Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

Knowledge Engineering in Chemistry: From Expert Systems to Agents of Creation

Aleksandar Kondinski [1], Jiaru Bai [1], Sebastian Mosbach [1], Jethro Akroyd [1,2], Markus Kraft [1,2,3,4]

released: December 5, 2022

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

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

[3] School of Chemical and Biomedical Engineering
Nanyang Technological University
62 Nanyang Drive
Singapore, 637459

[4] The Alan Turing Institute
London
United Kingdom

Preprint No. 300

UNIVERSITY OF CAMBRIDGE

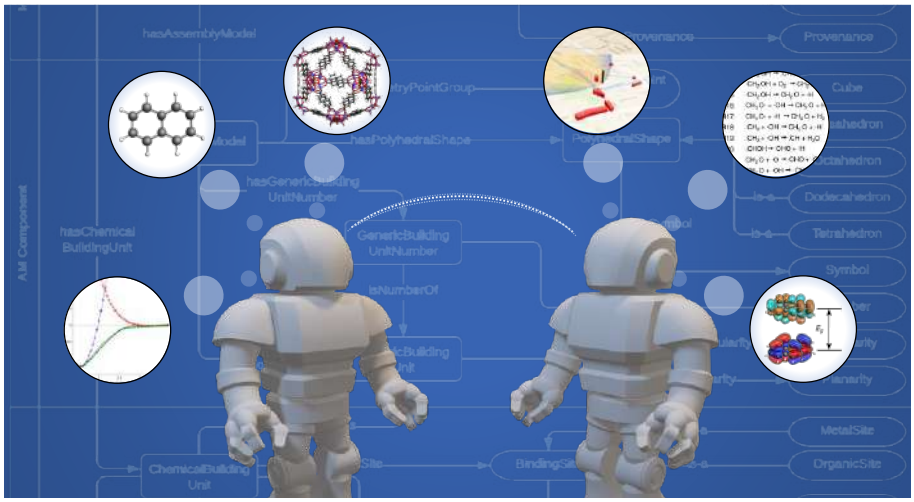
Keywords: Knowledge Engineering, Knowledge Graphs, Digital Chemistry, Chemical Reasoning, Agents

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

This account introduces the history, the core principles of KE, and its applications within chemical research and engineering. In this regard, we first discuss how chemical knowledge is formalised and how a chemist's cognition can be emulated with the help of reasoning algorithms. Following this, we discuss various applications of knowledge graph and agent technology used to solve chemistry-related problems related to molecular engineering, chemical mechanisms, multi-scale modelling, automation of calculations and experiments, and chemist-machine interactions. These developments are discussed in the context of a universal and dynamic knowledge ecosystem, referred to as The World Avatar (TWA).



Highlights

- Knowledge engineering (KE) is a branch of Artificial Intelligence (AI) that emulates the decision-making process of a human expert.
- In a KE approach, instances of knowledge are semantically described with the help of ontologies.
- Software agents are used to facilitate reasoning and execution of various operations.
- KE in chemistry covers many areas, such as rational design, multiscale modelling, lab automation and others.
- Broad chemical knowledge ecosystems are developed through multidisciplinary knowledge graphs.

Contents

1 Introduction

Knowledge is the focal subject of philosophical disciplines such as epistemology and metaphysics. When viewed from the perspective of information science, knowledge is described hierarchically and relative to data, information and wisdom (see Figure 1.a) [50]. The "DIKW" pyramid places data at the bottom of the hierarchy; thus, a data point such as "206.285" can exist without the necessity of having a meaning. A data point that is given relation can become meaningful and thus described as a piece of information. "206.285 g/mol" is a form of information that likely refers to some form of a molar mass. The collection of information in a way that becomes useful is regarded as knowledge. Thus, "ibuprofen" is a "drug" with formula "$C_{16}H_{18}O_{12}$", and molar mass of "206.285 g/mol" would be a form of (minimal) knowledge. Making reasoned and educated judgments or decisions based on knowledge is the basis of wisdom [50].

Knowledge of chemical processes has been documented since antiquity. However, most of this knowledge throughout most of time in human history has remained esoteric, poorly understood and shared among a tiny minority of people [7]. The scientific revolution introduced reasoned structuring of knowledge, which in chemistry followed by the adoption of common chemical representations (*e.g.* symbols, equations, structures), and standards in reporting new chemistry, thus making the subject more widely understandable [7]. Following the second world war, a number of visionary ideas such as the Turing test [57], general problem solvers [41], and universal constructors [26] appeared, that laid the foundations of artificial intelligence (AI), whose further sophistication was realized to depend not only on computing but also on advances in the understanding of human cognition [19].

Knowledge Engineering (KE) is one of the first and most successful branches of AI that emerged in the 1960s. The original aim of KE is to emulate the decision-making process of a human expert [56], consequently leading to the development of many commercialized expert and knowledge management systems, commonly referred to as knowledge based systems [23, 56]. A knowledge-based system is a fundamentally constructed knowledge base which documents knowledge in a machine-readable way and a reasoning component (*i.e.* an inference engine) that, following a request from a user, queries the knowledge base and provides reasoned answers. The knowledge base is commonly maintained by people with domain and knowledge engineering expertise (see Figure 1.b).

In the early days of KE, chemistry formulated problems that KE systems could address and demonstrate potential (*e.g.* meaningful hypothesis generation [58]). However, the AI winters in the 20th century and the general disinterest of the chemistry community in AI systems disrupted the continued development of such systems for chemistry research [25]. However, more recent technological advances (*e.g.*, inexpensive computational power, free software, popularization of programming) reopened interest in this field. A major game-changer was the conceptualization of the Semantic Web by Tim Berners-Lee in the late 1990's-early 2000s [6], which gradually transformed into a knowledge graph (KG) approach [9, 24]. KGs based on the Semantic Web can interlink heterogeneous data and make it accessible to (autonomous) software agents [56]. In addition to querying KGs, these agents were conceptualized as performing different tasks that involve reasoning, learning from humans, and operating on infrastructure to create new things (*e.g.* knowledge, services, and physical items, including chemicals). Owing to these qualities, agents

have been respectively referred to as "intelligent agents" [6], "disciple agents" [56], and "agents of creation" [1].

In this work, we first introduce the conceptual basics of KE: the formalization of chemical knowledge, reasoning, and the route to knowledge systems engineering. We then discuss the beginning of KE in chemistry through examples of legacy expert systems and proceed with the current implementations of a knowledge ecosystem where chemistry plays a central role. The latter technological implementation is illustrated with many examples of navigation through reaction complexity, multiscale modelling, rational design of self-assembled materials, and friendly interactions with chemical KG. Lastly, we outline existing challenges in capturing knowledge dynamics and provide a perspective for future developments.

