Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2024, **3**, 2070

Knowledge graph representation of zeolitic crystalline materials†

Aleksandar Kondinski,‡[a] Pavlo Rutkevych,‡[a] Laura Pascazio,[a] Dan N. Tran,[a] Feroz Farazi,[b] Srishti Ganguly[a] and Markus Kraft *[abcdef]

Zeolites are complex and porous crystalline inorganic materials that serve as hosts for a variety of molecular, ionic and cluster species. Formal, machine-actionable representation of this chemistry presents a challenge as a variety of concepts need to be semantically interlinked. This work demonstrates the potential of knowledge engineering in overcoming this challenge. We develop ontologies OntoCrystal and OntoZeolite, enabling the representation and instantiation of crystalline zeolite information into a dynamic, interoperable knowledge graph called The World Avatar (TWA). In TWA, crystalline zeolite instances are semantically interconnected with chemical species that act as guests in these materials. Information can be obtained *via* custom or templated SPARQL queries administered through a user-friendly web interface. Unstructured exploration is facilitated through natural language processing using the Marie System, showcasing promise for the blended large language model – knowledge graph approach in providing accurate responses on zeolite chemistry in natural language.

Received 20th June 2024
Accepted 6th September 2024

DOI: 10.1039/d4dd00166d

rsc.li/digitaldiscovery

1 Introduction

Zeolites are porous inorganic materials which have been of scientific interest since their first description by Fredrik Cronstedt in 1756.[1–3] Ancient applications of naturally occurring, mineralogical zeolites include water purification and use as construction materials.[3,4] Owing to their porosity, fine-tuning of chemical composition, size and topology of the internal channels and cavities, zeolites have been highly relevant in catalysis[5,6] and separation technologies.[7,8] Much of the interest in zeolites has been driven by their applications in domains such as crude oil cracking based on shape-selective Brønsted-acid catalysis,[9] separations of hydrocarbons[10] and the removal of water/CO_2 from natural gas.[11] Besides these energy-related domains, zeolites find further applications in ion exchange[12,13] and O_2/N_2 gas operations from air,[14] while new directions include the development of batteries and upcycling of carbon dioxide technologies.[15]

The porosity aspect of zeolites was inferred when, upon heating, certain mineralogical aluminium silicates released water vapour.[2,3] In addition to water molecules, zeolites are recorded to store a variety of other chemical species, including clusters and counter ions. Plenary zeolitic frameworks are typically described as having an ideal generic empirical formula $[TO_2]_n$, where the T-atom is a tetrahedrally coordinated framework-building element. Aluminosilicates are an example where the positions of the T-atoms are shared between T and T′ atoms, while the overall framework zeolite exhibits general formula $[T'_xT_{1-x}O_2]_n$. To completely balance the charge of the two oxo ligands per empirical formula unit, the T/T′ atoms are expected to be four valent (*e.g.* Si^{4+} or Ge^{4+}). However, when framework building centres with other oxidation states participate (*e.g.* Al^{3+} or $P^{4+}S$), the overall formal charge of the framework building element components may not be neutral, and thus it may need to be balanced by countercations which find a way in the structure through the network of channels and cavities. In this regard, most of the zeolite framework building elements are p-blocks (*e.g.*, Al, Si, Ga, Ge, P, Sn), s-block (*e.g.*, Li, Be), or d-block (*e.g.*, Ti, Fe). Oxygen atoms are the predominant complementary element for building zeolitic frameworks; however, other atoms such as N, S, or Se may take the position of oxygen in the construction of zeolitic materials.[16]

Zeolites precedent the development of other porous reticular materials, which obtain a broad prominence nowadays.[17] However, they still retain an enormous interest and fascination owing to their stability, market availability and industrial

[a] *CARES, Cambridge Centre for Advanced Research and Education in Singapore, 1 Create Way, CREATE Tower, #05-05, 138602, Singapore. E-mail: mk306@cam.ac.uk*

[b] *Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge, CB3 0AS, UK*

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

[d] *CMCL Innovations, Sheraton House, Castle Park, Cambridge CB3 0AX, UK*

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

[f] *The Alan Turing Institute, John Dodson House, 96 Euston Rd, London NW1 2DB, UK*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00166d

‡ These authors contributed equally to this work.

applications. Computational approaches have been expanding the frontier of research, especially in solving problems for which experimental design and validation can be challenging.[18,19] With the emergence and accessibility of applied AI, the field has been further advanced simply through data intelligence.[20–23] Similarly to medical and drug development research,[24,25] zeolite chemistry is highly interconnected to domains that may not be considered purely chemical in nature. Modelling of the interconnected nature is important to fully capitalise on machine intelligence and advance the field. In this regard, zeolite chemistry combines abstract aspects such as tiling of space and generic framework topologies,[26] with crystallographic information, and species/counterion information with its own chemistry in pores and framework directing effects.[27,28]

