Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

Natural Language Access Point to Digital Metal-Organic Polyhedra Chemistry in The World Avatar

Dan Tran[1], Simon D. Rihm[2], Aleksandar Kondinski[2], Laura Pascazio[1], Fabio Saluz[2,3], Sebastian Mosbach[1,2,4], Jethro Akroyd[1,2,4], Markus Kraft[1,2,4]

released: September 5, 2024

[1] CARES
Cambridge Centre for Advanced Research and Education in Singapore
1 Create Way
CREATE Tower, #05-05
Singapore, 138602

[2] Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

[3] D-MAVT
ETH Zurich
Rämistrasse 101
Zurich, CH-8092
Switzerland

[4] CMCL
No. 9, Journey Campus
Castle Park
Cambridge
CB3 0AX
United Kingdom

Preprint No. 327

UNIVERSITY OF CAMBRIDGE

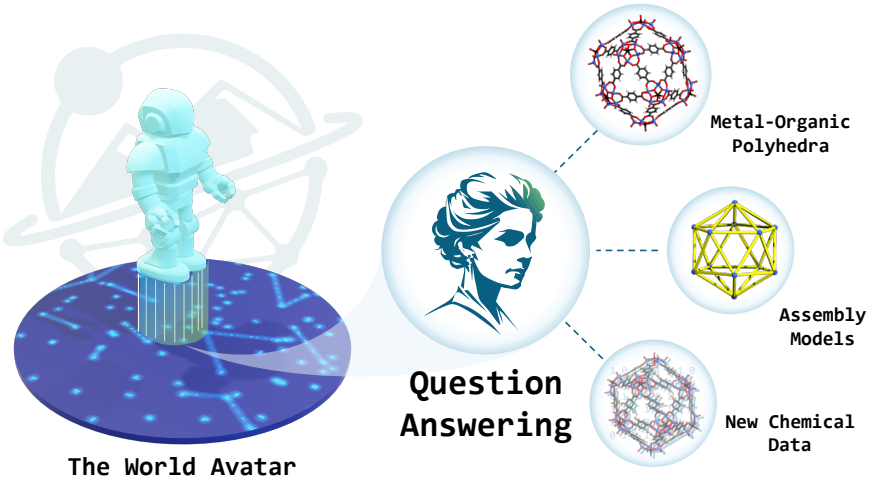
Keywords: metal-organic polyhedra, large languag models, question-answering systems, dynamic knowledge graphs, retrieval-augmented generation

Edited by
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

E-Mail: mk306@cam.ac.uk
World Wide Web: https://como.ceb.cam.ac.uk/

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 systematised known MOPs and expanded the explorable MOP space with novel targets. Although this data is 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 paper, 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 visualisation 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.



Highlights

- Metal-organic polyhedron (MOP) data made accessible via natural language.
- Visualisation of MOP geometries, including empirically verified and machine-predicted structures.
- Accelerated adaptation of the QA system to new data and domains using few-shot learning methods.

Contents

1 Introduction

Metal-organic polyhedra (MOPs) represent a class of materials characterised by their self-assembled, cage-like discrete nanomolecular architecture constructed from metal-based and organic building blocks [17, 33]. 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 [49, 60]. However, considering the chemical space that emerges by a 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 new material discovery [11, 26, 55], selection [20, 36] and synthesis [38], as well as the development of data infrastructures to support these tasks, including data cataloguing [41], mining [5], and accessing [26].

Considering the relatively lower sample size of MOPs in comparison to its extended metal-organic framework analogues, the development of data-driven digital tools for MOP discovery has remained challenging. This is simple because big data-driven that fits MOFs cannot be easily extended to to MOPs. In this regard, our group has developed new formal and semantic approaches to describe MOPs, including custom-designed inductive reasoning algorithms for new structure discovery. Thus, following a careful development of a knowledge model for MOP chemistry (*i.e.* a "OntoMOP" (KG)), we have instantiated 151 MOPs experimentally described MOPs, and based on them, our reasoning algorithm designed new 1,418 MOP instances that are rationally designed based on existing building units, following expert-like patterns of molecular engineering [29]. The overall research has been originally contextualised 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 very long time came with the requirement of some forms of cheminformatics knowledge [15]. In a similar line, acquiring chemical information instantiated in the form of a knowledge graph typically requires the use of querying tools such as SPARQL [48, 50], which may appear unintuitive and even cumbersome to operate by new users, thus, unfortunately, limiting the accessibility to chemical information. Despite our original success in describing MOP chemistry via a knowledge graph model and in developing agents for digital exploration of its chemical space, semantic query tools often appeared as a barrier for experimental chemists who may want to use the insights from our work towards the development of new materials. By noticing similar experiences along different chemistry domains, we were motivated to develop tools that integrate semantic querying with natural language processing, 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 public domain and therefore not accessible *via* traditional search engines or general-purpose large language models (LLMs) . One of TWA's core goals is to develop user-friendly interfaces that enable researchers and industry practitioners to efficiently interact with TWA's extensive data and leverage its powerful modelling and problem-solving capabilities. 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 [32, 47]. However, Marie's reliance on supervised fine-tuning in the development of 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 [23]. Lastly, specific to reticular chemistry is the problem of understanding complex information and structures, which calls for visualization.

The purpose of this paper 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 visualisation of not just empirically verified MOP structures but also those predicted by our "MOP Discovery" agent. 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.

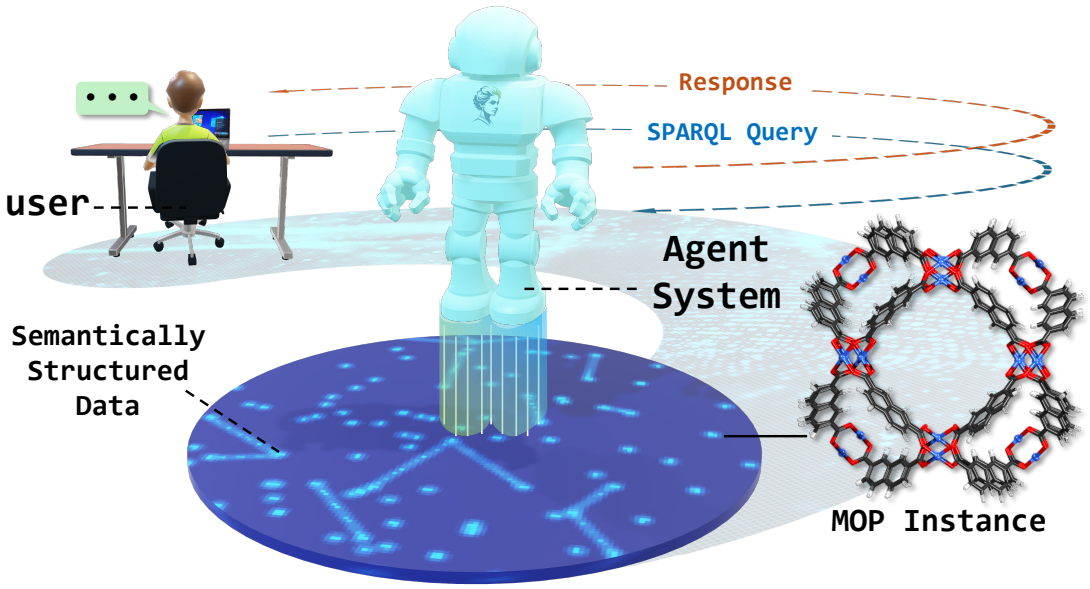


Figure 1: *Illustration of TWA's digital infrastructure that enables the retrieval of structured and validated MOP data via natural language requests.*

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 The World Avatar – 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 potentials of the Semantic Web to enhance cheminformatics and broader chemical applications [6, 43, 58]. Initially conceptualised in 2010, this initiative 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 [2], including power and heat network optimisations for CO_2 savings, environmental monitoring, and cross-domain climate resilience planning through the Climate Resilience Demonstrator [2, 3, 42]. TWA operates on the Semantic Web principles and adheres to the FAIR guidelines to ensure all data is findable, accessible, interoperable, and reusable [62]. It integrates software agents that manage information flows, interface with computational models, and continuously enhance TWA's KGs with new data [2, 68].

