original implementation, geometry data of MOPs and CBUs are provided as XYZ documents or XYZ-formatted strings in the KG. A potential limitation of this method is that a string can, in principle, exceed the stringent length limits imposed by the KG engine—for example, in the case of a very large

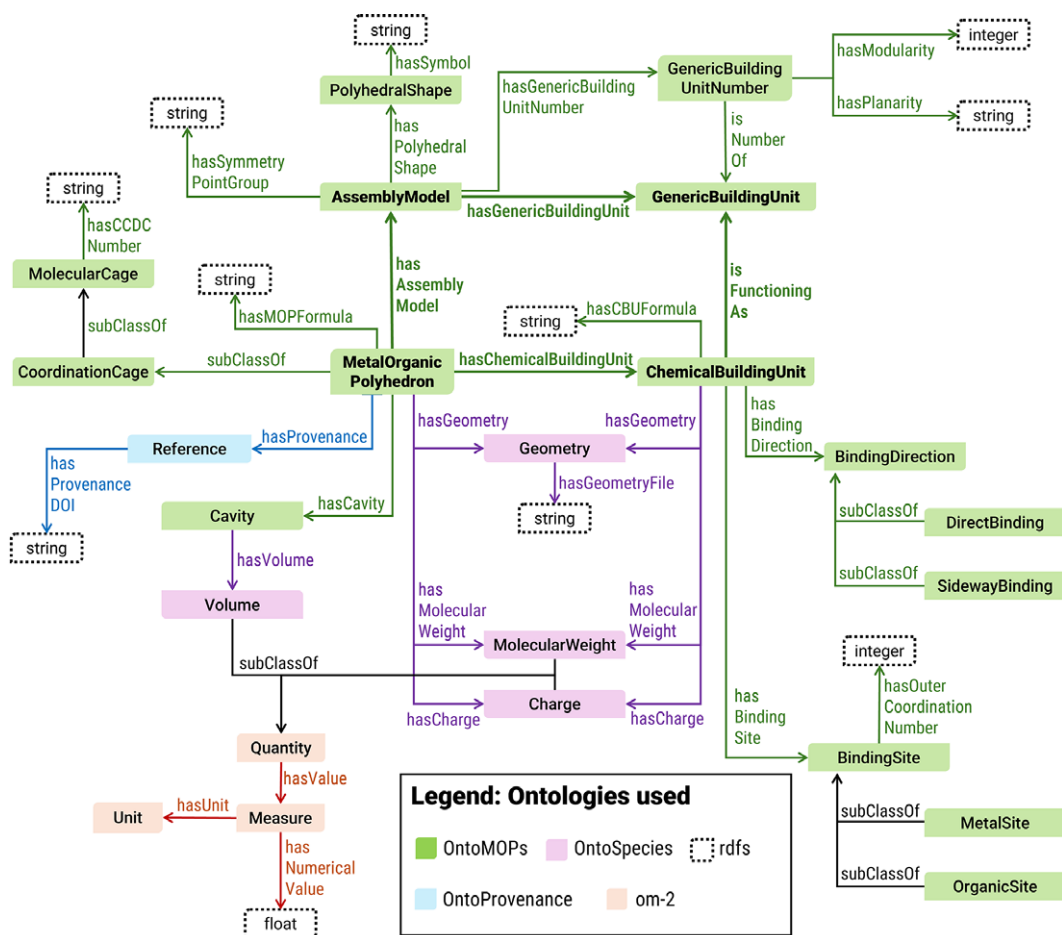


Figure 2. Illustration of the terminological component (TBox) of the MOP chemistry domain in TWA and its related ontologies. Core concepts are shown in bold.

chemical superstructure. An alternative implementation entails the instantiation of every atom in an MOP or CBU structure and linking these atoms to an intermediate Geometry node, which is then connected to an MOP or CBU instance via the `hasGeometry` predicate, as done in the `OntoSpecies` domain of TWA (Pascasio et al., 2023). However, doing so for large MOP structures could introduce an overwhelming number of triples and, consequently, may slow down KG operations. In this work, we make a compromise between limiting the number of instantiated triples and avoiding storing long strings directly in the KG by moving the storage of the geometry data to XYZ files on disk. These files are hosted on a web server so that they are accessible on the Internet via URLs, which are discoverable in TWA through `hasGeometryFile` links to Geometry nodes.

The original assertion component created redundancy in the assignment of IRIs for instances of assembly models and GBUs, necessitating the postprocessing of aggregate queries. In the new implementation, redundant entries were merged or removed, allowing for simpler traversal of the KG without lengthy queries. On a terminological level, our effort to increase interoperability and overlap between chemical TWA ontologies, particularly with the renewed implementation of `OntoSpecies` (Pascasio et al., 2023), has facilitated the reuse of general-purpose concepts. As illustrated in Figure 2, this reuse covers many concepts related to shared molecular properties and literature provenance. This way, we intensify the interlinkedness of TWA and simplify agentic data curation and the training of Marie modules.