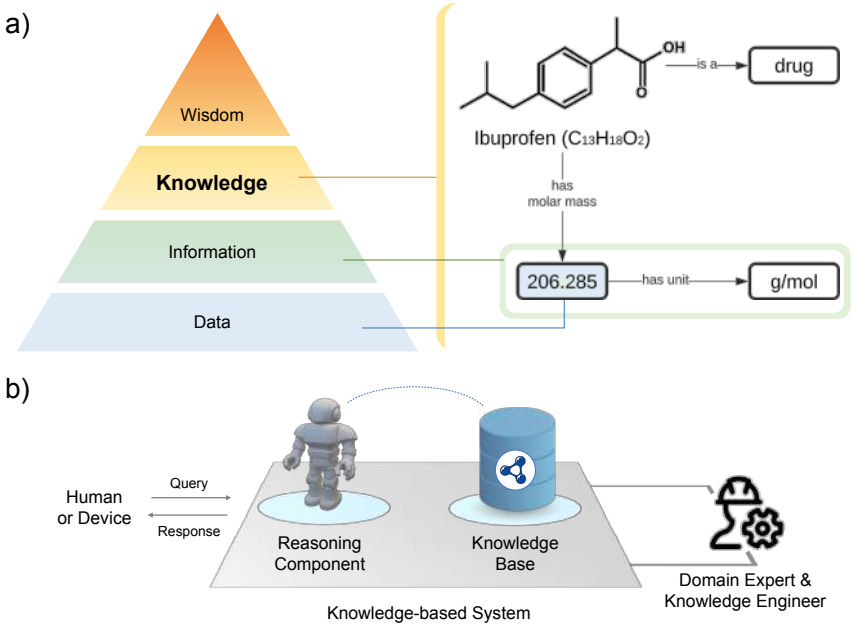


Figure 1: *a) Schematic representation of the "DIKW pyramid" illustrating the meaning of data, information, and knowledge in the chemical context; b) The minimal components of a knowledge-based system.*

2 Chemical Knowledge and Reasoning

How do KE systems emulate expert-like decision-making? To answer this question, we first look into the meaning of chemical knowledge formalization and the navigation through knowledge based on reasoning. Then we outline the main stages of KE project development.

2.1 Formal Representation of Chemical Knowledge

In order to map knowledge, a machine needs to ascribe meaning to data and find a relationship between data points. Documenting data in a relational format, that is, through many interconnected tables, is a straightforward but very restrictive format when it comes to changes in the knowledge structure [44]. Knowledge graphs (KGs) are a different approach where a data point can act as a node that links to other entities in the graph via well-defined relationships. New relations and data can be added to the KG without disturbing the preexisting knowledge structure. Structured data consistency in the framework of KGs is achieved using blueprint networks (*i.e.* schema) that describe how different concepts and properties link to one another. These forms of schemas are commonly referred to as ontologies, defining the terminological box (TBox) of a KG. Knowledge instantiated based on an ontology represents the assertion component (ABox), and it is used in the actual population of the KG. As an ontology, like any directed graph, can be represented as a collection of "triples", that is, subject-predicate-object statements, a database hosting (a part of) the KG is commonly referred to as a triple store. Although knowledge systems can solve real-world problems, many concepts they embody may vary in abstraction. A concept such as "chemical compound" has physical existence; however, "synthon" is a concept that refers to the mental imagery of a compound fragment. In other words, a synthon is not something one can purchase (see Figure 2). When using Semantic Web technology, all concepts and data are linked via unique Internationalized Resource Identifiers (IRIs), making them unambiguously identifiable [6].

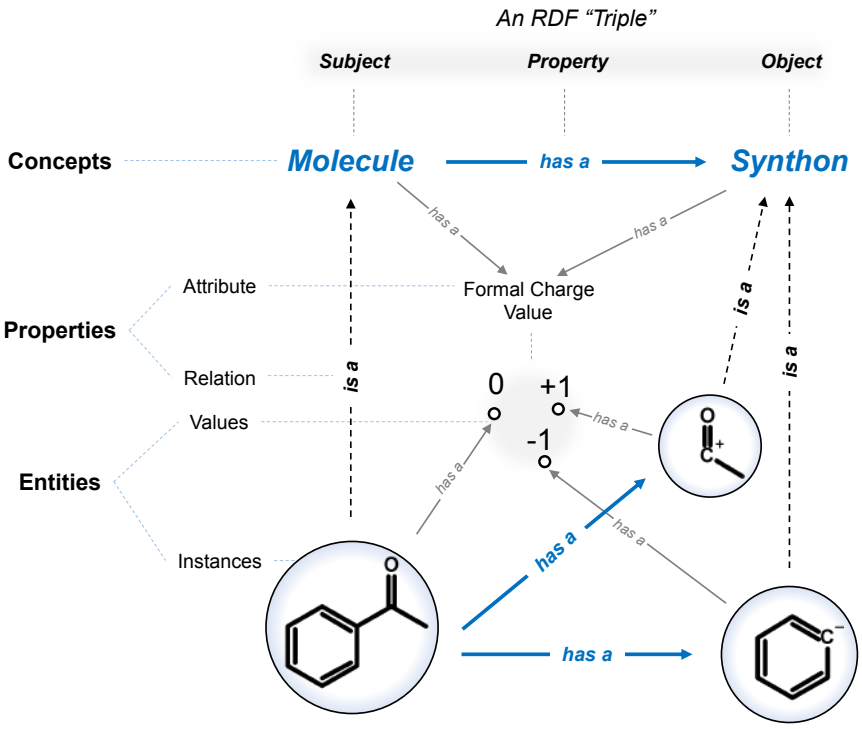


Figure 2: *Mapping the relationship between a molecule (chemical) and a synthon (abstract) concepts, and illustrating them with instances. Description of an RDF triple (top) and ontology stacking (left).*

2.2 Evidence-based Reasoning

Humans typically apply three forms of reasoning [56], such as: i) based on logic and fixed premises (*i.e.* deductive); ii) derived from statistical or anecdotal reference (*i.e.* inductive); iii) based on imagination and best guess (*i.e.* abductive) (see Figure 3.a). Abduction remains broadly accepted as the most challenging to be implemented in AI systems. The different forms of reasoning often manifest themselves in human cognition through mental shortcuts called heuristics [20]. When heuristics are used as part of programming, their utility is primarily to reveal a viable solution by disregarding unlikely solutions. In many expert systems, heuristics have been implemented as deductive reasoners (*i.e.* rules). In our view, this may not be the best practice as it blurs the line between a rule (*i.e.* guaranteed outcome) and a likely (*i.e.* not entirely certain) outcome. Consequently, "rules", especially those in the context of retrosynthetic analysis [53], risk becoming criticized for any possible shortcoming of an expert system implementation.

Over the past decade, machine learning (ML) has increased its dominance in extracting intelligence from chemical data [59]. However, this technique has been particularly successful in domains where clean data is plentiful [38]. ML makes inferences based on associations deriving from data; in principle, ML does not need knowledge or understanding of behaviour to make those associations. As associations are based on statistical significance, ML may also be viewed as a practical implementation of inductive reasoning [45]. On the other hand, KE is developed based on the knowledge and experiences of a domain expert. Thus, algorithms in KE do not need to be pre-trained with data, which is a way forward for cases where data is scarce. Our recent work in metal-organic polyhedra (MOPs) vividly illustrates this as the key algorithm embodies inductive reasoning through set operations, effectively deriving new and rational self-assemblies [31].