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 [30, 31]. This framework supports detailed data relationships and enhances interoperability, enabling multifaceted data usage and reducing ambiguities [2, 12]. Additionally, computational agents in TWA perform complex tasks such as calibrating kinetic mechanisms and automating discovery processes [31], exemplified by the development of novel MOPs [29] which, amongst a variety of applications, can be used for photocatalytic CO_2 reduction [1, 16].

The OntoMOPs ontology is designed to provide and enrich semantic relationships between MOPs, chemical building units (CBUs), and assembly models (AMs) [29]. This ontology enables advanced query capabilities for professionals engaged in the modelling and preparation of MOPs, supporting informed decision-making with detailed information on the construction and functionalities of these materials. OntoMOPs links MOP instances to crucial metadata such as molecular mass, charge, formulae, 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 recognised 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 natural language processing (NLP) has experienced a remarkable rise in popularity, primarily driven by the accessible deployment of large language models. The advent of LLMs is marked by their remarkable ability to tackle diverse knowledge-intensive tasks that range from the humanities to the sciences, including chemistry [45]. However, despite their impressive performance on standardised examinations, general-purposed LLMs like GPT-4 often struggle with more advanced and specialised requests, revealing their lack of in-depth understanding of the subject matter [22]. While fine-tuning is a possible remedy [65], 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 the realm of chemistry, LLMs are increasingly utilised for a variety of tasks, including data processing, engineering, inference, and augmentation, in conjunction with various computational tools [21, 25, 39]. Despite these advancements, concerns about the explainability of these technologies continue to persist [14], 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 [27, 66].

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 [35]. This allows RAG systems to generate more contextually relevant and accurate responses [26, 67]. Using knowledge graphs as the foundation for information retrieval (KG-RAG) has shown great promise in recent studies to reliably handle knowledge-intensive and cognitive tasks [56].

Another challenge for knowledge-intensive QA systems is the 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 generalisation ability and versatility are key to addressing these dual goals. For example, ChemCrow [39] is a tool-calling agent capable of incorporating information pulled from a mixture of public and priority data sources and computational tools, including the PubChem database and the RoboRXN platform by IBM Research [24]. It does so by employing an LLM pre-trained 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 [57].

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 [32, 47]. However, the previous version of Marie relies on supervised fine-tuning for its semantic parser, which necessitates re-training whenever it needs to integrate with a new knowledge domain in TWA. In contrast, the in-context learning capability of LLMs [7] 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 [44].

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

In order to include MOP knowledge in our existing KG-based chemistry QA system Marie and extend it for better user interactivity, the MOP knowledge base needs to be restructured and extended first. Firstly, 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. Furthermore, the MOP KG was enriched with 370 new geometries of machine-predicted MOPs in addition to the 151 existing geometries of previously synthesised MOPs. These molecular geometries were deduced from information represented in the KG and will help researchers to visualise these structures better and screen possible synthesis candidates.

The updated ontology is shown in Fig. 2. Its core concepts form a rectangle: MOPs can be classified by their geometric assembly models made up of distinct generic building units as which a variety of chemical building units can function [29]. These four core concepts now provide access to a range of geometric and molecular properties alike.

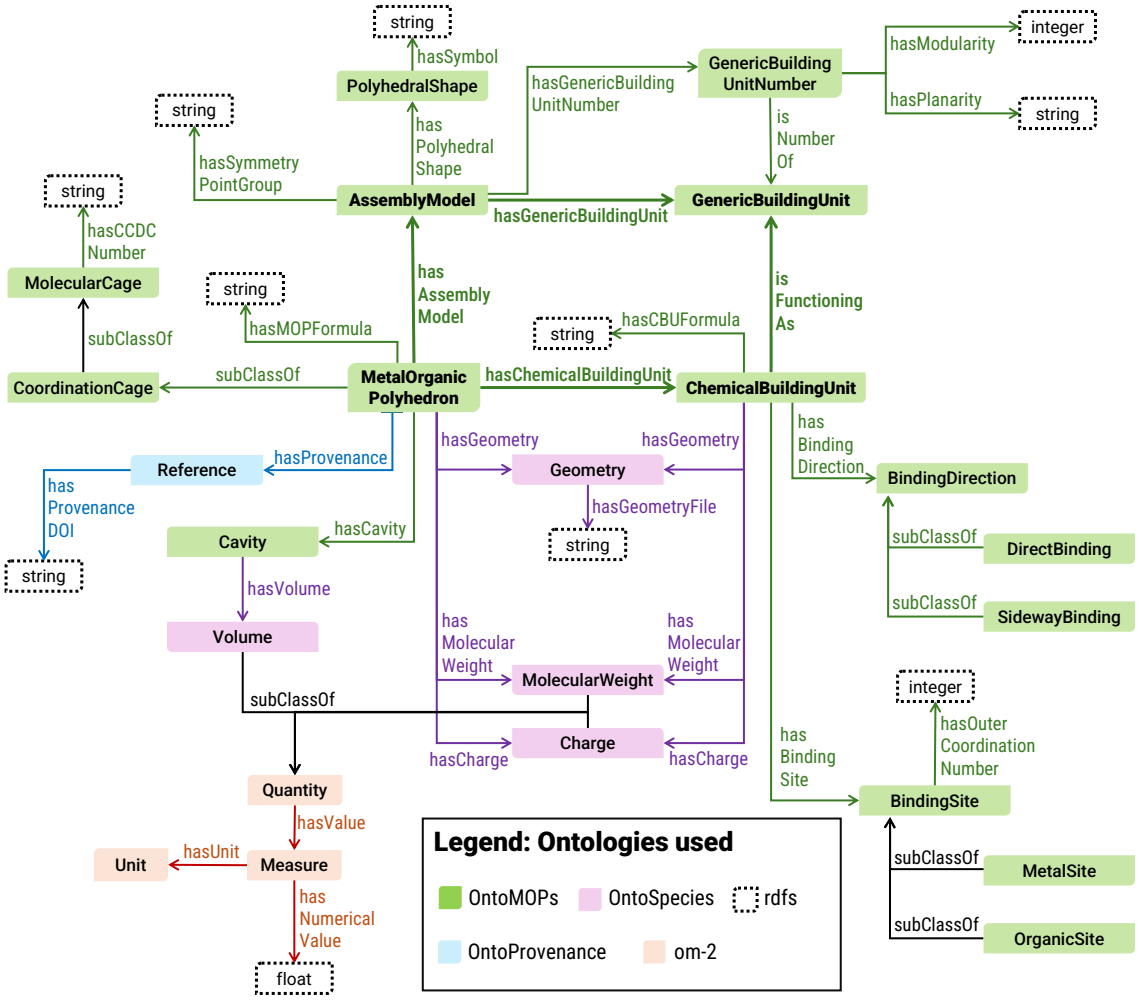


Figure 2: *Illustration of the terminological component (TBox) of the MOP chemistry domain in TWA and its related ontologies, core concepts are shown in bold.*

In the original implementation, geometry data of MOPs and CBUs is provided as an XYZ document or XYZ-formatted strings in the KG. The potential future limitation of this method is that a string length can, in principle, exceed the stringent length limits imposed by the KG engine for a very large chemical superstructure. An alternative implementation to this is to instantiate every atom of a MOP or CBU structure and link these atoms to the intermediate `Geometry` node, which is then connected to a MOP or CBU instance *via* the `hasGeometry` predicate, as done in the OntoSpecies domain of TWA [46]. 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 served on a web server so that they are accessible on the Internet *via* URLs, which are discoverable through `hasGeometryFile` links to `Geometry` nodes.