3.2. The architecture of Marie TWA

A QA system for TWA is not only required to map user intents to a machine-readable format accurately but it must also identify the correct data repository that contains the requested information. The latter stipulation arises from TWA's compartmentalization of its data into distinct triplestores to allow domain experts to own and manage them independently. Earlier versions of Marie struggle with the dynamic nature of TWA, as its semantic parser relied on supervised fine-tuning (Tran et al., 2024), requiring resource-intensive retraining. This limitation not only impedes Marie's scalability but also poses the risk of catastrophic forgetting (Hadsell et al., 2020). In contrast, the current version of Marie is designed with a more agile and adaptable architecture, ensuring continued support for existing chemical domains within TWA while seamlessly extending coverage to new domains, such as OntoMOPs.

To achieve this, we set up a KG-RAG system based on a modular architecture and adapted in-context few-shot learning methods to it. As depicted in Figure 3, Marie's online workflow comprises three main components:

1. The input rewriter aligns all physical quantities mentioned in the input question to the unit systems in our knowledge base.
2. The semantic parser jointly generates the logical form of a SPARQL query, detects the surface forms of entities present in the input question, and determines the triplestore to execute the query.
3. The response generator presents the structured SPARQL response and LLM-generated styled text, accompanied by visualization of the 3D structures of any invoked chemical entities.

Both the quantity recognizer and semantic parser are powered by LLMs prompted with in-context examples; the exact structure of these prompts is available in the [Supplementary Materials](#). While the LLM prompt for the physical quantity recognizer is fixed, the semantic parser dynamically adapts to the input question by incorporating only the $k_{\text{demonstrations}}$ most relevant semantic parsing demonstrations and $k_{\text{KG_relations}}$ most relevant KG relations. This approach is key to Marie's rapid integration with new knowledge domains because only a small number of semantic parsing demonstrations and KG relations need to be prepared, unlike the relatively larger training dataset required for supervised fine-tuning. Additionally, the on-demand retrieval of the most relevant elements for prompt construction ensures that the prompt is as compact as possible to fit within the context window of common LLMs while also saving

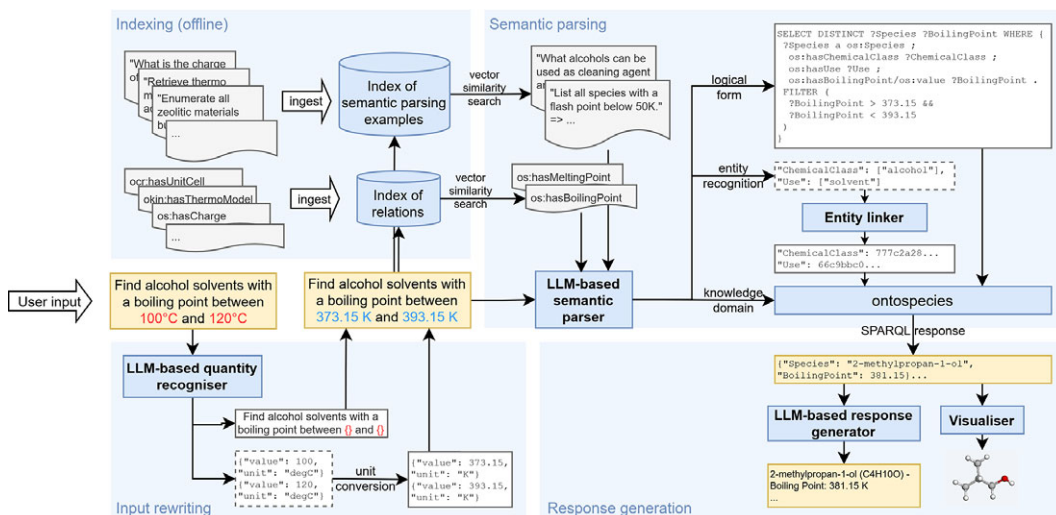


Figure 3. Architecture of "Marie," comprising one offline indexing stage and three online stages, namely input rewriting, semantic parsing, and response generation.

processing time. Relevance is measured by the cosine similarity of their Sentence-BERT embeddings (Reimers and Gurevych, 2019) using the all-mpnet-base-v2 variant. We use OpenAI's gpt-4o-mini-2024-07-18 model for in-context learning and the Redis Community Edition for all retrieval needs. These techniques allow us to construct Marie as a modular KG-RAG system that leverages pretrained LLMs without requiring fine-tuning or retraining itself. This not only drives down the cost of operating Marie as well as adding more subjects but also ensures response accuracy.

Marie's entity linking component uses different strategies depending on the entity class:

- Inverted index lookup for entities with well-defined labels, for example, chemical species with their IUPAC names, molecular formulas, and SMILES strings.
- Semantic search for entities that represent concepts or categories, for example, chemical classifications.
- RDF subgraph matching for more complex entities that are conceptually defined by their relationships with other entities, for example, assembly models composed of GBUs (Kondinski et al., 2022).

Compared to earlier versions, this multistrategy approach has been refined to accommodate the diverse and growing range of entities within TWA, particularly the complex entities in the MOP chemistry domain. In the [Supplementary Materials](#), we provide a summary of entity linking strategies and an illustration of RDF subgraph matching.

The response generation component in Marie has been enhanced to provide more comprehensive and user-friendly outputs. Marie's structured output is presented in both JSON and tabular format, allowing users to view the raw SPARQL response in JSON and the formatted version in a table. On top of this, the natural language text generated by an LLM explains the results in a more accessible manner. A major update in the current version is the visualization of intricate chemical structures like MOPs; this is done using the library 3Dmol.js (Rego and Koes, 2014). This feature not only broadens the utility of the QA system by making complex chemical data more tangible but also enhances the overall user experience, allowing researchers to engage with the data more interactively.