2.3 Stages in KE project development

A KE project starts with a genuine problem that a person or a team would like to solve and undergoes three general stages: specification, conceptualization, and implementation (see Figure 3.b) [35]. In the specification stage, the experts do what we would refer to as "epistemological reflection", formulating what they know and how they know it. The team then defines a list of competency questions that the desired KE system is meant to realistically solve. These two aspects effectively narrow down the main focus and goal of the KE system, and they lay the foundations for the conceptualization stage where concept maps are first formulated [52]. A concept map enables a semi-formal representation of knowledge and provides a preliminary insight into the type and number of involved entities. Experts may define or design an algorithm suitable for making inferences and tackling one or more competency questions in conjunction with the concept map [51].

During the implementation stage, the entities of the concept map are ontologized. Experts clean information and instantiate knowledge based on the ontological format. This completes the assertion component that populates the KG. Finally, based on the designed algorithm, an agent capable of traversing the KG and making inferences is programmed. The overall system is then tested and placed in use. Multiple iterations across the three stages are not uncommon, and they often contribute toward better project outcomes [51].

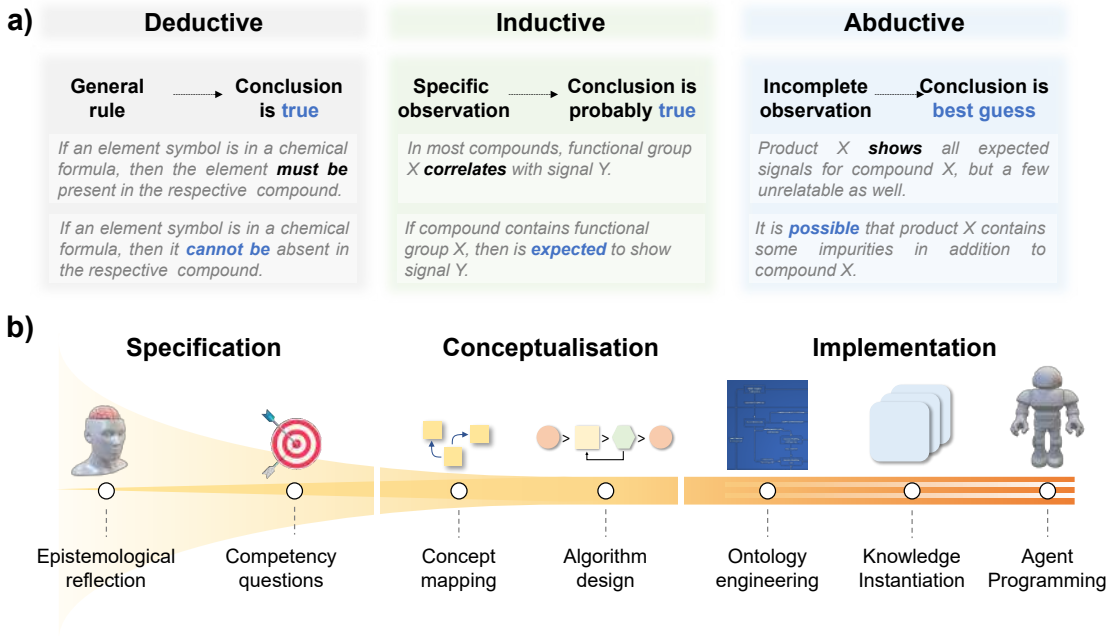


Figure 3: *a) The three main types of reasoning, illustrated with general case scenarios in chemistry. b) The three main stages in KE project development.*

3 Legacy Expert Systems

During the 1960s, two major legacy expert systems essentially pioneered KE. The Dendral project started in 1965 and was developed in the context of NASA's Mars exploration, where real-time molecule detection and elucidation systems were needed. This inspired a group of leading scientists at Stanford University, such as Carl Djerassi, Edward Feigenbaum, and Joshua Lederberg, to automate mass spectrometric species elucidation [34]. Regarding software architecture, Dendral was subdivided into Heuristic Dendral – a component that elucidates species, and Metadendral – a component that learns new rules on how species are fragmented [8]. The two components were envisioned to work in a way that ensures continuous learning. For the development of the Heuristic Dendral, the team developed a general workflow, integrating multiple algorithms for combinatorial exploration of the chemical space and a knowledge base of mass spectrometry fragmentation rules (see Figure 4.a). However, the development of Metadendral has remained challenging. One reason may be that the team attempted to tackle the problem of dynamic knowledge before practical implementation on how to achieve that could be possible.

In 1967, Elias Corey (Harvard University) conceptualized and structured retrosynthesis in the form of five general strategies [10]. In 1969 Corey and Wipke developed the first organic synthesis planning expert system [12] that later became better known as "Logic and Heuristics Applied to Synthetic Analysis" (LHASA) [48]. Over the past decades, LHASA boasted several design strategies and encoded group-protection information, and generally, it served as a blueprint for how to build retrosynthetic expert systems [53]. In 1990, Corey was awarded the Nobel Prize in Chemistry "for his development of the theory and methodology of organic synthesis", with the developments and usage of LHASA

playing an essential role in his Nobel Lecture (see Figure 4.b) [11].

These legacy expert systems in chemistry were followed by many other examples, beautifully discussed and illustrated in the books of Judson [25] and Hammer [23]. The expert systems also placed a technical necessity for finding efficient ways to store and share chemical information, which consequently laid a genuine purpose for developing cheminformatics [61]. Although not broadly acknowledged, some scientists, such as Corey himself, also appreciated the value of KE beyond its implementation. On a deeper level, KE requires chemists to think more generally about their subject and occasionally find more efficient ways to structure chemical knowledge [25, 48].

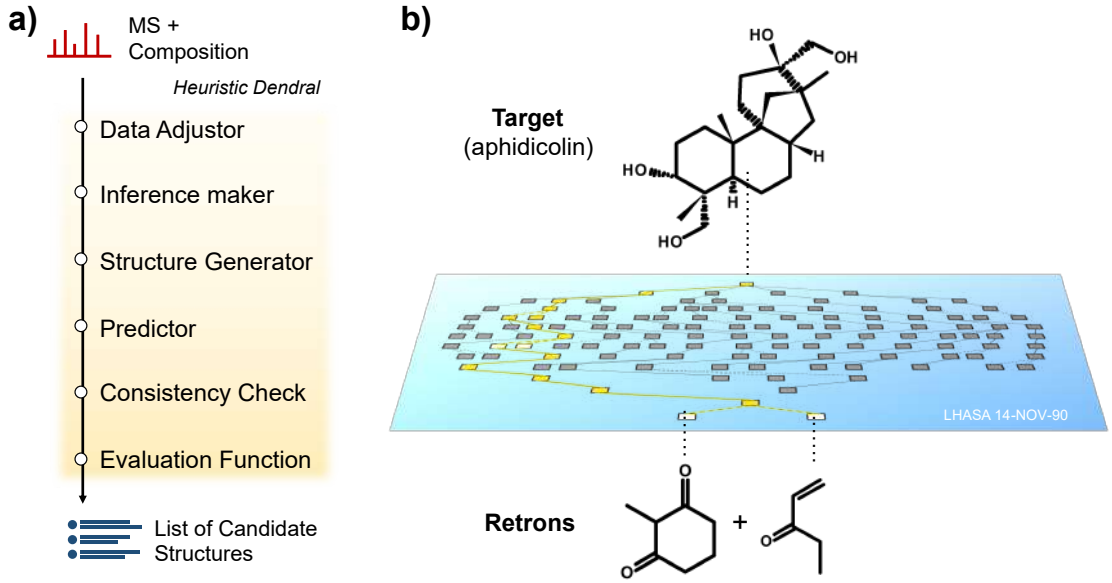


Figure 4: *a) The workflow of Heuristic Dendral; b) Deriving a retrosynthetic pathway to aphidicolin (an antibiotic) using LHASA as illustrated by Corey in his Nobel Lecture [11]*