The original assertion component created redundancy in the assignment of Internationalised Resource Identifiers (IRIs) for instances of assembly models and GBUs, necessitating postprocessing of aggregate queries. In the new implementation, this instantiation aspect has been omitted, allowing for simpler traversal of the knowledge graph (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 [46], has facilitated the reuse of general-purpose concepts. As illustrated in Fig. 2, this reuse covers many shared molecular properties and literature provenance, thereby simplifying the training of the Marie NLP agent.

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 compartmentalisation of its data into distinct triplestores to allow domain experts to own and manage it independently. Earlier versions of Marie struggled with the dynamic nature of TWA, as its semantic parser relied on supervised fine-tuning [59], requiring resource-intensive re-training. This limitation not only impeded Marie's scalability but also posed the risk of catastrophic forgetting [23]. In contrast, this 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 few-shot learning methods to it. As depicted in Fig. 3, Marie's online workflow comprises three main components: input rewriter, semantic parser, and response generator. The input rewriter aligns all physical quantities mentioned in the input question to the unit systems in our knowledge base. The semantic parser jointly generates the logical form of a SPARQL query, detects surface forms of entities present in the input question, and determines the triplestore to execute the query. Lastly, the response generator presents the structured SPARQL response and LLM-generated styled text, accompanied by visualisation of the 3D structures of any invoked chemical entities.

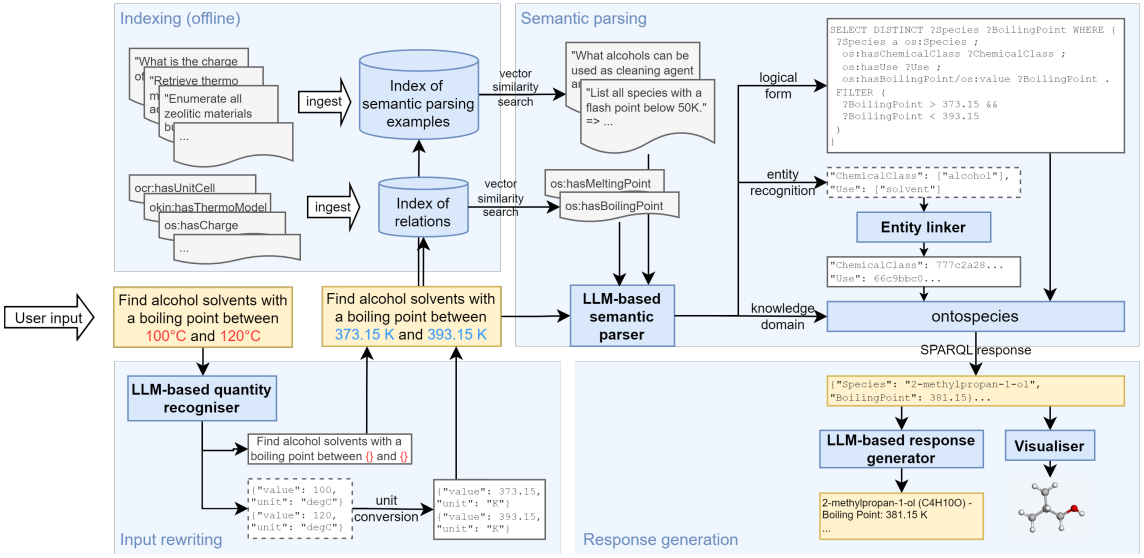


Figure 3: *Architecture of 'Marie', comprising one offline indexing stage and three online stages, namely input rewriting, semantic parsing, and response generation.*

Both the quantity recogniser and semantic parser are powered by LLMs prompted with in-context examples; the exact structure of the prompts is available in the Appendix A.3. While the LLM prompt for the physical quantity recogniser 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 processing time. Relevance is measured by the cosine similarity of their Sentence-BERT embeddings [52] using the `all-mpnet-base-v2` variant. We use OpenAI's `gpt-4o-mini-2024-07-18` model for in-context learning, and Redis Community Edition for all retrieval needs.

Marie's entity linking component mobilises multiple strategies, depending on the entity class. These include inverted index lookup for entities with well-defined labels, *e.g.* chemical species with their IUPAC names, molecular formulae, and SMILES strings; semantic search for entities that represent concepts or categories, *e.g.* chemical classifications; and RDF subgraph matching for more complex entities that are conceptually defined by their relationships with other entities, *e.g.* assembly models composed of GBUs [29]. Compared to earlier versions, this multi-strategy 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 Appendix A.3.3, 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 visualisation of intricate chemical structures like MOPs; this is done using the library 3Dmol.js [51]. 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 successfully created a functioning KG-RAG system for MOPs-related research. The information retrieval is thereby not limited to a simple database lookup; Marie has access to deep domain knowledge of MOPs, including their underlying structures, components, and design principles. Fig. 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 Fig. 5, factual answers can be given in natural language combined with integrated 3D visualisations. The adapted architecture of Marie, utilising 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' [53] by providing an assistant with which researchers can have a productive conversation to aid them in the scientific discovery process [28], as shown in Fig. 6.

Fig. 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 chemical building units are used as 2-linear generic building units?". The question-answering process follows the general flow chart given in Fig. 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 is able to find the instance of `GenericBuildingUnit` with the required unique combination of `hasModularity` and `hasPlanarity` properties *via* RDF subgraph matching. The recognised entity serves as a starting point for traversing the knowledge graph *via* SPARQL query. The prediction of such a query is invoked through the first part of the question, asking for entities of type `ChemicalBuildingUnit` that are linked to the previously recognised entity (and thereby its IRI) *via* a `isFunctioningAs` predicate. As the query was 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 formulae. 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 formulae are often not enough for a human user to understand the presented