4. Results and discussion

By integrating the OntoMOPs knowledge domain and its semantically structured data with our QA system Marie, we have successfully created a functioning KG-RAG system for MOP-related research. Not limited to a simple database lookup, Marie can access deep domain knowledge of MOPs, including their underlying structures, components, and design principles. [Figure 4](#) illustrates how the modular architecture of Marie facilitates a powerful KG-RAG system that can reliably traverse a complex KG. Retrieving different kinds of data, including molecular geometries, enables informative multilayered output: as shown in [Figure 5](#), factual answers can be given in natural language combined with integrated 3D visualizations. The adapted architecture of Marie, utilizing in-context prompting coupled with entity recognition techniques, enables shorter development cycles for new RAG systems. Moreover, it allows for iterative extension beyond their common scope to more niche domains like MOPs. This brings us a step closer to creating a "Digital Research Scientist" (Rihm et al., 2024) by providing an assistant with which researchers can have a productive conversation to aid them in the scientific discovery process (Klami et al., 2024), as shown in [Figure 6](#).

[Figure 4](#) demonstrates the usability of our QA system and the functions of its components with a rundown of Marie's handling of an exemplary query in the domain of MOP chemistry, "Which CBUs are used as two-linear GBUs?" The QA process follows the general flowchart given in [Figure 3](#): as no quantities are detected, unit conversion and input rewriting are not needed in this case, so only the processes related to semantic parsing and response generation are triggered. The invocation of a particular GBU in the second part of the question triggers Marie's entity recognition and linking module, which identifies the exact IRI that corresponds to the mentioned entity. In this case, Marie can find the instance of GenericBuildingUnit with the required unique combination of hasModularity and hasPlanarity properties

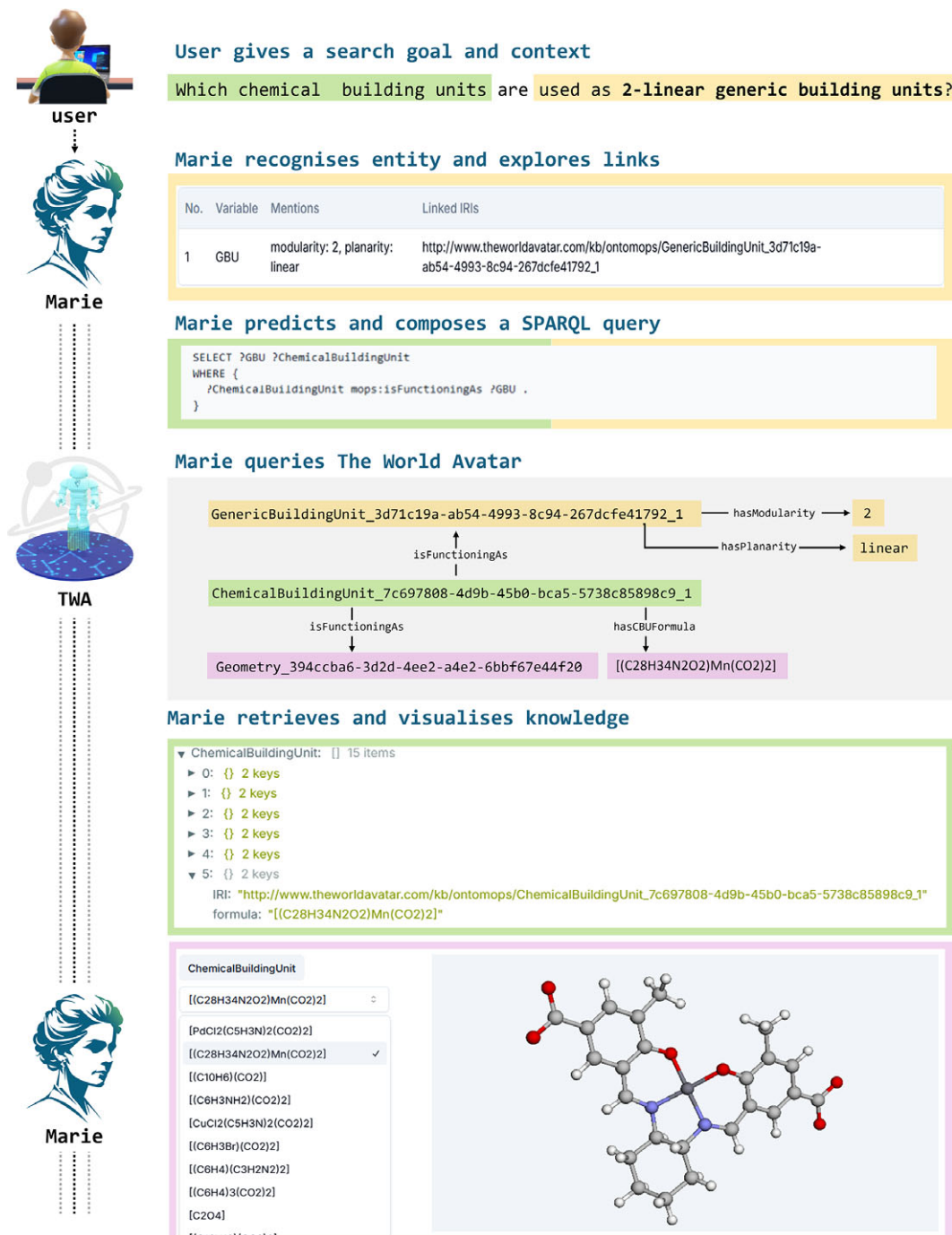




Figure 4. Processing steps to respond to a natural language question in the MOP chemistry domain as implemented in Marie. These steps are displayed on the Marie page and can be retraced for every question.