Over the past decade, our group has investigated the intersection of knowledge engineering (also known as knowledge AI) and chemistry.[29] Starting with the development of automated discovery and structure elucidation of organic species, along with retrosynthetic analysis by expert systems,[29] knowledge engineering has deepened our understanding of pure chemistry by helping chemists stipulate formal relationships between concepts,[30] examine cognitive decision-making,[31] and inspire new fundamental studies through playful interactions with these knowledge systems.[32,33] Knowledge engineering often relies on semantic web technology that enables efficient machine actionable retrieval and navigation of interconnected information, coupled with dynamic knowledge growth and decision-making facilitated by agent reasoning.[34,35] In terms of chemical and materials informatics, zeolite chemistry overarches chemical and crystalline material concepts, typically described in different data formats (see Fig. 1), making it a subject of fundamental and practical interest. Further on, zeolites are involved in forms of "host–guest" chemistry, and thus, their semantic representation is an effort towards developing more general models for simultaneous multi-component information representation in digital chemistry.[36]

In this study, we address the challenge of making zeolite chemistry machine-actionable and subsequently ensure that information can be retrieved in a structured and unstructured manner. This implies that information on zeolite material instances is integrated with information on zeolite topologies and their construction, crystalline information and information on non-framework chemical species functioning as guest or charge-balancing ions inside the framework cavities. These types of information are currently found through different research data resources (see Section 3.4 for more details†), and face interoperability challenges. To overcome these challenges, in this work, we apply knowledge engineering to develop two interconnected ontologies, namely "OntoZeolite" and "OntoCrystal", that deal with zeolitic and crystalline information, respectively. Concepts of these ontologies are semantically interconnected with "OntoSpecies" ontology,[37] that has been previously developed by us and used in the semantic representation of chemical species relevant in domains such as chemical kinetics,[38,39] reticular chemistry,[31] and experiment automation.[40,41] Following the integration of the new ontologies with the overall semantic world model of The World Avatar (TWA), we instantiate and interconnect curated zeolite, crystal and species data. On a basic level, TWA, as a progression of knowledge graph (triplestore), differs from traditional databases by storing data as triples of subject–predicate–object,

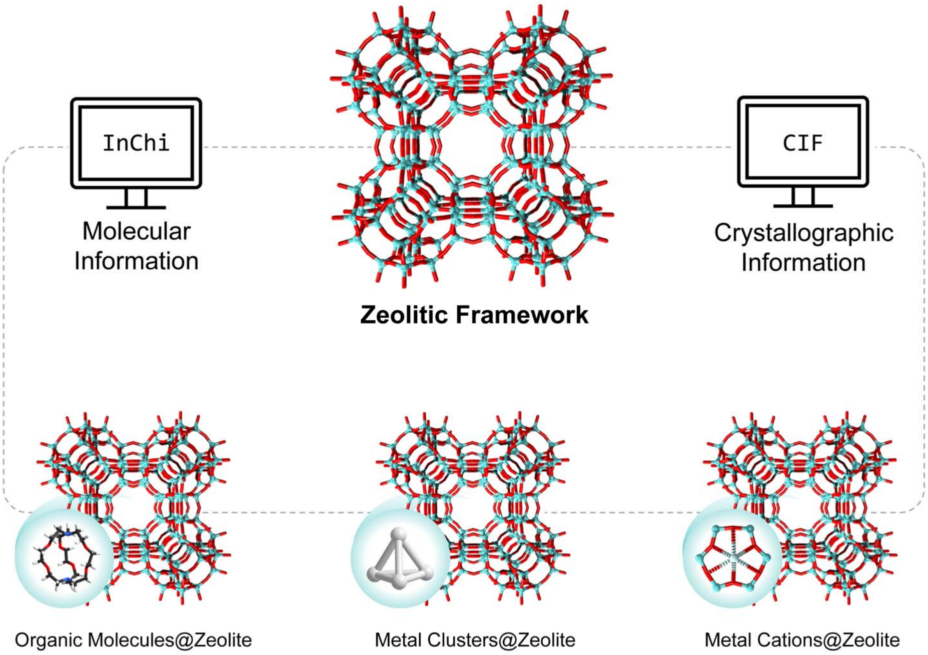


Fig. 1 In terms of information modelling, zeolite chemistry bridges information related to framework topologies, chemical species and crystalline materials.