4 The World Avatar – A Universal Word Model

Not long after conceptualizing the Semantic Web [6], leading cheminformatics researchers realized how beneficial this technology could be to chemists [40, 55]. However, how we can make the broader community benefit from the semantic instantiation of chemistry was envisioned by us in 2010 [32]. In this regard, we outlined the necessity for semantic instantiating of the chemical industry complex and the environmental impact from combustion as two very relevant subjects able to bridge molecular-scale chemistry to real-world macroscale phenomena with socioeconomic, environmental and health impacts. Our early vision was practically implemented as part of our effort to digitalize eco-friendly chemical industry parks [46, 47], such as the one located on Jurong Island (Singapore). The latter attempt initially led to the foundations of the "J-Park Simulator" (JPS) [46, 47]. JPS

embodied many aspects beyond chemical engineering affecting productivity and environment, such as logistics, infrastructure, energy usage, and waste among others [14, 63, 64]. By building digital tools to represent these aspects, it was realized that they are more widely applicable than just to chemical parks but more broadly to the world at large, leading to the extension and transition of the JPS to the ongoing "The World Avatar" project (TWA), an effort to create an all encompassing universal world model [2, 15]. Although TWA (see Figure 5) at first sight may appear as a bold and over-ambitious project, recently more leading figures in computer science and environmental studies have embraced the world-centric idea as a necessity for the progression of contemporary AI [5, 62].

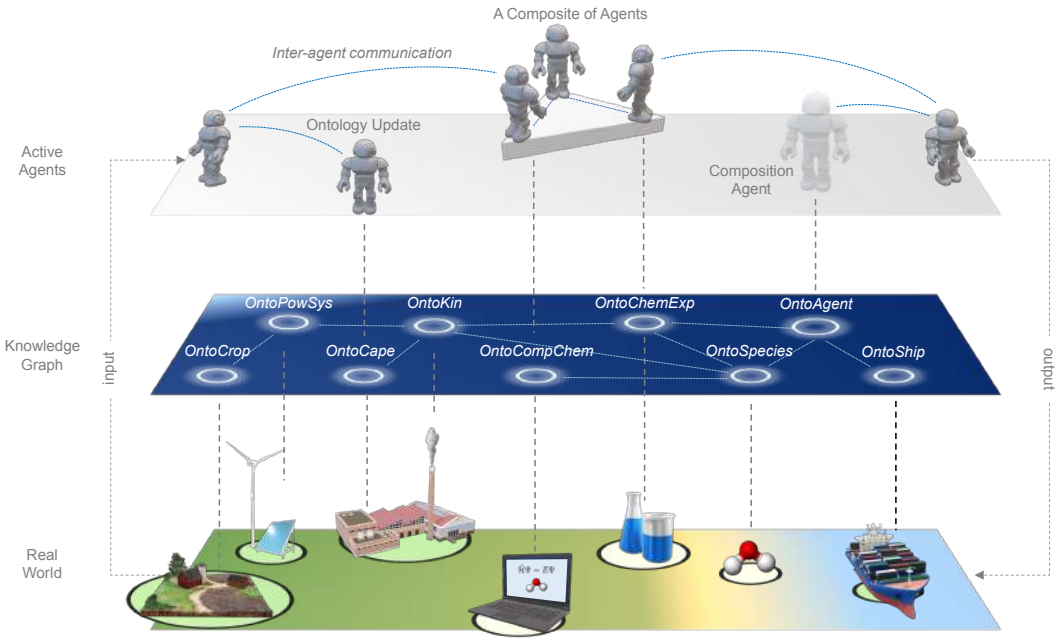


Figure 5: *The three layers of TWA (www.theworldavatar.com) digital twin of the real world.*

Digital twins are an emerging technology that provides a real-time representation of real-world phenomena, assisting decision-making by exploration of what-if scenarios [54]. In this regard, TWA (see Figure 5) has been conceptualized as a universal digital twin based on the Semantic Web, where a universal KG maps the real world. TWA follows the FAIR principles of linked data, that is, all stored knowledge is findable, accessible, interoperable and reusable [60]. On the "top" of the knowledge layer, TWA integrates a layer of active agents that operate on it [64]. These agents differ from the classical inference engines employed in expert systems, and they have a number of different tasks such as i) implementing information pipelines; ii) sending signals back to the real world; iii) providing an interface to computational models; iv) restructuring the KG by adding instances, concepts and relationships; v) discovering, combining, and composing new agents capable of performing new and on-demand tasks [2, 64]. At the same time, agents are also represented through concepts, instances and properties in the knowledge graph. The latter feature makes agents findable and enables the possibility of solving complex

tasks through inter-agent communication and collaboration [64].

5 Chemistry as Part of a Knowledge Ecosystem

Currently, a number of high-tech companies, Google, IBM, Microsoft, Facebook, and eBay, have been implementing KGs on an industrial scale [42]. In the context of the pharmaceutical industry, AstraZeneca is a company that openly leads the way on KGs as part of their drug discovery [21]. This section discusses the development of a chemistry KG and related agents as part of TWA knowledge ecosystem [14, 15].

5.1 Chemical Species

OntoSpecies is an ontology that describes unique chemical species and their chemical properties in TWA. In TWA, species are assigned IRIs, allowing their unique identification [17]. OntoSpecies plays a central role, enabling the linking of species to instances and concepts deriving from other ontologies in TWA KG (see Figure 6.a). A chemical species in OntoSpecies has a recorded molecular formula, charge, molecular weight, and spin multiplicity. Species that are based on different isotopes, charges, and spin states are treated as different chemical species. By assigning different IRIs to species, OntoSpecies becomes relevant for the digital representation of isotope labelling experiments, redox and electrochemically driven processes, and photochemistry. Considering reactor simulations, OntoSpecies records standard enthalpy of formation along with its contextual information such as reference temperature, state and provenance [17].