via RDF subgraph matching. The recognized entity serves as a starting point for traversing the KG with a SPARQL query. The prediction of such a query is invoked by the first part of the question, asking for entities of type ChemicalBuildingUnit that are linked to the previously recognized entity (and thereby its



user

"What MOPs and respective assembly models are described in the paper with DOI 10.1038/nchem.803?"



Marie TWA

The paper with DOI 10.1038/nchem.803 describes the following Metal-Organic Polyhedra (MOPs) and their respective assembly models:

- MOP Formula:** [Cu2]12[((C6H3)C((CH3)3)(CO2)2]24

 - Assembly Model Name: (4-planar)x12(2-bent)x24
 - Polyhedral Shape: Anticuboctahedron
 - Symmetry Point Group: D3h
- MOP Formula:** [Cu2]12[(C6H3SO3)(CO2)2]24

 - Assembly Model Name: (4-planar)x12(2-bent)x24
 - Polyhedral Shape: Anticuboctahedron
 - Symmetry Point Group: D3h
- MOP Formula:** [Cu2]12[(C6H3OH)(CO2)2]24

 - Assembly Model Name: (4-planar)x12(2-bent)x24
 - Polyhedral Shape: Anticuboctahedron
 - Symmetry Point Group: D3h

All MOPs share the same assembly model characteristics, indicating they belong to a specific structural family with the same assembly model configuration and symmetry.

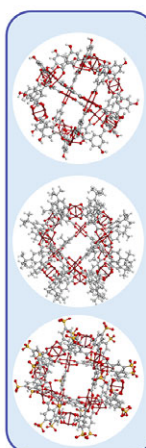


Figure 5. Example of a multilayered response by Marie, combining a natural language summary of data retrieved from the knowledge graph with 3D visualization of chemical structures.

IRI) via an `isFunctioningAs` predicate. As the query is valid, it is automatically extended before execution so that the results returned are not only machine-readable IRIs of appropriate CBUs but also scientifically meaningful identifiers that can be supplied to the user, such as chemical formulas. The retrieved CBUs and associated data (here in JSON format) can now be used to generate tabular overviews or natural language responses. In the case of reticular chemistry, lengthy formulas are often not enough for a human user to understand the presented structures intuitively. For this reason, the geometries of certain entity types are retrieved as well and structures are visualized in an interactive 3D viewer, giving users a tangible means of comprehending the results. Notably, the presentation of Marie's internal workings—including its entity linking, SPARQL query formulation, and retrieval of node IRIs—contributes to the system's interpretability and users' confidence in its accuracy. Sanity checks can also be performed at any step by looking up intermediate values directly in the RDF triplestore.

Figure 5 demonstrates how the structure visualization combined with natural language responses based on knowledge retrieval can be especially valuable for complex MOP structures. In the illustrated example, a user enquires about MOPs described in a specific scientific paper—a typical question a chemist would try to answer when reviewing publications reporting different types of MOPs. This can be quite an extensive task when done by hand, especially when trying to compare structural similarity in terms of assembly models and symmetry. Even when consulting a dedicated review or, in this case, a single work that includes a collection of MOPs and their properties, it is hard to successfully keep track of and distinguish these MOPs. Their formulas are often insufficient for human users to construct a mental image of the MOPs, and although they can be broadly described in terms of polyhedral shapes, the vast variability in geometric shapes means that even morphology experts might not be able to immediately conceive of the structures. By providing interactive visualization of these structures enabling 3D rotation,