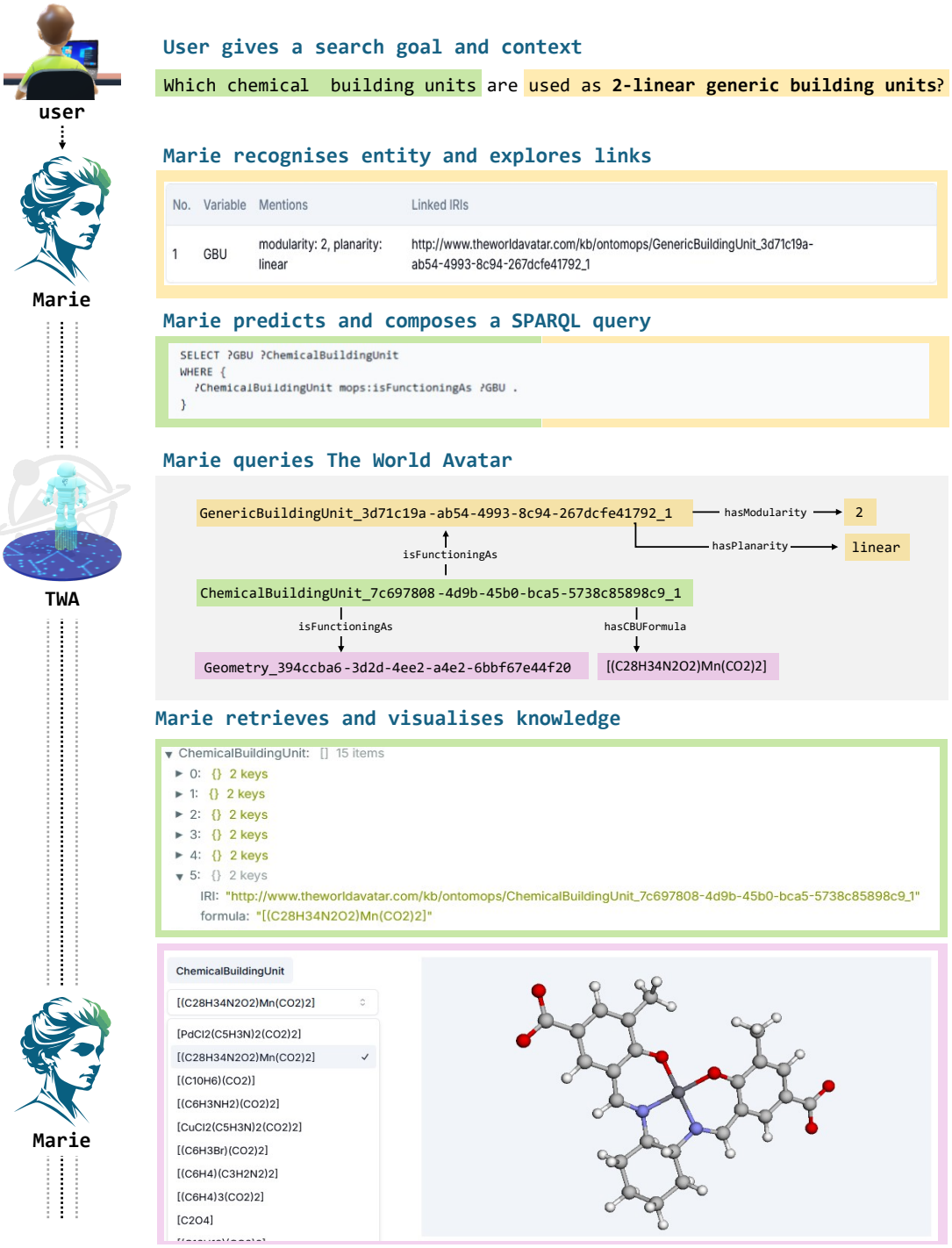


Figure 4: *Processing steps to respond to a natural language question in the MOP chemistry domain as implemented in Marie. These steps are displayed on the Marie page and can be retraced for every question.*

structures intuitively. For this reason, the geometries of certain entity types are retrieved as well, and structures are visualised 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 the system's correctness. Sanity checks can also be performed at any step by looking up intermediate values directly in the RDF store.

Fig. 5 demonstrates how the structure visualisation 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 formulae 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 visualisation of these structures enabling 3D rotation, 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. Lastly, a summarising sentence as shown in Fig. 5 can provide instant comprehension, even when the number of items returned might be much larger for some queries.

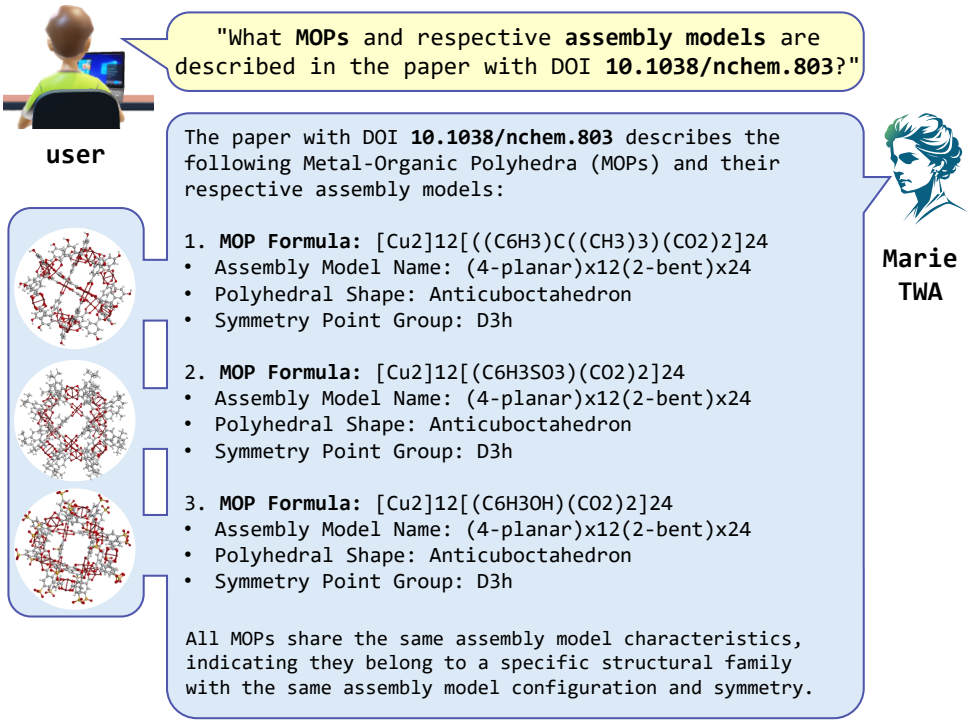


Figure 5: *Example of a multilayered response by Marie, combining a natural language summary of data retrieved from the knowledge graph with 3D visualisation of chemical structures.*

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 synthesise MOPs with certain properties and need to probe potential candidates. Fig. 6 illustrates how this use case can be realised 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 5-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 it 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.