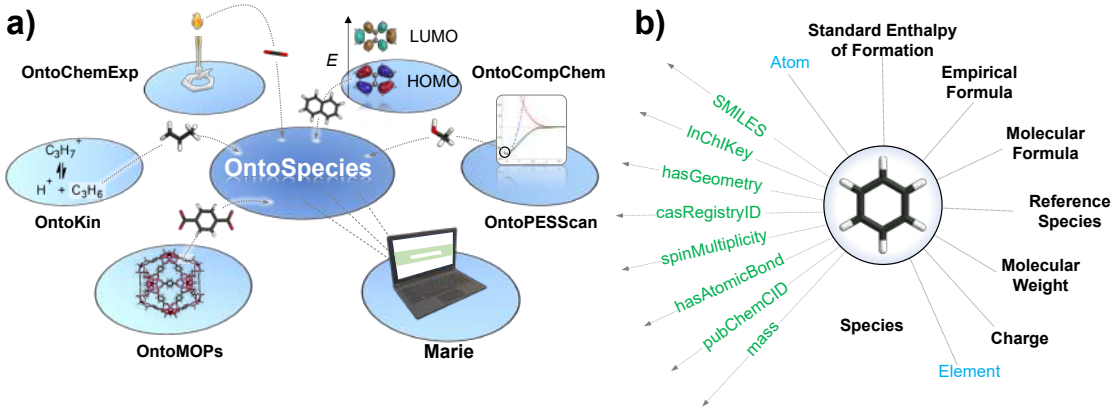


Figure 6: *a) Connection of OntoSpecies to other segments of TWA KG; b) Key OntoSpecies (black) and external (blue) concepts, along with a number of properties (green) used to describe chemical species in TWA KG.*

In addition to the IRIs, chemical species in TWA are labelled with common cheminformatic identifiers [37], such as InChI, InChIKey, CAS registry number, PubChemCID, and SMILES (see Figure 6.b). These labelling identifiers facilitate searching for additional

information on external resources. OntoSpecies also records the molecular geometry of different species semantically, meaning that every bond and atom is uniquely identified with an IRI. The information on molecular geometry can be used as an initial guess of the geometry for quantum chemical calculations, while unique identification of bonds and atoms is used for the identification of geometric changes between calculations. For many organics, the geometric information can be automatically generated by translation from InChI or SMILES identifiers using OpenBabel [43] and by pre-optimization using force fields. However, the latter approach is not always suitable for inorganics and thus, storing a pre-curated geometry can be an advantage.

5.2 Navigating Reaction Complexity

In chemistry, many reactions and self-assembly processes starting with simple molecular precursors often lead to a rich variety of chemical species and (meta-stable) intermediates. The speciation of molecular metal oxides in solution [28], or the formation of nanoparticulate carbonaceous materials [36] are examples of such chemistries. Understanding and modelling these chemistries require a grasp of kinetic and thermodynamic factors. Motivated to model these factors semantically on chemical species, our group developed and interlinked the OntoKin [16] and OntoCompChem [33] ontologies (see Figure 7.a).

OntoKin is an ontology that represents reaction mechanisms in alignment with nomenclature standards used in computer-aided process design [16]. In a chemical process, a reaction mechanism constitutes a set of stochiometric reactions involving different chemical species. In OntoKin, a reaction is described through products and reactants that are further described through different thermodynamic and transport model concepts and identified via OntoSpecies IRIs. Depending on where the reaction occurs, OntoKin introduces further specifications (*e.g.* in gas, on the surface, *etc.*). The reaction rate of each reaction is presented based on Arrhenius-type rate models, which are used to compute rate coefficients. As a single reaction mechanism may consist of many different reactions, OntoKin, in conjunction with OntoSpecies, can provide a facile and unambiguous comparison between other kinetic, thermodynamic, or transport models reported in the literature [18].

The OntoCompChem ontology represents the input and output of density functional theory (DFT), currently mainly focused on molecular systems [33]. OntoCompChem has been developed based on the semantic concepts specified in the CompChem convention of Chemical Markup Language (CML) [49]. A calculation in OntoCompChem is described in terms of a) its objective (*e.g.* single point calculation, geometry optimization, or a frequency calculation); b) the software it uses (*e.g.* Gaussian16); c) the employed theoretical level in terms of functional and basis set (*e.g.* B3LYP, 6-31G(d)); d) the overall charge, and spin polarization. The ontology also represents the calculated frontier orbitals and the final converged self-consistent field (SCF) energy. For geometry optimizations, the final optimized geometry is represented, while for frequency calculations, it stores the zero-point energy correction and a full list of the computed vibrational frequencies linking back to the stationary geometry and calculation it corresponds to.

A linking agent automates the creation of links between reactions, species, and DFT calculations [17]. Such an agent is needed because a reaction mechanism in OntoKin can

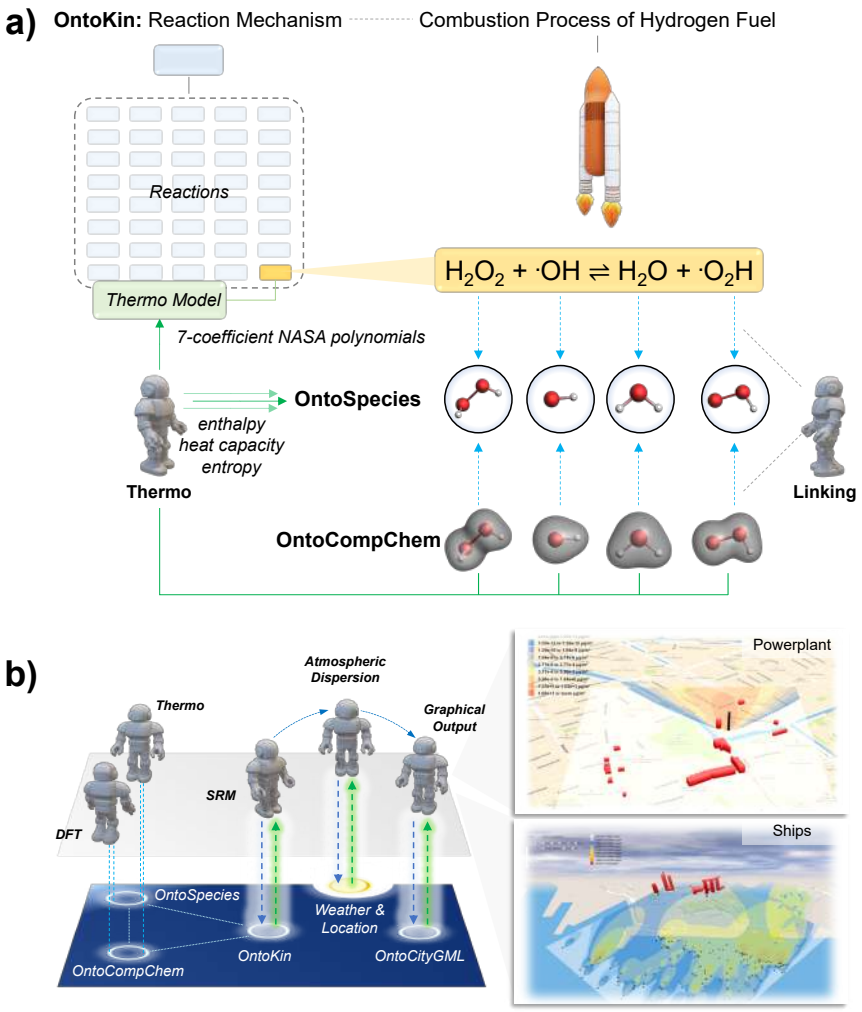


Figure 7: *a) Automated linking between OntoSpecies, OntoKin, and OntoCompChem. b) Multiscale modelling of pollution starting with fuel molecules stored in OntoKin.*

easily involve thousands of species and tens of thousands of reactions [18]. The linking allows zooming into a mechanism, its reactions, and involved species. An example may be the combustion of clean hydrogen fuel used as a rocket propellant, which involves 10 species and 40 elementary reactions, one of which is $2H_2 + O_2 \rightleftharpoons H_2O$ (see Figure 7.a). For existing DFT calculations, a Thermo agent instantiates enthalpy, heat capacity, and entropy factors back to the involved species and 7-coefficient NASA polynomials to the reaction. If experimental data is provided as a concept, reaction mechanisms can be linked to it, and agents wrapping our custom-made software can do sensitivity analysis and calibration, providing a quantitative explanation of experimental phenomena [3]. Finally, a workflow of agents (see Figure 7.b) that perform: i) DFT calculations; ii) thermodynamic data analysis; iii) stochastic model calculations predicting particle formation from fuels in engines; iv) atmospheric dispersion modelling based on real-time weather data, and graphical output based on physical infrastructure are showcased to predict the dispersion

of particle pollution in urban areas [39]. The relevance of such systems is in digital urban planning.