facilitating semantic reasoning and schema flexibility, whereas traditional databases use fixed schemas and focusing on structured data retrieval without inherent relational inference.[29] Using tailored SPARQL queries, we showcase how interconnected information that is necessary for answering complex chemistry questions can be seamlessly retrieved. Using the TWA capability for question-answering (QA) through its "Marie" system, we open the possibility of zeolite information query using natural language. The application of large language models (LLMs) in chemistry has attracted attention for their potential utility, yet the persistent challenge remains in accurately assessing their performance.[42,43] Therefore, using Marie herein, we provide a blended approach combining the accuracy of knowledge graphs with the natural language understanding of LLMs with the intention to continue the development of QA systems that are explainable, track provenance and adapt to changes in their knowledge-base.[44]

2 Background

2.1 Zeolite architectures and their chemistry

Owling to their highly porous framework topologies, zeolites are significantly less dense than other silicate-based minerals (*e.g.* quartz). However, this aspect often increases their crystallographic complexity and the level of their configurational entropy.[45,46] Standardising the description of these frameworks has been one of the main focal points of the *International Zeolite Association*,[47] which has developed a variety of industry and research standards for zeolite chemistry, including codes of formally recognised zeolite materials, synthesis and characterisation references, among others. The association recognises over 250 topologically different zeolitic frameworks designated with three-letter codes. For instance, "Linde Type A"-LTA is one of the very commonly studied and described zeolite frameworks (see Fig. 2). Although one can build an LTA framework solely of