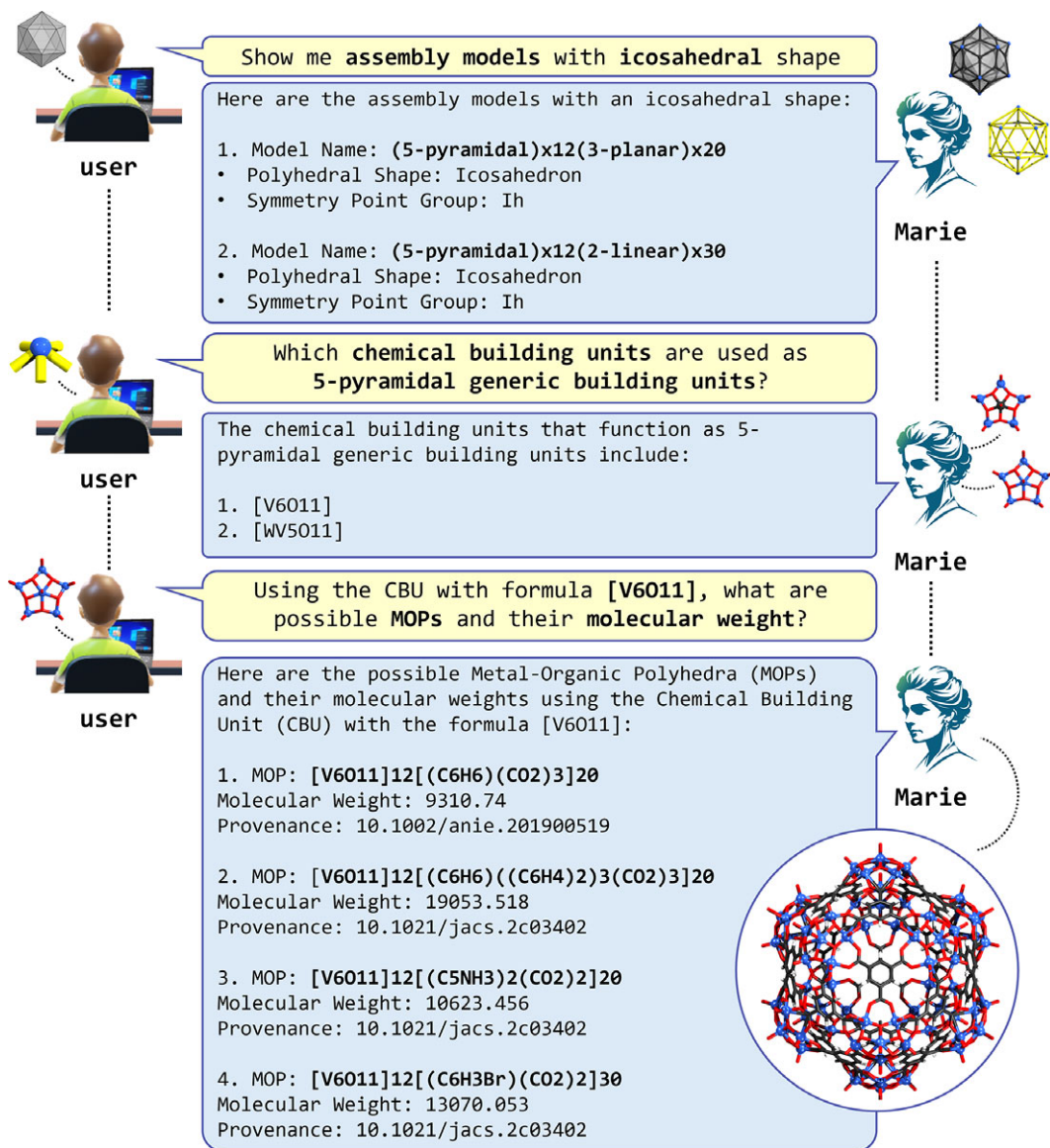


Figure 6. Example of a conversation with Marie via chained questions.

our platform not only aids the understanding of MOP topologies but also enhances the output provided by Marie by rendering it more intuitive and accessible. Finally, a summarizing sentence as shown in Figure 5 can provide instant comprehension, even when the number of items returned might be much larger for some queries.

Marie's detailed responses and interactive usage enable users to navigate the knowledge base of MOPs efficiently. This could prove useful for chemists who look to synthesize MOPs with certain properties and need to probe potential candidates. Figure 6 illustrates how this use case can be realized with Marie. Starting with a desired structural shape, the chemist may use Marie to retrieve all assembly models that exhibit this geometry. Subsequently, the frequent occurrence of the five-pyramidal GBU among retrieved assembly models may prompt the chemist to search for CBUs that can function as such. Marie identifies two potential CBUs, of which the chemist chooses one to focus on, querying for MOPs that contain them

and checking for their molecular weight, to which Marie responds with a comprehensive list of materials. The results are not limited to MOPs that have been previously reported in the literature but also include machine-predicted ones, enabling the chemist to explore potential synthesis targets thoroughly.

With a question chain as illustrated in Figure 6, users can traverse the KG step-by-step, using each response as additional information to base the next question on. With these three questions, the user was able to explore the KG across the four core concepts highlighted in Figure 2: starting at an assembly model (via entity recognition and query prediction), the user picks a GBU for which Marie provides appropriate CBUs. The user picks a CBU and asks Marie for MOPs, which Marie returns and augments with molecular data, provenance information, and structural geometry.

In verifying the QA system with both the examples presented and other sample questions, Marie's responses were consistently accurate and repeatable. However, Marie does have two modes of failure: (1) if a question is misunderstood, it can lead to an incorrect KG query being generated and (2) if the requested information is not available in the knowledge base, the query response will be empty. In both cases, no data are retrieved, and the user will be notified accordingly. Marie is limited to the specific knowledge base it operates on, making it more suitable for highly specialized topics, such as MOPs and their properties. In contrast, a general-purpose LLM can answer questions across various domains but lacks in-depth knowledge in niche areas and can potentially introduce hallucinations. Within the scope of its intended topic, previous work has demonstrated that Marie outperforms general-purpose QA systems like ChatGPT (Pascazio et al., 2024). A similar evaluation for the updated version of Marie, along with a more quantitative assessment of the QA system's accuracy and repeatability, can be found in the [Supplementary Materials](#).