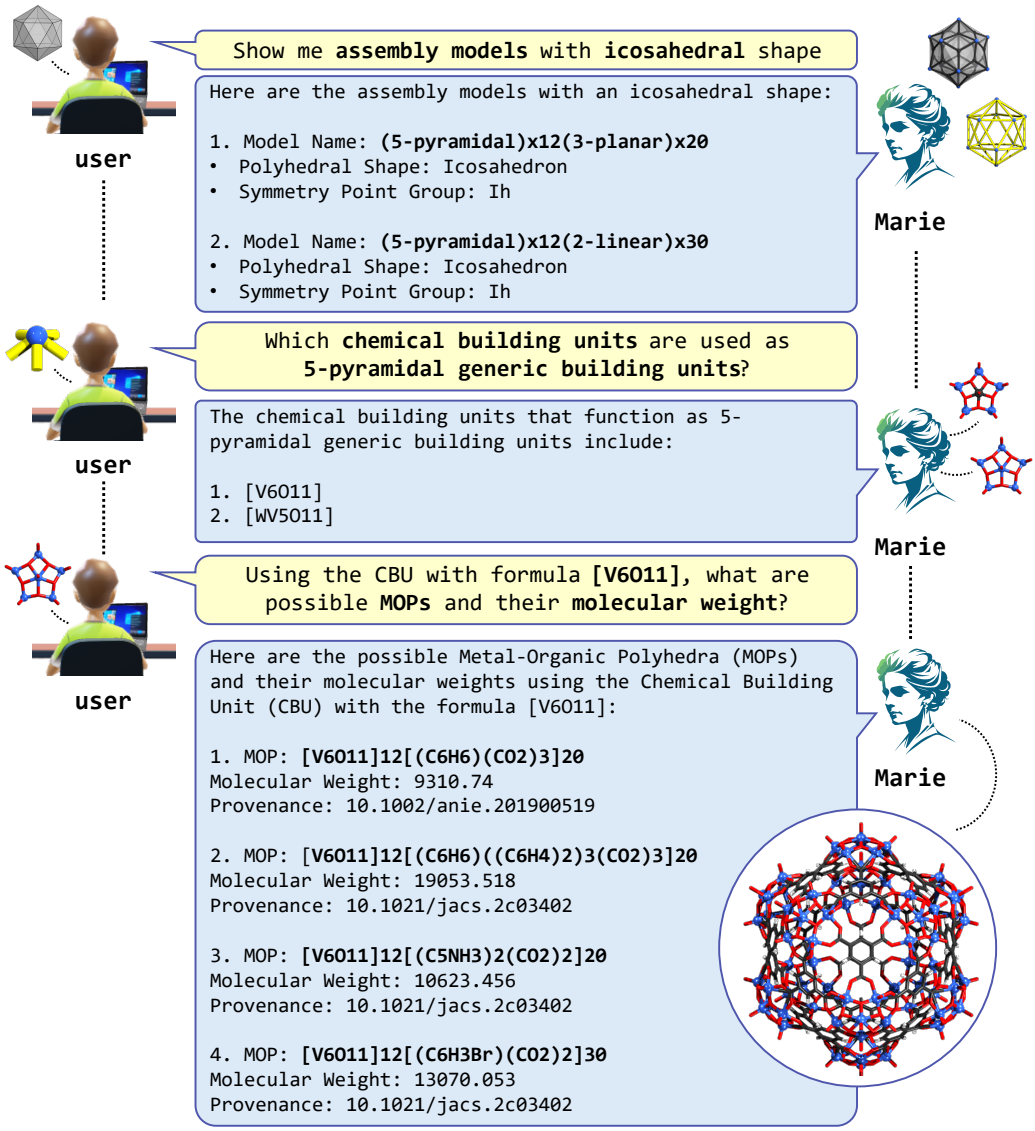


Figure 6: *Example of a conversation with Marie via chained questions.*

With such a question chain as illustrated in Fig. 6, users can traverse the knowledge graph 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 Fig. 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 metal-organic polyhedra, which Marie returns and augments with molecular data, provenance information, and structural geometry.

5 Conclusion

This paper 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 visualising 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 – 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 optimise 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 specialised 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. Future efforts will focus on integrating Marie with automated synthesis planning tools to enable the swift design and optimisation of new MOPs with targeted functionalities [31, 53].

Looking forward, 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, democratising access to cutting-edge research and fostering innovation. Future work could explore the application of this architecture to other specialised 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.

Nomenclature

BERT Bidirectional Encoder Representations from Transformers

CBU Chemical Building Unit

CCDC Cambridge Crystallographic Data Centre

DOI Digital Object Identifier

GBU Generic Building Unit

IRI Internationalised Resource Identifier

IUPAC International Union of Pure and Applied Chemistry

JSON JavaScript Object Notation

KG Knowledge Graph

LLM Large Language Model

MOP Metal-Organic Polyhedron

NLP Natural Language Processing

QA Question-Answering

RAG Retrieval-Augmented Generation

SMILES Simplified Molecular Input Line Entry System

SPARQL SPARQL Protocol and RDF Query Language

TBox Terminological and Assertion components

TWA The World Avatar

Acknowledgements

This research was supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Data and code availability

Marie is available at https://theworldavatar.io/demos/marie/
The source code for Marie is available at https://github.com/cambridge-cares/TheWorldAvatar/tree/main/QuestionAnswering/QA_ICL.

A Appendix

A.1 Retrieval-augmented generation (RAG)

In RAG systems, LLM generation needs not rely exclusively on facts implicitly encoded in model weights, but instead can make use of external data sources [34]. Specifically, the RAG architecture generally comprises two separate components: a *retriever* and a *reader*, also known as *generator*. The retriever gathers relevant facts from external data sources and the reader, which can be any text-to-text model, processes the retrieved data to respond to input questions. As data sources can be updated at a relative low cost and independently from LLM training, RAG systems are capable of incorporating new data in their responses without needing to update LLM weights.

The performance of RAG systems is generally sensitive to retrieval success and the ability of the LLM-reader to process long context windows. Ideally, the retrieved data should contain all the necessary information to answer a given question (high recall) and minimal amount of irrelevant data (high precision). The simplest data store is a collection of unstructured text documents. Despite advances in text search methods such as BM25 [54] and semantic search [9], this type of data repository suffers from relatively low retrieval recall and precision, which force the reader component to fall back on its internal knowledge. Structured data stores such as knowledge graphs and RDBMS databases can enable more targeted retrieval, but they require annotation efforts for data instantiation and more elaborate retrieval methods. With semi-structured data, a mix of retrieval strategies can be leveraged and retrieval performance can be tuned accordingly.

Of relevance to TWA are retrieval methods for knowledge graphs, which fall under two broad categories: semantic parsing and information retrieval. Semantic parsing-based systems convert user queries in natural language to a query language compatible with the underlying data store, such as SPARQL[1], Cypher [13], S-expression [18], KoPL [8], or formulated as a program in a general-purpose language like Python [44, 61]. Meanwhile, information retrieval-based systems extract fragments of knowledge graphs by performing relation classification [27, 63] and vector similarity search [64]. Although the information retrieval approach is generally simpler to implement, it is unable to capture higher-order query operations and often suffers from low recall. Meanwhile, the approach of semantic parsing can capture complex constraints such as numerical comparisons and global-level operations like min and max, but the development of an accurate semantic parser is not as straightforward.

Advances in NLP capabilities of LLMs have made the development of semantic parsers easier, rendering semantic parsing the generally preferred approach. Exemplary methods include logical query construction as a multi-step search problem [19] and draft-then-refine, whereby a candidate logical form is first generated via zero-shot [4] or in-context few-shot learning [10, 37, 44] and is then refined to improve alignment to query intent and the knowledge base's schema.

One example of a RAG system for reticular chemistry is MOF Chatbot [26]. However, it operates on the level of documents and requires LLMs that can handle long context

[1] https://www.w3.org/TR/sparql11-query/

windows. It also might not be able to handle multi-hop questions due to the inherent limitations of document-RAG.

A.2 In-context learning

A technique commonly employed in this usage of LLMs is **in-context few-shot learning** [7], whereby LLMs are provided with a few examples of text input-output pairs at test time, which help align LLM's behaviour with user expectation. In this section, we provide an overview of in-context learning, including its definition, variants of its setup, and factors influencing its performance.

A.2.1 Formulation

Few-shot in-context learning, or in-context learning for short, refers to the ability of a model to learn to perform tasks when provided with few demonstrations, also known as examples, and without updating its weights. Formally, given an input text x, text generation model f, instruction I, and demonstration set $D = \{(x_1, y_1), ..., (x_k, y_k)\}$, the model outputs the label y of x as $y = f(I, D, x)$. Although in-context learning was first reported in the GPT-3 model as part of the line of research that experiments with model scaling [7], it has been shown that smaller models can also be trained to perform in-context learning [40].

A.2.2 Variations

Common variations in in-context learning setups differ in the construction of demonstration set D, output processing, and instruction construction.

A **demonstration set** can be fixed at inference time, generally for simple tasks that can be handled with a small number of demonstrations. Retrieval of a demonstration subset is required for complex tasks that are accompanied by a training set that cannot be reasonably fit into the context window of an LLM. Even for LLMs with long context windows, indiscriminate inclusion of demonstrations in LLM prompts might slow down or introduce noise to inference.

Output processing is any additional treatment applied to LLM output meant to obtain a better result than direct LLM decoding. For example, in a workflow that employs the self-consistency check, an LLM is invoked multiple times to obtain a set of n outputs, of which the majority vote decides the final result [37, 44].

Construction of instruction input into LLMs can be fixed or dynamically adapt to every input query to enhance semantic parsing quality. For instance, knowledge base relations that are the most semantically similar to user questions may be inserted into the LLM prompt to minimise hallucination of non-existent schema elements [44].

A.3 Marie's implementation

Unlike standalone large language model (LLM)-based chatbots that tend to hallucinate scientific facts in low-resource domains, 'Marie' provides fact-oriented responses by augmenting LLM generation with data retrieved from TWA [32, 47]. Furthermore, 'Marie' displays a fine-grained understanding of user intent owing to its semantic parsing component, which can accurately represent logical expressions such as 'boiling point greater than 100°C'—an ability that embedding-based methods found in conventional vector-based retrieval-augmented generation (RAG) systems lack.