5.3 Automating Rational Design of Self-Assembled Materials

Metal-organic polyhedra (MOPs) are assemblies made of organic and metal-based chemical building units (CBUs) resembling the shape of regular polyhedra [22]. MOPs and other cage-like structures are rationally designed by domain experts. To design new MOPs, an expert requires the consideration of both chemical and spatial complementarity factors. Insights from didactical research with toys have shown that children do not need any formal foreknowledge on geometric aspects to build polyhedral models [29, 30], which implied that some form of mental imagery is involved as part of the overall reasoning. These considerations inspired the conceptualization of assembly models (AMs) and generic building units (GBUs) as mental blueprints involved in the rational design of MOPs from sets of available CBUs [31]. The latter concepts were encoded in the OntoMOPs ontology, where the CBUs were further instantiated as species based on the OntoSpecies ontology. The MOP discovery agent was based on an algorithm that performs set operations revealing which CBUs can be meaningfully combined without causing undesired strains. The study involved 151 experimentally reported MOPs built from 137 unique CBUs, which were effectively clustered in 18 AMs and 7 GBUs, respectively. The MOP discovery agent showed that up to 1418 new MOP instances could be rationally designed, some of which are confirmed by domain experts. The latter aspect is a considerable advantage as it allows more focused and efficient exploration of chemical spaces through calculations and experiments [31]. The rational projection estimate is a significant reduction in the combinatorial chemical space, which in this case amounts to about 80 000 possibilities [31].

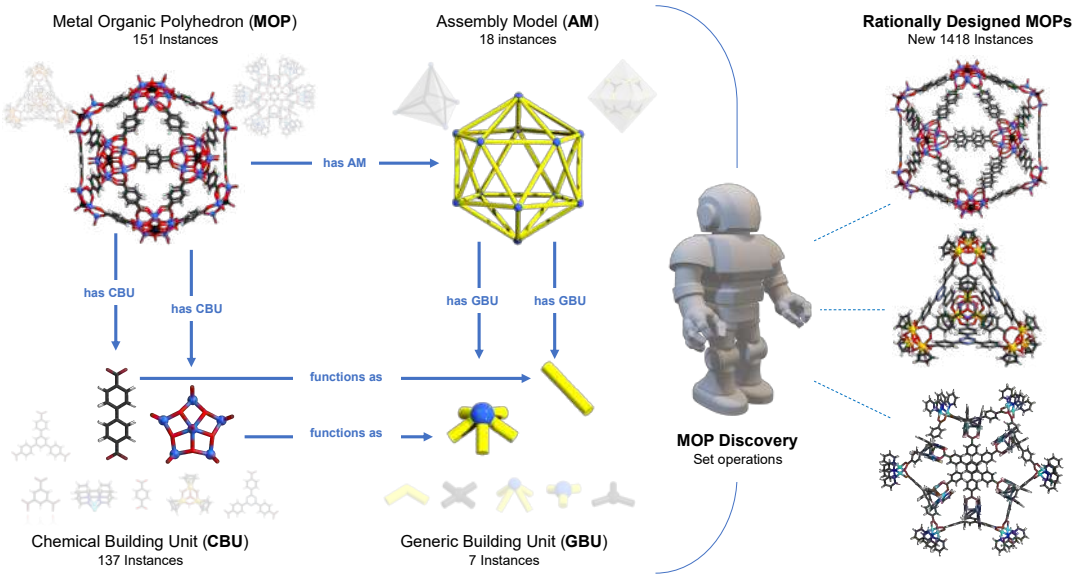


Figure 8: *Key concepts in OntoMOPs (left) and examples of newly rationally designed MOPs (right).*

5.4 Marie – Enabling User-friendly Interaction with TWA KG

Querying a KG requires the use of a query language (*e.g.* SPARQL) and awareness of how the knowledge has been structured in that KG. These factors make exploration of the KG less convenient for users who lack the foreknowledge; thus, a more user-friendly interface with the KG is desired. In the context of chemistry within TWA, "Marie" is a question-answering interface that is aimed at allowing users to type their questions in their natural language, which are then translated behind the scenes into machine readable queries [65, 66]. To achieve this, Marie implements natural language processing (NLP), and a network of agents that can identify the topic, the type of question and the entities the user is asking about. Once clarified, the agents pass the information to an ontology lookup agent that passes the information to a SPARQL construction agent, which then queries the KG and returns information to the user [65]. A typical example is when a user asks Marie to show models of aromatic hydrocarbons (see Figure 9) [65].

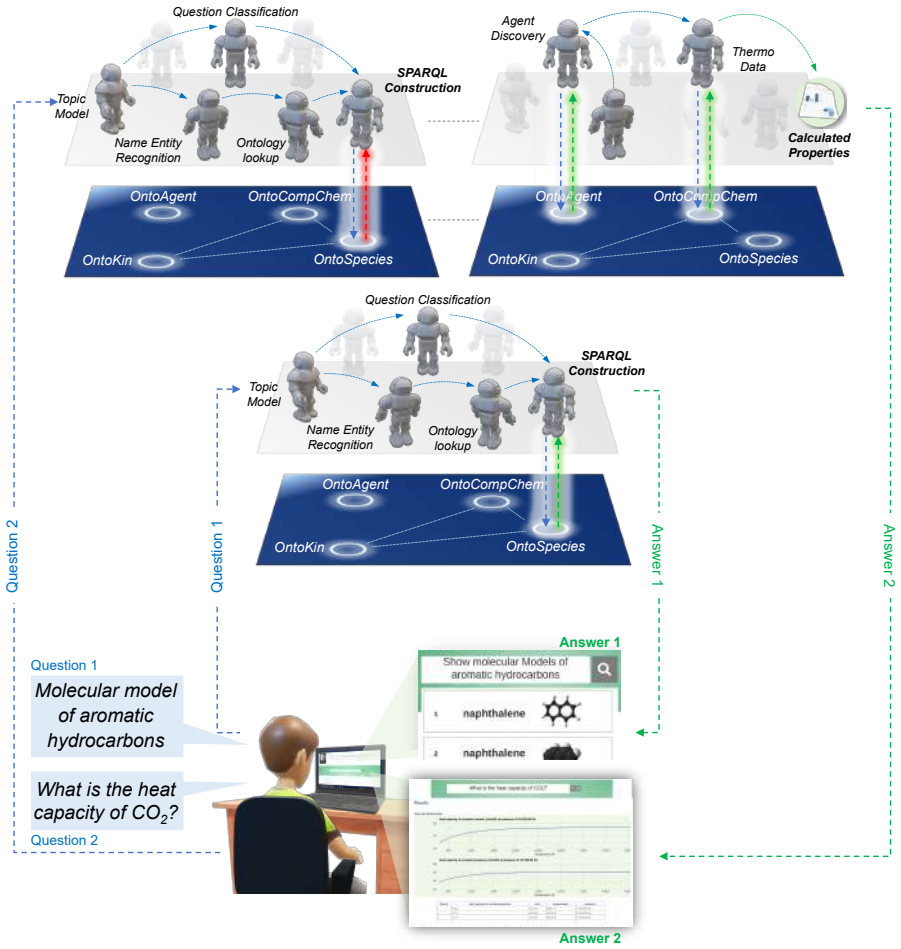


Figure 9: *Marie's back-end operations involved in querying information that is in TWA KG and one that it is generated through agent operation.*

As it is challenging to store all knowledge, while much knowledge can be indirectly inferred or calculated, Marie takes a different circuit when the answer is not found in the

KG. In this case, an agent that discovers agents is activated, who then allocates an appropriate agent for the task. The appropriate agent can then query the graph and calculate information. An example would be a question to display the heat capacity of CO_2, where the Thermo agent can calculate it from instances in OntoCompChem.