5. Conclusion

This article presents a QA system tailored for MOP chemistry, backed by the MOP knowledge base of empirically verified and machine-predicted instances enriched with geometry data. Our work focused on overcoming three critical pain points: the difficulty of navigating complex and domain-specific concepts not fully accessible by general-purpose LLMs, the challenge of effectively understanding and visualizing complex information and structures in reticular chemistry, and the need to accelerate development cycles for QA systems by reducing model (re-)training requirements. To address these issues, we introduced several key innovations, including the integration of MOP data into an existing KG-integrating QA system called Marie, the incorporation of multilayered output incorporating visual, textual, and tabular hyperlinked outputs to enhance the interpretation of complex data, and the adaptation of few-shot learning techniques to optimize the system's performance in new domains. Through these advancements, we have demonstrated notable improvements in the capability and efficiency of the Marie QA system within the specialized context of MOPs, paving the way for more effective and accessible scientific inquiry in this field.

Our work demonstrates the use of natural language to efficiently navigate TWA's vast repository of MOPs, which can aid chemists in rapidly screening for synthesis targets with desired properties. These enhancements broaden the scope of exploration within the MOP space and provide a visual interface that makes the data more tangible. This marks a significant step forward in making MOP data more accessible and actionable for researchers, ultimately supporting ongoing efforts in MOP design and application. The architecture we have developed for the Marie QA system holds significant potential for broader applications in scientific research. Enabling the simple and resourceful creation of KG-RAG models based on a body of knowledge described in individual papers, collections of papers, or comprehensive scientific databases can help gather insights from large data sources and drastically increase the accessibility of scientific knowledge. This flexibility allows researchers to quickly adapt the system to emerging fields or specific niches, democratizing access to cutting-edge research and fostering innovation.

The scope of this work is limited to the retrieval of facts from a KG, but answering more complex questions based on the available information is an important research question on the path to creating increasingly autonomous "digital research scientists." Marie's knowledge base consists of different topical KGs to which the QA system has access. The accuracy of information inside them is currently

ensured by human and agentic curation. While it can be restored from local copies at any time, data protection needs to be addressed. Going forward, expanding the KG while ensuring its accuracy and security will be crucial, as data safety and security in TWA are areas of ongoing research. Future efforts will also focus on integrating Marie with automated synthesis planning tools to enable the swift design and optimization of new MOPs with targeted functionalities (Rihm et al., 2024; Kondinski et al., 2024a). Additionally, future work could explore the application of the presented architecture to other specialized domains, further refining the integration of multilayered outputs and combining the use of in-context prompting and query prediction with embedding methods to enhance the efficiency of KG-RAG-based QA systems.

Abbreviations

CBU	chemical building unit
GBU	generic building unit
IRI	Internationalized Resource Identifier
JSON	JavaScript Object Notation
KG	knowledge graph
LLM	large language model
MOF	metal–organic framework
MOPs	metal–organic polyhedra
NLP	natural language processing
QA	question answering
RAG	retrieval-augmented generation
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
TBox	terminological component
TWA	The World Avatar (project)

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/dce.2025.12>.

Data availability statement. Marie is available at <https://theworldavatar.io/demos/marie> for testing. The source code for Marie is available at https://github.com/cambridge-cares/TheWorldAvatar/tree/main/QuestionAnswering/QA_ICL. Additional data for deployment of the presented Marie version (including model weights, in-context questions, etc.) are available upon request.

Acknowledgments. Markus Kraft gratefully acknowledges the support of the Alexander von Humboldt Foundation.

Author contribution. Conceptualization: D.N.T., S.D.R., M.K., L.P.; Data curation: A.K., F.S., S.D.R.; Funding acquisition: M.K., J.A.; Investigation: D.N.T., F.S.; Methodology: D.N.T.; Project administration: S.D.R.; Resources: S.M., M.K.; Software: D.N.T., X.D., L.P.; Supervision: M.K., L.P., S.M., J.A.; Validation: D.N.T., S.D.R., X.D.; Visualization: A.K., S.D.R.; Writing—original draft: D.N.T., S.D.R., F.S.; Writing—review and editing: S.D.R., A.K., M.K. All authors approved the final submitted draft.

Funding statement. This research was supported by the National Research Foundation, Prime Minister’s Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program. Simon D. Rihm acknowledges financial support from Fitzwilliam College, Cambridge, and the Cambridge Trust.

Competing interest. The authors declare none.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

- Adeola AO, Ighalo JO, Kyesmen PI and Nomngongo PN (2024) Metal-organic polyhedra (MOPs) as an emerging class of metal-organic frameworks for CO₂ photocatalytic conversions: Current trends and future outlook. *Journal of CO₂ Utilization* 80, 102664.
- Akroyd J, Bhave A, Brownbridge G, Christou E, Hillman MD, Hofmeister M, Kraft M, Lai J, Lee KF, Mosbach S, Nurkowski D and Parry O (2022) CReDo technical report 1: Building a cross-sector digital twin. Technical report, Centre for Digital Built Britain (CDBB).