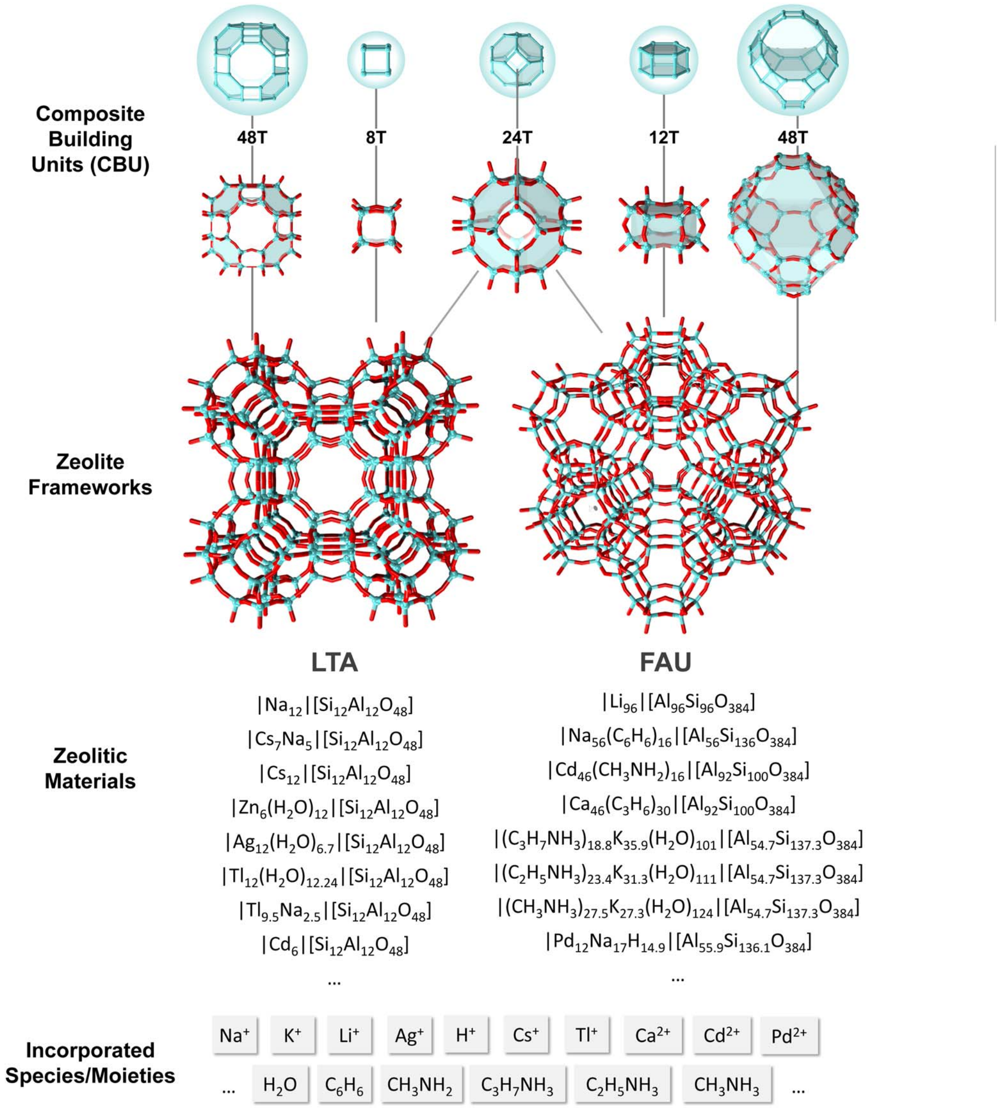

Composite Building Units (CBU)
48T
8T
24T
12T
48T
Zeolite Frameworks
LTA
FAU
Zeolitic Materials
$|Na_{12}|[Si_{12}Al_{12}O_{48}]$
$|Cs_7Na_5|[Si_{12}Al_{12}O_{48}]$
$|Cs_{12}|[Si_{12}Al_{12}O_{48}]$
$|Zn_6(H_2O)_{12}|[Si_{12}Al_{12}O_{48}]$
$|Ag_{12}(H_2O)_{6.7}|[Si_{12}Al_{12}O_{48}]$
$|Tl_{12}(H_2O)_{12.24}|[Si_{12}Al_{12}O_{48}]$
$|Tl_{9.5}Na_{2.5}|[Si_{12}Al_{12}O_{48}]$
$|Cd_6|[Si_{12}Al_{12}O_{48}]$
...
$|Li_{96}|[Al_{96}Si_{96}O_{384}]$
$|Na_{56}(C_6H_6)_{16}|[Al_{56}Si_{136}O_{384}]$
$|Cd_{46}(CH_3NH_2)_{16}|[Al_{92}Si_{100}O_{384}]$
$|Ca_{46}(C_3H_6)_{30}|[Al_{92}Si_{100}O_{384}]$
$|(C_3H_7NH_3)_{18.8}K_{35.9}(H_2O)_{101}|[Al_{54.7}Si_{137.3}O_{384}]$
$|(C_2H_5NH_3)_{23.4}K_{31.3}(H_2O)_{111}|[Al_{54.7}Si_{137.3}O_{384}]$
$|(CH_3NH_3)_{27.5}K_{27.3}(H_2O)_{124}|[Al_{54.7}Si_{137.3}O_{384}]$
$|Pd_{12}Na_{17}H_{14.9}|[Al_{55.9}Si_{136.1}O_{384}]$
...
Incorporated Species/Moieties
Na^+
K^+
Li^+
Ag^+
H^+
Cs^+
Tl^+
Ca^{2+}
Cd^{2+}
Pd^{2+}
...
H_2O
C_6H_6
CH_3NH_2
$C_3H_7NH_3$
$C_2H_5NH_3$
CH_3NH_3
...

Fig. 2 Illustration of key concepts defining zeolite chemistry (top to bottom): CBUs describe the topology of frameworks, while based on a common framework, different formulations/materials can be derived. These materials differ in terms of the reported species/moieties they incorporate.

Si as T atoms,[48] this material is more commonly built of Si and Al atoms in equal ratios. In the latter case, owing to the Al-presence, the overall framework becomes formally negatively charged, and thus it attracts countercations in its pores. In the case of sodium cation incorporation, one forms zeolite material formulations of the type $|Na_{12}(H_2O)_{27}|[Al_{12}Si_{12}O_{48}]$. The crystallographic unit cell of this zeolite is cubic ($a = 24.61$ Å) with $Fm\bar{3}c$ symmetry. The LTA framework has eight-member oxygen ring pores with a size of around 4.3 Å.

Another example of a zeolite framework is FAU (Fig. 2), whose three-letter code derives from the mineral faujasite. The naturally occurring faujasite exhibits a framework construction formula described as $[Al_7Si_{17}O_{48}]^{7-}$, which requires to be counterbalanced by cations. In the natural form, this can be based on Na^+, Ca^{2+} and Mg^{2+}, which collectively counter the charge, although their relative contributions can vary and may differ between samples. In synthetically formed FAU, the silica-to-alumina ratios may differ, while increased stability favours Si-rich frameworks. Furthermore, in synthetic FAU systems, the countercations can be similarly exchanged, leading to a plethora of different formulations. The unit cell of FAU zeolites is cubic with $a = 24.65$ Å and $Fm\bar{3}c$ space-group symmetry. When comparing both framework types, one can notice particular similarities. First, the T-atoms virtually describe polyhedral cages that share polyhedral corners, edges and faces with their respective neighbours. These types of virtual framework building fragments are often referred to as composite building units and, in principle, can be discrete (*e.g.* rings and polyhedra) but also continuous (*e.g.* chains).[49,50] When examining LTA and FAU frameworks, we notice that they both share structural arrangements, such as the sodalite cage made of 24 T atoms. This aspect is quite interesting as different fragments of the zeolite framework may be responsible for different functionalities. However, their description and existence provide a possibility for cross-structural comparisons. In addition to the composite building unit description, a more general description with mathematical tiling has evolved, which describes zeolitic topologies as three-dimensional structures made of polygonal faces that are commonly referred to as "Natural Building Units", which do not necessarily need to be flat.[51–53]