6 Real-time Knowledge Dynamics

Many discoveries or outcomes in chemical research depend on other outcomes in the field or, more generally, from the real world. For instance, when a chemist plans the synthesis and the characterization of a new chemical, what instrumental infrastructure will be used is dependent on the nature of the chemical target. Further on, the discovery of new self-assembled material may depend on the discovery of a suitable building block precursor. Navigating dependencies is a complex and challenging task; however, its successful emulation provides an opportunity to realize autonomous laboratory systems [4], and even future AI Scientists [27].

The dynamic data-driven applications systems (DDDAS) [13], which originated from control systems, is a research paradigm focused on tackling this challenge. It seeks to provide data context to improve decision-making in dynamic and complex environments. Using the KE approach, our group has worked on a derived information framework as a knowledge-graph-native solution to represent how pieces of information depend on others in a dynamic knowledge graph. The framework represents complex and interconnected phenomena as a directed graph of computational or physical activities, with agents serving as executable knowledge components. Once dependencies between objects are created, the framework propagates the effects induced by changes in the source information. We envisage this framework providing solutions to the aforementioned difficulties in the chemistry domain.

7 Summary and Outlook

In this account article, we have summarised the developments of KE in chemical research. From its beginnings, KE in chemistry has been going through a very challenging path and generally has retained its relevance through the engineering of expert and knowledge management systems [25]. The beginning of the Semantic Web opened a new paradigm for KE, effectively removing any boundary in terms of knowledge representation and reasoning. By building a KG that includes agents, a new ecosystem for chemical knowledge creation and exploitation has been enabled, allowing the implementation of inductive and, hopefully, over time, abductive reasoning algorithms as well. These aspects have been recently showcased to scale up discoveries [31], and likely are a path forward to chemical intelligence amplification.

Through multiple examples of our work, we see that a KG with its agents can combine complex decision-making processes with the generation of new knowledge from calculations, external sources, and in the near future, autonomous experiments [4]. Based on

this, it is not difficult to envision more sophisticated combinations of agents involved in the conceptualisation and creation of new molecules and materials in near future. Making chemical knowledge part of a single knowledge ecosystem enables efficient inferencing across disciplines, scales, and depths in terms of chemical space exploration. In this regard, KE can be a true enabler of systems-level research frontiers such as materiomics and systems chemistry. Much of the success in the latter will be critically dependent on the wisdom of the human experts in structuring knowledge and their capability of developing rational agents. Finally, we expect that the enormous progress in machine learning combined with ideas of KE will further expand the knowledge space.

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. AK and MK thank the Humboldt Foundation (Berlin, Germany) and the Isaac Newton Trust (Cambridge, UK) for a Feodor Lynen Fellowship. JB acknowledges financial support provided by CSC Cambridge International Scholarship from Cambridge Trust and China Scholarship Council. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

[1] Agents of creation. *The Economist*, 2003. URL https://www.economist.com/science-and-technology/2003/10/09/agents-of-creation.

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

[3] J. Bai, R. Geeson, F. Farazi, S. Mosbach, J. Akroyd, E. J. Bringley, and M. Kraft. Automated calibration of a poly(oxymethylene) dimethyl ether oxidation mechanism using the knowledge graph technology. *J. Chem. Inf. Model*, 61(4):1701–1717, 2021.

[4] J. Bai, L. Cao, S. Mosbach, J. Akroyd, A. A. Lapkin, and M. Kraft. From platform to knowledge graph: evolution of laboratory automation. *JACS Au*, 2(2):292–309, 2022.

[5] P. Bauer, B. Stevens, and W. Hazeleger. A digital twin of earth for the green transition. *Nature Climate Change*, 11(2):80–83, 2021.

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

[7] W. H. Brock. *Norton history of chemistry*. WW Norton, 1993.

[8] B. G. Buchanan and E. A. Feigenbaum. DENDRAL and Meta-DENDRAL: Their applications dimension. *Artif. Intell.*, 11(1-2):5–24, 1978.

[9] X. Chen, S. Jia, and Y. Xiang. A review: Knowledge reasoning over knowledge graph. *Expert Syst. Appl.*, 141:112948, 2020.

[10] E. J. Corey. General methods for the construction of complex molecules. *Pure Appl. Chem.*, 14(1):19–38, 1967.

[11] E. J. Corey. The logic of chemical synthesis: multistep synthesis of complex carbogenic molecules (Nobel lecture). *Angew. Chem. Int. Ed.*, 30(5):455–465, 1991.

[12] E. J. Corey and W. T. Wipke. Computer-assisted design of complex organic syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. *Science*, 166(3902):178–192, 1969.

[13] F. Darema. Dynamic data driven applications systems: A new paradigm for application simulations and measurements. In *International conference on computational science*, pages 662–669. Springer, 2004.

[14] A. Eibeck, M. Q. Lim, and M. Kraft. J-Park Simulator: An ontology-based platform for cross-domain scenarios in process industry. *Comput. Chem. Eng.*, 131:106586, 2019.

[15] A. Eibeck, A. Chadzynski, M. Q. Lim, K. Aditya, L. Ong, A. Devanand, G. Karmakar, S. Mosbach, R. Lau, I. A. Karimi, E. Y. Foo, and M. Kraft. A parallel world framework for scenario analysis in knowledge graphs. *Data-Centric Eng.*, 1, 2020.

[16] F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca, and M. Kraft. Ontokin: An ontology for chemical kinetic reaction mechanisms. *J. Chem. Inf. Model*, 60(1):108–120, 2019.

[17] 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, 2020.

[18] F. Farazi, M. Salamanca, S. Mosbach, J. Akroyd, A. Eibeck, L. K. Aditya, A. Chadzynski, K. Pan, X. Zhou, S. Zhang, M. Q. Lim, and M. Kraft. Knowledge graph approach to combustion chemistry and interoperability. *ACS omega*, 5 (29):18342–18348, 2020.

[19] R. M. French. The Turing test: the first 50 years. *Trends in cognitive sciences*, 4(3): 115–122, 2000.

[20] G. Gigerenzer. Why heuristics work. *Perspectives on psychological science*, 3(1): 20–29, 2008.