- Akroyd J, Mosbach S, Bhawe A and Kraft M** (2021) Universal digital twin – a dynamic knowledge graph. *Data-Centric Engineering* 2, e14.
- Bai X, Xie Y, Zhang X, Han H and Li J-R** (2024) Evaluation of open-source large language models for metal–organic frameworks research. *Journal of Chemical Information and Modeling* 64(13), 4958–4965.
- Berners-Lee T, Hendler J and Lassila O** (2001) The semantic web. *Scientific American* 284(5), 34–43.
- Bran AM, Cox S, Schilter O, Baldassari C, White AD and Schwaller P** (2024) Augmenting large language models with chemistry tools. *Nature Machine Intelligence* 6(5), 525–535.
- Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D** (2020) Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20 Curran Associates, Inc.
- Chong S, Lee S, Kim B and Kim J** (2020) Applications of machine learning in metal–organic frameworks. *Coordination Chemistry Reviews* 423, 213487.
- Farazi F, Krdzavac NB, Akroyd J, Mosbach S, Menon A, Nurkowski D and Kraft M** (2020) Linking reaction mechanisms and quantum chemistry: An ontological approach. *Computers & Chemical Engineering* 137, 106813.
- Gallegos M, Vassilev-Galindo V, Poltavsky I, Martín Pendás Á and Tkatchenko A** (2024). Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. *Nature Communications* 15(1), 4345.
- Gasteiger J** (2016) Chemoinformatics: Achievements and challenges, a personal view. *Molecules* 21(2), 151.
- Ghosh AC, Legrand A, Rajapaksha R, Craig GA, Sassoce C, Balázs G, Farrusseng D, Furukawa S, Canivet J and Visser FM** (2022) Rhodium-based metal–organic polyhedra assemblies for selective CO₂ photoreduction. *Journal of the American Chemical Society* 144(8), 3626–3636.
- Gosselin AJ, Rowland CA and Bloch ED** (2020) Permanently microporous metal–organic polyhedra. *Chemical Reviews* 120(16), 8987–9014.
- Guan J, Huang T, Liu W, Feng F, Japip S, Li J, Wu J, Wang X and Zhang S** (2022) Design and prediction of metal organic framework-based mixed matrix membranes for CO₂ capture via machine learning. *Cell Reports Physical Science* 3(5), 100864.
- Guo T, Guo K, Nan B, Liang Z, Guo Z, Chawla NV, Wiest O and Zhang X** (2024) What can large language models do in chemistry? A comprehensive benchmark on eight tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23 Curran Associates, Inc.
- Guo T, Nan B, Liang Z, Guo Z, Chawla N, Wiest O, Zhang X et al** (2023) What can large language models do in chemistry? A comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems* 36, 59662–59688.
- Hadsell R, Rao D, Rusu AA and Pascanu R** (2020) Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences* 24(12), 1028–1040.
- IBM** (2021) Roborxn. Available at <https://research.ibm.com/science/ibm-roborxn/>. (accessed 2 September 2024).
- Jablunka KM, Schwaller P, Ortega-Guerrero A and Smit B** (2024) Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* 6(2), 161–169.
- Kang Y and Kim J** (2024) ChatMOF: An artificial intelligence system for predicting and generating metal–organic frameworks using large language models. *Nature Communications* 15(1), 4705.
- Kim J, Kwon Y, Jo Y and Choi E** (2023) KG-GPT: A general framework for reasoning on knowledge graphs using large language models. In Bouamor H, Pino J and Bali K (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 9410–9421.
- Klami A, Damoulas T, Engkvist O, Rinke P and Kaski S** (2024) Virtual laboratories: Transforming research with AI. *Data-Centric Engineering* 5, e19.
- Kondinski A, Bai J, Mosbach S, Akroyd J and Kraft M** (2023) Knowledge engineering in chemistry: From expert systems to agents of creation. *Accounts of Chemical Research* 56(2), 128–139.
- Kondinski A, Menon A, Nurkowski D, Farazi F, Mosbach S, Akroyd J and Kraft M** (2022) Automated rational design of metal–organic polyhedra. *Journal of the American Chemical Society* 144(26), 11713–11728.
- Kondinski A, Mosbach S, Akroyd J, Breeson A, Tan YR, Rihm S, Bai J and Kraft M** (2024a) Hacking decarbonization with a community-operated creatorspace. *Chem* 10(4), 1071–1083.
- Kondinski A, Rutkevych P, Pascasio L, Tran DN, Farazi F, Ganguly S and Kraft M** (2024b) Knowledge graph representation of zeolitic crystalline materials. *Digital Discovery* 3, 2070–2084.
- Lee S, Jeong H, Nam D, Lah MS and Choe W** (2021) The rise of metal–organic polyhedra. *Chemical Society Reviews* 50(1), 528–555.
- Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktächel T, et al** (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33, 9459–9474.
- Li L, Zhou T, Li J and Wang X** (2022) A machine learning-based decision support framework for energy storage selection. *Chemical Engineering Research and Design* 181, 412–422.
- Luo Y, Bag S, Zaremba O, Cierpka A, Andreo J, Wuttke S, Friederich P and Tsotsalas M** (2022) MOF synthesis prediction enabled by automatic data mining and machine learning. *Angewandte Chemie International Edition* 61(19), e202200242.