The zeolite crystal structures often display many species found in their cavities. These species may have entered the zeolite cavities through "post-synthetic" modifications such as ion exchange. Calcination is a process that normally removes internal species, but the charge balance is maintained through (partial) protonation. During the synthesis of zeolites, chemical species may play a role in directing the chemical outcome. However, their role may be conceptualised as a rigid templating effect, as it can be the case that a zeolitic framework can be synthesized in the presence of many different species.[54] Finally, complex zeolitic structures can also tightly incorporate complementary cluster materials that form simultaneously with the zeolite formation.[55]

2.2 Crystallographic information

The CIF (Crystallographic Information File) is a structured text file format designed and maintained by the International Union of Crystallography (IUCr) for the storage of crystal structure data as well as information relating to the actual crystallographic measurement.[56,57] The CIF contains different data blocks with array-like structuring covering information on atomic coordinates, lattice parameters, Miller indices, coordinate transformation matrices, Cromer–Mann scattering-factor coefficients *etc.*[58,59] The core CIF dictionary is rich in terms of data names that enable convenient archiving and exchanging of raw and processed crystallographic data. This dictionary covers several thousand data properties; however, only 30 are sufficient to represent the crystallographic information involved in the virtual building of zeolitic models (see Fig. S2 in ESI†). Many of the concepts (*i.e.* tags) covered by the core CIF and its related dictionaries relate to publication information, sample preparation, experimental conditions and techniques used, and audit and revision history, which are not involved in crystallographic model building but provide process information for reproducibility and data integrity purposes. These concepts are useful for practical guidance on other integrated knowledge graphs relevant to experimental material design and laboratory automation.[60–63]

Attempts to represent chemical crystallographic information with the help of the semantic web technologies have been reported;[64] however, the respective ontologies have not reached a maturity level to provide detailed representation for the complex query of crystals at the atomic level. The reason for this may be that to make meaningful queries, the data of the CIF has to undergo vector and matrix transformations, taking into consideration the overall crystallographic symmetry. In this work, we develop a new crystal structure describing ontology OntoCrystal, which includes classes that facilitate operations suitable for semantic storage of data as well as visualisation.

2.3 Digital chemistry in The World Avatar

The World Avatar (TWA) is an open, dynamic world model built upon the semantic web stack (Fig. 3). It encapsulates a comprehensive representation of diverse domains, including power and heat network optimisation, environmental monitoring, and climate resilience, as demonstrated through the Climate Resilience Demonstrator (CReDo) project.[65,66] Central to TWA's functionality is its focus on chemicals and processes, underpinned by interlinked ontologies such as OntoSpecies, OntoKin, OntoCompChem, and OntoPESScan.[38,67,68] These ontologies provide a semantic framework for representing chemical species, reaction mechanisms, quantum chemistry calculations, and potential energy surface scans. Through its carefully designed interconnectivity, TWA promotes data interoperability and reduces ambiguity across previously isolated data silos.[29,69–71]

Semantic agents play a vital role within TWA, managing information flow and executing complex tasks. These agents perform essential functions, such as the calibration of kinetic mechanisms[40] and the automated design of metal–organic polyhedra (MOPs) based on inductive reasoning algorithms.[31,72] To facilitate user interaction, TWA employs a question-answering system named "Marie", which leverages advanced