A.3.1 Input rewriter

Our system relies on in-context learning to detect physical quantities broken down into magnitudes, units, and quantity types. Refer to Fig. 7 for the structure of the prompt. Unit conversion is done using the Pint library[2], whereby the target unit is looked up based on the quantity type; if no target unit is registered for a quantity type, the quantity will be converted to the SI base units.

Instruction:
Your task is to detect physical quantities in natural language texts based on the examples given. Please ignore physical quantities with no units and respond with a single JSON object exactly, or 'null' if no physical quantities are present.

Input-output examples:
"Find all chemical species with boiling point above 50°C."

```
{
 "template": "Find all chemical species with boiling point
 ↪  above {}",
 "quantities": [
  {"type": "boiling_point", "value": 50, "unit": "degC"}
 ]
}
```

"What is the solubility of C6H6?"
...

Input:
"Find alcohol solvents with a boiling point between 100°C and 120°C."

Figure 7: *LLM prompt for physical quantity detection.*

[2]https://pint.readthedocs.io/en/stable/

Table 1: *Characteristics of questions found in our semantic parsing dataset.*

Criterion	Variants	Example
Answer set cardinality	single	What is the reference zeolite of framework ABW?
	multiple	Find all steroids with molecular weight around 200 g/mol.
Number of constraints	single	Retrieve all MOPs known to have geometric structure (2-bent)x3(3-pyramidal)x2.
	multiple	Find all polymer lubricants with melting point between 200 K and 300 K.
Hop distance	1	Find transport model of oxygen radical.
	2	What are the boiling points of alkenes?
	3	Show all transport models of species that can be used as fuels, indicate which reaction mechanisms the data are derived from.
Query federation	yes	Compare thermo models of hydrocarbons across all mechanisms they appear in.
	no	Show the optimized geometry of H2 calculated using MP2.

A.3.2 Semantic parser

The semantic parsing dataset are manually crafted to cater to diverse information needs within the chemical realm of TWA. The examples display varying levels of complexities, such as single- and multi-hop questions, single- and multi-constraint questions. Notably, we include queries that require federation over multiple SPARQL endpoints, such as "Compare thermo models of hydrocarbons across all mechanisms they appear in.". Here, the information of which species are classified as hydrocarbon is located in the `ontospecies` triplestore, while thermo model data are stored in the `ontokin` triplestore. For a full analysis of question characteristics, see Table 1.

Our prompt template contains three slots for the input query, relevant knowledge base relations and semantic parsing demonstrations. Both semantic parsing examples, and knowledge base relations are retrieved on-demand by vector similarity search, with the user question as the search query. Each semantic parsing demonstration (x_i, y_i) is represented by the embedding of the input query, *i.e.* $f_{\text{S-BERT}}(x_i)$, while each relation r is represented by the embedding of its formatted `rdfs:label` and `rdfs:comment` attributes, *i.e.* $f_{\text{S-BERT}} \circ f_{\text{format}}(a_{\text{rdfs:label}}, a_{\text{rdfs:comment}})$. We retrieve $k_{\text{KG\_relations}} = 10$ relations and $k_{\text{demonstrations}} = 10$ demonstrations without tuning these parameters. See Fig. 8 for an example of how our prompt is constructed.

Instruction:
Your task is to translate the input question to an executable data request based on the provided relations and semantic parsing examples. Please respond with a single JSON object exactly.

Relations:

```
{
 "IRI": "os:hasUse",
 "comment": "A relation between a species and its uses or
 ↪ applications."
}
```

...

Input-output examples:
"What are some common usages of aromatic compounds?"

```
{
 "var2cls": { "ChemicalClass": "os:ChemicalClass", "Use":
 ↪ "os:Use" },
 "entity_bindings": { "ChemicalClass": ["aromatic
 ↪ compound"] },
 "triplestore": "ontospecies",
 "query": "SELECT DISTINCT ?ChemicalClass ?Use WHERE {
 ↪ ?Species os:hasChemicalClass/rdfs:subClassOf*
 ↪ ?ChemicalClass . ?Species os:hasUse ?Use . }"
}
```

"What chemicals can be used to regulate pH?" => ...
...

Input:
"Find chemicals commonly used as fuels"

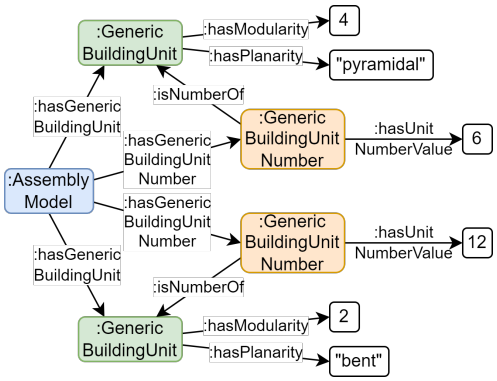
Figure 8: *LLM prompt for semantic parsing.*

A.3.3 Entity linking

Table 2 summarises strategies for entity linking used in Marie, and Fig. 9 illustrates the entity linking logic for entities of class `AssemblyModel`.

Table 2: *Summary of entity linking strategies and their corresponding illustrative examples. Entities to be linked are in bold.*

	Example	
Strategy	Input question	Entity linking logic
Inverted index lookup	"What is the charge of **benzene**?"	Match against all `rdfs:label`, `skos:altLabel`, IUPAC names, molecular formulae, and SMILES strings of `Species` nodes.
Semantic search	"What chemicals can be used to **regulate pH**?"	Perform semantic search over the labels of all `Use` nodes; entity with label "pH regulator" is matched.
RDF subgraph matching	"Find MOPs with assembly model **(4-pyramidal)x6(2-bent)x12**?"	Execute RDF graph query that matches any `AssemblyModel` entities that are linked to exactly six units of 4-pyramidal and twelve units of 2-bent `GenericBuildingUnit` nodes.



(a)

```
PREFIX : <https://www.theworldavatar.com/
↪ kg/ontomops/>

SELECT DISTINCT ?AM WHERE {
  ?AM :hasGenericBuildingUnit ?GBU1;
      :hasGenericBuildingUnitNumber ?GBUNum1.
  ?GBU1 :hasModularity 4; :hasPlanarity
  ↪ "pyramidal".
  ?GBUNum1 :isNumberOf ?GBU1 ;
  ↪ :hasUnitNumberValue 6.

  ?AM :hasGenericBuildingUnit ?GBU2;
      :hasGenericBuildingUnitNumber ?GBUNum2.
  ?GBU2 :hasModularity 2; :hasPlanarity
  ↪ "bent".
  ?GBUNum2 :isNumberOf ?GBU2;
  ↪ :hasUnitNumberValue 12.

  FILTER NOT EXISTS {
    ?AM :hasGenericBuildingUnit ?GBUExclude.
    FILTER ( ?GBUExclude NOT IN ( ?GBU1 ,
    ↪ ?GBU2 ) )
  }
}
```

(b)

Figure 9: *Illustration of RDF subgraph matching as a strategy for entity linking. (a) The ABox subgraph that defines the assembly model (4-pyramidal)x6(2-bent)x12. (b) the SPARQL query to determine the IRI of this entity.*

References

[1] A. O. Adeola, J. O. Ighalo, P. I. Kyesmen, and P. N. Nomngongo. Metal-organic polyhedra (MOPs) as an emerging class of metal-organic frameworks for CO2 photocatalytic conversions: Current trends and future outlook. *J. CO2 Util.*, 80:102664, February 2024. doi:10.1016/j.jcou.2023.102664.