- Min S, Lewis M, Zettlemoyer Land Hajishirzi H** (2022) MetalCL: Learning to learn in context. In Carpuat M, de Marneffe M-C and Meza Ruiz IV (eds), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, USA, 2791–2809.
- Moghadam PZ, Li A, Wiggin SB, Tao A, Maloney AGP, Wood PA, Ward SC and Fairen-Jimenez D** (2017) Development of a cambridge structural database subset: A collection of metal–organic frameworks for past, present, and future. *Chemistry of Materials* 29(7), 2618–2625.
- Mosbach S, Menon A, Farazi F, Krdzavac N, Zhou X, Akroyd J and Kraft M** (2020) Multiscale cross-domain thermochemical knowledge-graph. *Journal of Chemical Information and Modeling* 60(12), 6155–6166.
- Murray-Rust P** (2008) Chemistry for everyone. *Nature* 451(7179), 648–651.
- Nie Z, Zhang R, Wang Z and Liu X** (2024) Code-style in-context learning for knowledge-based question answering. *Proceedings of the AAAI Conference on Artificial Intelligence* 38(17), 18833–18841.
- OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman A-L, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung HW, Cummings D, Currier J, Dai Y, Decareaux C, Degry T, Deutsch N, Deville D, Dhara A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu SS, Guo Y, Hallacy C, Han J, Harris J, He Y, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S, Hu X, Huizinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser L, Kamali A, Kanitscheider I, Keskar NS, Khan T, Kilpatrick L, Kim JW, Kim C, Kim Y, Kirchner JH, Kiros J, Knight M, Kokotajlo D, Kondraciuk L, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li CM, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney SM, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O’Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, Peres FdAB, Petrov M, Pinto HPdO, Michael P, Pokrass M, Pong VH, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such FP, Summers N, Sutskever I, Tang J, Tezak N, Thompson MB, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe JFC, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang JJ, Wang A, Wang B, Ward J, Wei J, Weinmann C, Welihinda A, Welinder P, Weng J, Weng L., Wiethoff M., Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S, Zheng T, Zhuang J, Zhuk W, and Zoph B** (2023) GPT-4 Technical report.
- Pascasio L, Rihm S, Naseri A, Mosbach S, Akroyd J and Kraft M** (2023) Chemical species ontology for data integration and knowledge discovery. *Journal of Chemical Information and Modeling* 63(21), 6569–6586.
- Pascasio L, Tran D, Rihm SD, Bai J, Mosbach S, Akroyd J and Kraft M** (2024) Question-answering system for combustion kinetics. *Proceedings of the Combustion Institute* 40(1), 105428.
- Pérez J, Arenas M and Gutierrez C** (2009) Semantics and complexity of sparql. *ACM Transactions on Database Systems* 34(3), 1–45.
- Perry Iv JJ, Perman JA and Zaworotko MJ** (2009) Design and synthesis of metal–organic frameworks using metal–organic polyhedra as supermolecular building blocks. *Chemical Society Reviews* 38(5), 1400–1417.
- Quilitz B and Leser U** (2008) Querying distributed rdf data sources with sparql. In *The Semantic Web: Research and Applications*. Springer, 524–538
- Rego N and Koes D** (2014). 3Dmol.js: Molecular visualization with WebGL. *Bioinformatics* 31(8), 1322–1324.
- Reimers N and Gurevych I** (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui K, Jiang J, Ng V and Wan X (eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 3982–3992.
- Rihm SD, Bai J, Kondinski A, Mosbach S, Akroyd J and Kraft M** (2024) Transforming research laboratories with connected digital twins. *Nexus* 1(1), 100004.
- Rosen AS, Iyer SM, Ray D, YaoZ, Aspuru-Guzik A, Gagliardi L, Notestein JM and Snurr RQ** (2021). Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter* 4(5), 1578–1597.
- Sanmartin D** (2024). KG-RAG: Bridging the gap between knowledge and creativity. arXiv preprint, arXiv:2405.12035v1.
- Schick T, Dwivedi-Yu J, Dessi R, Raileanu R, Lomeli M, Hambro E, Zettlemoyer L, Cancedda N and Scialom T** (2023) Toolformer: Language models can teach themselves to use tools. In Oh A, Naumann T, Globerson A, Saenko K, Hardt M, and Levine S (eds), *Advances in Neural Information Processing Systems*, Curran Associates, Inc. vol. 36, 68539–68551.

- Taylor KR, Gledhill RJ, Essex JW, Frey JG, Harris SW and De Roure D C** (2006) Bringing chemical data onto the semantic web. *Journal of Chemical Information and Modeling* 46(3),939–952.
- Tran D, Pascazio L, Akroyd J, Mosbach S and Kraft M** (2024) Leveraging text-to-text pretrained language models for question answering in chemistry. *ACS Omega* 9(12),13883–13896.
- Vardhan H, Yusubov M and Verpoort F** (2016) Self-assembled metal–organic polyhedra: An overview of various applications. *Coordination Chemistry Reviews* 306, 171–194.
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, ‘t Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J and Mons B** (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3(1), 160018.
- Zhang Y, Dai H, Kozareva Z, Smola A and Song L** (2018) Variational reasoning for question answering with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence* 32(1).
- Zhang D, Liu W, Tan Q, Chen J, Yan H, Yan Y, Li J, Huang W, Yue X, Ouyang W, Zhou D, Zhang S, Su M, Zhong H-S and Li Y** (2024) ChemLLM: A chemical large language model. arXiv preprint, arXiv:2402.06852.
- Zheng Z, Zhang O, Nguyen HL, Rampal N, Alawadhi AH, Rong Z, Head-Gordon T, Borgs C, Chayes JT and Yaghi OM** (2023) ChatGPT research group for optimizing the crystallinity of MOFs and COFs. *ACS Central Science* 9(11), 2161–2170.
- Zhou X, Eibeck A, Lim MQ, Krdzavac NB and Kraft M** (2019) An agent composition framework for the J-Park Simulator – a knowledge graph for the process industry. *Computers & Chemical Engineering* 130, 106577.