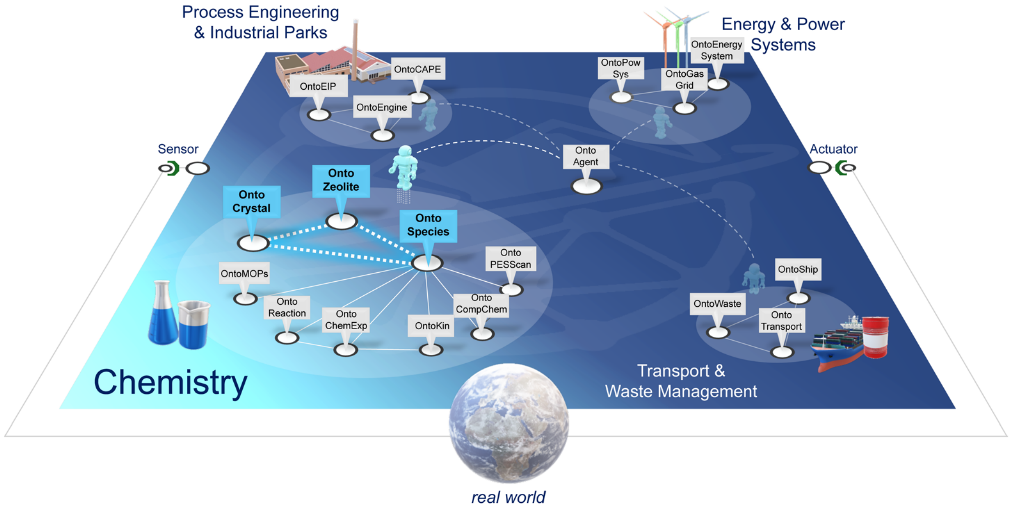


Fig. 3 A selection of ontologies and their connectivity that have been integrated in TWA. OntoCrystal, OntoZeolite, and OntoSpecies are part of the digital chemistry domain.

natural language processing to provide real-time responses.[73–75] The output agents that form the Marie functionality map natural language question to machine-readable SPARQL commands that retrieve the relevant information from TWA.[29]

3 Methodology

In the ontological context, the TBox (Terminological Box) organises and hierarchically categorises concepts while defining inter-domain associations through object properties. This can be represented through the help of description logic (see Section SI.5 in the ESI†). In contrast, the ABox (Assertional Box) leverages the TBox structure to instantiate these concepts with specific entities and their interrelations, as well as relevant data. Together, they enable precise data querying, individual entry access, and consistency checks.[76] The Hermit reasoning tool[77] checks the consistency of the TBox and ensures that the data types in the ABox align with the definitions provided in the TBox.

Prior to the creation of an ontology, we developed competency questions (see Section SI.3 in the ESI†) to determine the scope of the ontology and ensure the ontological model captures complex domain interconnections. This section summarises the development of three critical ontologies: OntoZeolite, OntoCrystal, and OntoSpecies, each crucial for integrating domain-specific knowledge coherently.

The current approach aligns with trends in chemistry and materials information science, aiming to make knowledge machine-actionable and openly accessible.[78–81] Unfortunately, many existing datasets for zeolite chemistry remain siloed and are primarily accessible only to experts. By employing the TWA method, which combines natural language processing and semantic graph instantiation, we ensure that these datasets become interconnected with general chemistry knowledge, a development that is generally well-received by chemists beyond the zeolite community.

3.1 OntoZeolite

The OntoZeolite ontology provides a structured framework that contextualises zeolites-related knowledge. This ontology introduces 26 classes, 28 object properties and 30 data properties. One of the central classes in this ontology is the zeolite framework. This class is used to instantiate information about individual framework types (*e.g.* FAU, LTA, NAT *etc*). A zeolite framework may be described separately or in combination through a set of topological properties. Thus, the topological properties class further connects to classes such as occupiable volume, accessible area, framework density, ring sizes and other provide different information about the properties that define the frameworks, but also can provide qualitative information on what forms of guest species can access the porous areas of the zeolite.

The class zeolite framework also connects to the class zeolitic material. The latter class is introduced to represent different zeolite instances that have been synthesised or discovered in nature. On practical grounds, for every zeolite material, we further represent the elements and their count involved in the description of the framework structure. In the ontology, this is being implemented through the class framework component, which allows querying of materials based on elemental composition and relative compositions. Considering that within the zeolitic material, there can be different chemical species, they are represented as such through the class species in the OntoSpecies ontology. As zeolitic material and zeolite framework are crystalline in nature, they further connect to the class crystal information defined by the OntoCrystal ontology. All zeolitic frameworks and materials are linked to the document class. This class connects them to relevant bibliographic details using the BIBO ontology.[82] Considering the growing interest in the digital exploration of the synthesis of new zeolite materials,[83] our ontology also introduces a link between the zeolitic material and recipe classes, followed by connections to precursor chemicals and chemical species for future studies.

The OntoZeolite ontology depicts the relationships between zeolite materials and their frameworks, defined by unique tiling elements and symmetry. While frameworks may share tiling elements, differences in connectivity result in distinct topologies and porosities. The knowledge graph captures these nuances, showing how materials with similar compositions can have varying structural properties. It includes crystallographic data to differentiate materials based on recognised zeolitic topologies. Although semantic agents, in principle, can be developed to classify new materials, the formal recognition of zeolitic frameworks is managed by the International Zeolite Association.[47]

3.2 OntoCrystal

The OntoCrystal ontology provides a semantic representation of crystallographic data (see Fig. 4). This ontology encompasses 18 classes, 43 object properties and 25 data properties. Physical properties with unit reuse concepts defined by the Ontology of Units of Measure (OM) version 2.0.[84,85] Crystal Information Files (CIF) allow and often use measured values with uncertainties, which are not currently supported by OM. For such data, we introduced a new concept MeasureWithUncertainty in OntoCrystal (see ESI Section SI.1,† Fig. 1 for further details).