[21] A. Gogleva, D. Polychronopoulos, M. Pfeifer, V. Poroshin, M. Ughetto, M. J. Martin, H. Thorpe, A. Bornot, P. D. Smith, B. Sidders, J. R. Dry, M. Ahdesmäki, U. McDermott, E. Papa, and K. C. Bulusu. Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer. *Nature Communications*, 13:1667, 2022.

[22] A. J. Gosselin, C. A. Rowland, and E. D. Bloch. Permanently microporous metal–organic polyhedra. *Chem. Rev.*, 120(16):8987–9014, 2020.

[23] M. C. Hemmer. *Expert systems in chemistry research*. CRC Press, 2007.

[24] S. Ji, S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Netw. Learn. Syst.*, 33(2):494–514, 2021.

[25] P. Judson. *Knowledge-based Expert Systems in Chemistry: Artificial Intelligence in Decision Making*, volume 15. Royal Society of Chemistry, 2019.

[26] J. G. Kemeny. The universal constructor: Theory of self-reproducing automata. John von Neumann. Edited by Arthur W. Burks. University of Illinois Press, Urbana, 1966. 408 pp., illus. $10. *Science*, 157(3785):180–180, 1967.

[27] H. Kitano. Nobel turing challenge: creating the engine for scientific discovery. *npj Systems Biology and Applications*, 7(1):1–12, 2021.

[28] A. Kondinski. Metal–metal bonds in polyoxometalate chemistry. *Nanoscale*, 13 (32):13574–13592, 2021.

[29] A. Kondinski and T. N. Parac-Vogt. Programmable interlocking disks: bottom-up modular assembly of chemically relevant polyhedral and reticular structural models. *J. Chem. Educ.*, 96(3):601–605, 2019.

[30] A. Kondinski, J. Moons, Y. Zhang, J. Bussé, W. De Borggraeve, E. Nies, and T. N. Parac-Vogt. Modeling of nanomolecular and reticular architectures with 6-fold grooved, programmable interlocking disks. *J. Chem. Educ.*, 97(1):289–294, 2020.

[31] 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.*, 2022.

[32] M. Kraft and S. Mosbach. The future of computational modelling in reaction engineering. *Philos. Trans. R. Soc. A*, 368(1924):3633–3644, 2010.

[33] N. Krdzavac, S. Mosbach, D. Nurkowski, P. Buerger, J. Akroyd, J. Martin, A. Menon, and M. Kraft. An ontology and semantic web service for quantum chemistry calculations. *J. Chem. Inf. Model*, 59(7):3154–3165, 2019.

[34] J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi. Applications of artificial intelligence for chemical inference. I. Number of possible organic compounds. Acyclic structures containing carbon, hydrogen, oxygen, and nitrogen. *J. Am. Chem. Soc.*, 91(11):2973–2976, 1969.

[35] M. F. López, A. Gómez-Pérez, J. P. Sierra, and A. P. Sierra. Building a chemical ontology using methontology and the ontology design environment. *IEEE Intell. Syst.*, 14(1):37–46, 1999.

[36] J. W. Martin, M. Salamanca, and M. Kraft. Soot inception: Carbonaceous nanoparticle formation in flames. *Prog. Energy Combust. Sci.*, 88:100956, 2022.

[37] A. Menon, N. B. Krdzavac, and M. Kraft. From database to knowledge graph—using data in chemistry. *Curr. Opin. Chem.*, 26:33–37, 2019.

[38] J. B. Mitchell. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 4(5):468–481, 2014.

[39] 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.

[40] P. Murray-Rust. Chemistry for everyone. *Nature*, 451(7179):648–651, 2008.

[41] A. Newell, J. C. Shaw, and H. A. Simon. Report on a general problem solving program. In *IFIP congress*, volume 256, page 64. Pittsburgh, PA, 1959.

[42] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done. *Queue*, 17(2):48–75, 2019.

[43] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An open chemical toolbox. *J. Cheminform.*, 3(1):1–14, 2011.

[44] T. O'Donnell. *Design and use of relational databases in chemistry*. CRC Press, 2008.

[45] E. Orłowska. Semantic analysis of inductive reasoning. *Theoret. Comput. Sci.*, 43: 81–89, 1986.

[46] M. Pan, J. Sikorski, C. A. Kastner, J. Akroyd, S. Mosbach, R. Lau, and M. Kraft. Applying Industry 4.0 to the Jurong Island eco-industrial park. *Energy Procedia*, 75:1536–1541, 2015.

[47] M. Pan, J. Sikorski, J. Akroyd, S. Mosbach, R. Lau, and M. Kraft. Design technologies for eco-industrial parks: From unit operations to processes, plants and industrial networks. *Appl. Energy*, 175:305–323, 2016.

[48] D. A. Pensak and E. J. Corey. LHASA — logic and heuristics applied to synthetic analysis. In W. T. Wipke and W. J. Howe, editors, *Computer-Assisted Organic Synthesis*, chapter 1, pages 1–32. ACS Publications, 1977.

[49] W. Phadungsukanan, M. Kraft, J. A. Townsend, and P. Murray-Rust. The semantics of chemical markup language (CML) for computational chemistry: CompChem. *J. Cheminform.*, 4(1):1–16, 2012.

[50] J. Rowley. The wisdom hierarchy: representations of the DIKW hierarchy. *J. Inf. Sci.*, 33(2):163–180, 2007.

[51] S. Staab and R. Studer. *Handbook on ontologies*. Springer Science & Business Media, 2010.

[52] R. R. Starr and J. M. P. De Oliveira. Concept maps as the first step in an ontology construction method. *Inf. Syst.*, 38(5):771–783, 2013.

[53] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, and B. A. Grzybowski. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.*, 55(20):5904–5937, 2016.

[54] F. Tao and Q. Qi. Make more digital twins. *Nature*, 573:490–491, 2019.

[55] 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.

[56] G. Tecuci, D. Marcu, M. Boicu, and D. A. Schum. *Knowledge engineering: Building cognitive assistants for evidence-based reasoning*. Cambridge University Press, 2016.

[57] A. M. Turing. Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer, 2009.

[58] D. Waltz and B. G. Buchanan. Automating science. *Science*, 324(5923):43–44, 2009.

[59] J. M. Weber, Z. Guo, C. Zhang, A. M. Schweidtmann, and A. A. Lapkin. Chemical data intelligence for sustainable chemistry. *Chem. Soc. Rev.*, 2021.

[60] 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*, 3:160018, 2016.

[61] P. Willett. Chemoinformatics: a history. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(1):46–56, 2011.

[62] M. Wooldridge. What is missing from contemporary AI? The World. *Intell. Comp.*, 9847630, 2022.

[63] L. Zhou, C. Zhang, I. A. Karimi, and M. Kraft. An ontology framework towards decentralized information management for eco-industrial parks. *Comput. Chem. Eng.*, 118:49–63, 2018.

[64] 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.

[65] X. Zhou, D. Nurkowski, S. Mosbach, J. Akroyd, and M. Kraft. Question answering system for chemistry. *J. Chem. Inf. Model*, 61(8):3868–3880, 2021.

[66] X. Zhou, D. Nurkowski, A. Menon, J. Akroyd, S. Mosbach, and M. Kraft. Question answering system for chemistry—a semantic agent extension. *Digit. Chem. Eng.*, 3: 100032, 2022.