[2] J. Akroyd, S. Mosbach, A. Bhave, and M. Kraft. Universal digital twin – a dynamic knowledge graph. *Data-Centric Eng.*, 2, 2021. doi:10.1017/dce.2021.10.

[3] J. Akroyd, A. Bhave, G. Brownbridge, E. Christou, M. D. Hillman, M. Hofmeister, M. Kraft, J. Lai, K. F. Lee, S. Mosbach, D. Nurkowski, and O. Parry. CReDo technical report 1: Building a cross-sector digital twin. Technical report, Centre for Digital Built Britain (CDBB), 2022. URL https://doi.org/10.17863/CAM.81779.

[4] D. Allemang and J. Sequeda. Increasing the LLM accuracy for question answering: Ontologies to the rescue!, 2024. URL https://arxiv.org/abs/2405.11706.

[5] X. Bai, Y. Xie, X. Zhang, H. Han, and J.-R. Li. Evaluation of open-source large language models for metal–organic frameworks research. *J. Chem. Inf. Model.*, 2024. doi:10.1021/acs.jcim.4c00065.

[6] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Sci. Am.*, 284(5): 34–43, 2001. doi:10.1038/scientificamerican052001-yL7Vw7HIOZ4iSjlnEeVsJ.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, 2020. ISBN 9781713829546. doi:10.48550/arXiv.2005.14165.

[8] S. Cao, J. Shi, L. Pan, L. Nie, Y. Xiang, L. Hou, J. Li, B. He, and H. Zhang. KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6101–6119, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.422.

[9] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi:10.18653/v1/S17-2001.

[10] S. Cheng, Z. Zhuang, Y. Xu, F. Yang, C. Zhang, X. Qin, X. Huang, L. Chen, Q. Lin, D. Zhang, S. Rajmohan, and Q. Zhang. Call me when necessary: LLMs can efficiently and faithfully reason over structured environments, 2024. URL https://arxiv.org/abs/2403.08593.

[11] S. Chong, S. Lee, B. Kim, and J. Kim. Applications of machine learning in metal-organic frameworks. *Coord. Chem. Rev.*, 423:213487, Nov. 2020. doi:10.1016/j.ccr.2020.213487.

[12] F. Farazi, N. B. Krdzavac, J. Akroyd, S. Mosbach, A. Menon, D. Nurkowski, and M. Kraft. Linking reaction mechanisms and quantum chemistry: An ontological approach. *Comput. Chem. Eng.*, 137:106813, June 2020. doi:10.1016/j.compchemeng.2020.106813.

[13] N. Francis, A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault, S. Plantikow, M. Rydberg, P. Selmer, and A. Taylor. Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD/PODS '18. ACM, May 2018. doi:10.1145/3183713.3190657.

[14] M. Gallegos, V. Vassilev-Galindo, I. Poltavsky, Á. Martín Pendás, and A. Tkatchenko. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. *Nat. Commun.*, 15(1):4345, 2024. doi:10.1038/s41467-024-48567-9.

[15] J. Gasteiger. Chemoinformatics: Achievements and challenges, a personal view. *Molecules*, 21(2):151, 2016. doi:10.3390/molecules21020151.

[16] A. C. Ghosh, A. Legrand, R. Rajapaksha, G. A. Craig, C. Sassoye, G. Balázs, D. Farrusseng, S. Furukawa, J. Canivet, and F. M. Wisser. Rhodium-based metal–organic polyhedra assemblies for selective CO_2 photoreduction. *J. Am. Chem. Soc.*, 144(8): 3626–3636, 2022. doi:10.1021/jacs.1c12631.

[17] A. J. Gosselin, C. A. Rowland, and E. D. Bloch. Permanently microporous metal–organic polyhedra. *Chem. Rev.*, 120(16):8987–9014, 2020. doi:10.1021/acs.chemrev.9b00803.

[18] Y. Gu, S. Kase, M. Vanni, B. Sadler, P. Liang, X. Yan, and Y. Su. Beyond i.i.d.: Three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, WWW '21, page 3477–3488, 2021. ISBN 9781450383127. doi:10.1145/3442381.3449992.

[19] Y. Gu, X. Deng, and Y. Su. Don't generate, discriminate: A proposal for grounding language models to real-world environments. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4928–4949, July 2023. doi:10.18653/v1/2023.acl-long.270.

[20] J. Guan, T. Huang, W. Liu, F. Feng, S. Japip, J. Li, J. Wu, X. Wang, and S. Zhang. Design and prediction of metal organic framework-based mixed matrix membranes for CO_2 capture via machine learning. *Cell Rep. Phys. Sci.*, 3(5), 2022. doi:10.1016/j.xcrp.2022.100864.

[21] T. Guo, B. Nan, Z. Liang, Z. Guo, N. Chawla, O. Wiest, X. Zhang, et al. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. *Adv. Neural Inf. Process. Syst.*, 36:59662–59688, 2023. doi:10.48550/arXiv.2305.18365.

[22] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N. V. Chawla, O. Wiest, and X. Zhang. 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, 2024. doi:10.5555/3666122.3668729.

[23] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu. Embracing change: Continual learning in deep neural networks. *Trends Cogn. Sci.*, 24(12):1028–1040, Dec. 2020. doi:10.1016/j.tics.2020.09.004.

[24] IBM. Roborxn. `https://research.ibm.com/science/ibm-roborxn/`, 2021. Last accessed 2 September 2024.

[25] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, and B. Smit. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.*, 6(2):161–169, 2024. doi:10.1038/s42256-023-00788-1.

[26] Y. Kang and J. Kim. ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat. Commun.*, 15(1):4705, 2024.

[27] J. Kim, Y. Kwon, Y. Jo, and E. Choi. KG-GPT: A general framework for reasoning on knowledge graphs using large language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9410–9421, dec 2023. doi:10.18653/v1/2023.findings-emnlp.631.

[28] A. Klami, T. Damoulas, O. Engkvist, P. Rinke, and S. Kaski. Virtual Laboratories: Transforming research with AI. *Data-Centric Eng.*, 5:e19, 2024. doi:10.1017/dce.2024.15.

[29] A. Kondinski, A. Menon, D. Nurkowski, F. Farazi, S. Mosbach, J. Akroyd, and M. Kraft. Automated rational design of metal–organic polyhedra. *J. Am. Chem. Soc.*, 144(26):11713–11728, June 2022. doi:10.1021/jacs.2c03402.

[30] A. Kondinski, J. Bai, S. Mosbach, J. Akroyd, and M. Kraft. Knowledge engineering in chemistry: From expert systems to agents of creation. *Acc. Chem. Res.*, 56(2): 128–139, 2023. doi:10.1021/acs.accounts.2c00617.

[31] A. Kondinski, S. Mosbach, J. Akroyd, A. Breeson, Y. R. Tan, S. Rihm, J. Bai, and M. Kraft. Hacking decarbonization with a community-operated creatorspace. *Chem*, 10(4):1071–1083, 2024. doi:10.1016/j.chempr.2023.12.018.

[32] A. Kondinski, P. Rutkevych, L. Pascazio, D. Tran, F. Farazi, S. Ganguly, and M. Kraft. Knowledge graph representation of zeolitic crystalline materials. Technical Report 321, c4e-Preprint Series, Cambridge, 2024. URL https://como.ceb.cam.ac.uk/preprints/321/. Submitted for publication.

[33] S. Lee, H. Jeong, D. Nam, M. S. Lah, and W. Choe. The rise of metal–organic polyhedra. *Chem. Soc. Rev.*, 50(1):528–555, 2021. doi:10.1039/D0CS00443J.

[34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020. doi:10.48550/arXiv.2005.11401.

[35] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktächel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.*, 33:9459–9474, 2020. doi:10.5555/3495724.3496517.