The central class in the OntoCrystal ontology, CrystalInformation, is used to store fundamental crystallographic information and aggregates data from five key classes: unit cell, XRD spectrum, atomic structure, coordinate transformation, and tiled Structure. The unit cell class provides metrics on unit cell dimensions, including lengths, vectors, angles, and volume. Atomic structure details the arrangement of atoms within the crystal lattice. The atom site information consists of the atom type, the absolute and relative positions, and the site occupancy. The coordinate transformation class incorporates transformation vectors and matrices to convert relative within

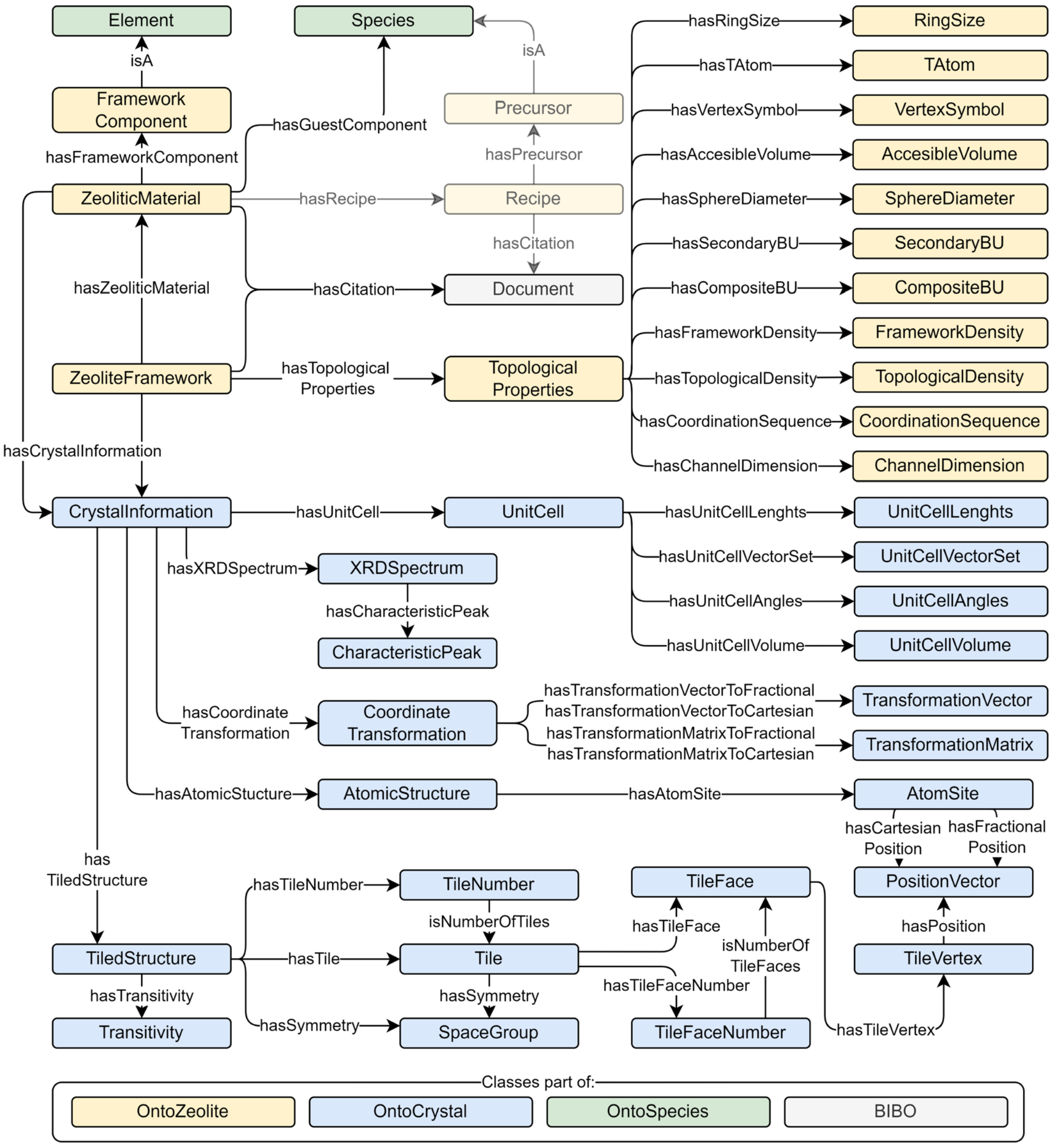


Fig. 4 Overview of the main classes, properties and interconnectivity between OntoZeolite, OntoCrystal, OntoSpecies and BIBO ontologies.