[36] L. Li, T. Zhou, J. Li, and X. Wang. A machine learning-based decision support framework for energy storage selection. *Chemical Engineering Research and Design*, 181:412–422, 2022. doi:10.1016/j.cherd.2022.04.023.

[37] T. Li, X. Ma, A. Zhuang, Y. Gu, Y. Su, and W. Chen. Few-shot in-context learning on knowledge base question answering. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6966–6980, July 2023. doi:10.18653/v1/2023.acl-long.385.

[38] Y. Luo, S. Bag, O. Zaremba, A. Cierpka, J. Andreo, S. Wuttke, P. Friederich, and M. Tsotsalas. MOF synthesis prediction enabled by automatic data mining and machine learning. *Angew. Chem. Int. Ed.*, 61(19):e202200242, 2022. doi:10.26434/chemrxiv-2021-kgd0h.

[39] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.*, 6(5): 525–535, May 2024. doi:10.1038/s42256-024-00832-8.

[40] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi. MetaICL: Learning to learn in context. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, July 2022. doi:10.18653/v1/2022.naacl-main.201.

[41] P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward, and D. Fairen-Jimenez. Development of a cambridge structural database subset: A collection of metal–organic frameworks for past, present, and future. *Chem. Mater.*, 29(7):2618–2625, Mar. 2017. doi:10.1021/acs.chemmater.7b00441.

[42] S. Mosbach, A. Menon, F. Farazi, N. Krdzavac, X. Zhou, J. Akroyd, and M. Kraft. Multiscale cross-domain thermochemical knowledge-graph. *J. Chem. Inf. Model.*, 60(12):6155–6166, 2020. doi:10.1021/acs.jcim.0c01145.

[43] P. Murray-Rust. Chemistry for everyone. *Nature*, 451(7179):648–651, 2008. doi:10.1038/451648a.

[44] Z. Nie, R. Zhang, Z. Wang, and X. Liu. Code-style in-context learning for knowledge-based question answering. *Proc. AAAI Conf. Artif. Intell*, 38(17):18833–18841, Mar. 2024. doi:10.1609/aaai.v38i17.29848.

[45] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Ł. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Ł. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. d. A. B. Peres, M. Petrov, H. P. d. O. Pinto, P. Michael, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong,

L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph. GPT – 4 Technical Report, 2023.

[46] L. Pascazio, S. Rihm, A. Naseri, S. Mosbach, J. Akroyd, and M. Kraft. Chemical species ontology for data integration and knowledge discovery. *J. Chem. Inf. Model.*, 63(21):6569–6586, Oct. 2023. doi:10.1021/acs.jcim.3c00820.

[47] L. Pascazio, D. Tran, S. D. Rihm, J. Bai, S. Mosbach, J. Akroyd, and M. Kraft. Question-answering system for combustion kinetics. *Proc. Combust. Inst.*, 40(1): 105428, 2024. doi:https://doi.org/10.1016/j.proci.2024.105428.

[48] J. Pérez, M. Arenas, and C. Gutierrez. Semantics and complexity of sparql. *ACM Trans. Database Syst.*, 34(3):1–45, 2009. doi:10.1145/1567274.1567278.

[49] J. J. Perry Iv, J. A. Perman, and M. J. Zaworotko. Design and synthesis of metal–organic frameworks using metal–organic polyhedra as supermolecular building blocks. *Chem. Soc. Rev.*, 38(5):1400–1417, 2009. doi:10.1039/B807086P.

[50] B. Quilitz and U. Leser. Querying distributed rdf data sources with sparql. In *The Semantic Web: Research and Applications*, pages 524–538. Springer, 2008.

[51] N. Rego and D. Koes. 3Dmol·js: molecular visualization with WebGL. *Bioinformatics*, 31(8):1322–1324, Dec. 2014. doi:10.1093/bioinformatics/btu829.

[52] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Nov. 2019. doi:10.18653/v1/D19-1410.

[53] S. D. Rihm, J. Bai, A. Kondinski, S. Mosbach, J. Akroyd, and M. Kraft. Transforming Research Laboratories with Connected Digital Twins. *Nexus*, 1(1):100004, 2024. doi:10.1016/j.ynexs.2024.100004.

[54] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009. doi:10.1561/1500000019.

[55] A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, and R. Q. Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021. doi:/10.1016/j.matt.2021.02.015.

[56] D. Sanmartin. KG-RAG: bridging the gap between knowledge and creativity, 2024. URL https://arxiv.org/abs/2405.12035.

[57] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551, 2023.

[58] K. R. Taylor, R. J. Gledhill, J. W. Essex, J. G. Frey, S. W. Harris, and D. C. De Roure. Bringing chemical data onto the semantic web. *J. Chem. Inf. Model*, 46(3):939–952, 2006. doi:10.1021/ci050378m.

[59] D. Tran, L. Pascazio, J. Akroyd, S. Mosbach, and M. Kraft. Leveraging text-to-text pretrained language models for question answering in chemistry. *ACS Omega*, 9 (12):13883–13896, Mar. 2024. doi:10.1021/acsomega.3c08842.

[60] H. Vardhan, M. Yusubov, and F. Verpoort. Self-assembled metal–organic polyhedra: An overview of various applications. *Coord. Chem. Rev.*, 306:171–194, 2016. doi:10.1016/j.ccr.2015.05.016.

[61] X. Wang, Q. Yang, Y. Qiu, J. Liang, Q. He, Z. Gu, Y. Xiao, and W. Wang. KnowledGPT: Enhancing large language models with retrieval and storage access on knowledge bases, 2023. URL https://arxiv.org/abs/2308.11761.

[62] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, 3(1), Mar. 2016. doi:10.1038/sdata.2016.18.

[63] Y. Wu, N. Hu, G. Qi, S. Bi, J. Ren, A. Xie, and W. Song. Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering. In *Proceedings of The 12th International Joint Conference on Knowledge Graphs*, 2023. doi:10.48550/arXiv.2309.11206.

[64] R. Yang, H. Liu, E. Marrese-Taylor, Q. Zeng, Y. H. Ke, W. Li, L. Cheng, Q. Chen, J. Caverlee, Y. Matsuo, and I. Li. KG-Rank: Enhancing large language models for medical QA with knowledge graphs and ranking techniques, 2024. URL https://arxiv.org/abs/2403.05881.

[65] D. Zhang, W. Liu, Q. Tan, J. Chen, H. Yan, Y. Yan, J. Li, W. Huang, X. Yue, W. Ouyang, D. Zhou, S. Zhang, M. Su, H.-S. Zhong, and Y. Li. ChemLLM: A Chemical Large Language Model, 2024. URL https://arxiv.org/abs/2402.06852.

[66] Y. Zhang, H. Dai, Z. Kozareva, A. Smola, and L. Song. Variational reasoning for question answering with knowledge graph. *Proc. AAAI Conf. Artif. Intell*, 32(1), Apr. 2018. doi:10.1609/aaai.v32i1.12057.

[67] Z. Zheng, O. Zhang, H. L. Nguyen, N. Rampal, A. H. Alawadhi, Z. Rong, T. Head-Gordon, C. Borgs, J. T. Chayes, and O. M. Yaghi. ChatGPT research group for

optimizing the crystallinity of MOFs and COFs. *ACS Cent. Sci.*, 9(11):2161–2170, 2023. doi:10.1021/acscentsci.3c01087.

[68] X. Zhou, A. Eibeck, M. Q. Lim, N. B. Krdzavac, and M. Kraft. An agent composition framework for the J-Park Simulator – a knowledge graph for the process industry. *Comput. Chem. Eng.*, 130:106577, 2019. doi:/10.1016/j.compchemeng.2019.106577.