the unit cell to real Cartesian coordinates, and *vice versa*. The XRD spectrum class models the X-ray Powder Diffraction spectrum, quantifying X-ray diffraction intensity across diffraction angles and is represented in a "2θ plot", which can be derived from experimental or simulated data. Apart from the full plot data represented as plot *XY* this class stores the same information as a list of peaks. The characteristic peak class is tailored for fingerprint analysis, facilitating the assessment of peak characteristics, including position, intensity, and width, critical for comparative crystallography. In most cases, the processed data in terms of characteristic peak saves storage, and the full plot data is omitted in this case.

Natural tiling of space is a practical way of describing zeolite frameworks; however, its relevance is far more generally applicable to crystalline materials. Natural tiling involves the concept of tile, which is also considered by the CIF standards and described in a separate topology dictionary.[57] Thus, as part of OntoCrystal, we included tiled structure that defines the tiling patterns and includes the transitivity class, which reflects on the uniformity and the description of the allowed transformations through symmetry operations. Tiled structure further connects to the classes tile, tile number and space group that define the geometric properties of tile faces, the count of tiles and the space groups associated with each tile configuration.

3.3 OntoSpecies

OntoSpecies, an integral ontology within the TWA framework, catalogues distinct chemical species and their properties, each assigned a unique Internationalized Resource Identifier (IRI) to ensure unambiguous identification.[37] This ontology works in tandem with OntoZeolite to facilitate the precise identification of chemical species in zeolite structures through OntoSpecies IRIs, thereby enabling detailed exploration of their interconnected properties. OntoSpecies is crucial in linking species to instances and concepts from other ontologies within the TWA chemistry domain. It also incorporates common cheminformatic identifiers such as InChI, InChIKey, CAS registry numbers, PubChem CID, and SMILES, which are used for retrieving external information. The molecular geometry is meticulously documented within the ontology, making the data useable for quantum chemical calculations, with each bond and atom distinctly identified by an IRI. The OntoSpecies ontology encompasses a broad spectrum of chemical and physical properties, classifications, applications, and spectral data for each species. It includes detailed provenance and attribution metadata to ensure the reliability and traceability of the data. Most of the chemical species information is sourced from a variety of open chemical databases, rendering OntoSpecies as a unifying ontology for chemical informatics.

The OntoSpecies ontology is semantically interoperable with the OntoCompChem ontology,[67] facilitating the semantic description of computational chemistry data for species and materials. Future efforts could enable the instantiation of existing calculated (quantum)-mechanical information on zeolites[86] and their instances, as well as the use of semantic agents to perform new on-demand calculations based on user requests.

3.4 Data curation and instantiation

Information on zeolites has been guided by the IZA structural dataset in conjunction with zeolite framework and material descriptions published as original research.[50,87] From the original literature, we have acquired information on mineralogical and synthetically reported zeolites, which includes their chemical formula, crystallographic information and relation/incorporation of chemical species/counterions in their porous structure. Additional information on zeolite materials, their chemical formulae, their relation to crystallographic systems, and their bibliographic information were sourced from previously published and peer-reviewed datasets.[88,89] Considering the integration of zeolitic material instances from the last two sources, our current implementation features more material instances displayed in the traditional IZA structure dataset, which is not surprising as the IZA resource is mainly focused on the detained description of zeolitic frameworks.[47] Manual cross-checking of papers was required to confirm the presence of chemical species/ions, and further collection on the properties of these chemical species and ions was performed through programmatic queries from the PubChem database. In a few instances, PubChem info was absent (*e.g.* for cluster and organometallic structures), and thus such instances were added manually.

The original data were derived from various file formats, including CSV, CIF, JSON, BIB, and TXT, among others. Following this, as outlined in our workflow (see Fig. 5a), we augmented, corrected, and supplemented missing data as necessary. For XRD spectra, we extracted the 2θ positions and their relative intensities, preparing them for instantiation. Information on zeolite formulae has been cross-checked with the original literature, which typically derives it with consideration of multiple characterisation techniques. Owing to different limitations in real experiments, formula content ascribed to the linked crystallographic information may sometimes differ due to various factors (*e.g.* disorder of the chemical guest species, no detection of light elements such as hydrogen, *etc.*). For authenticity reasons, such crystallographic data is not

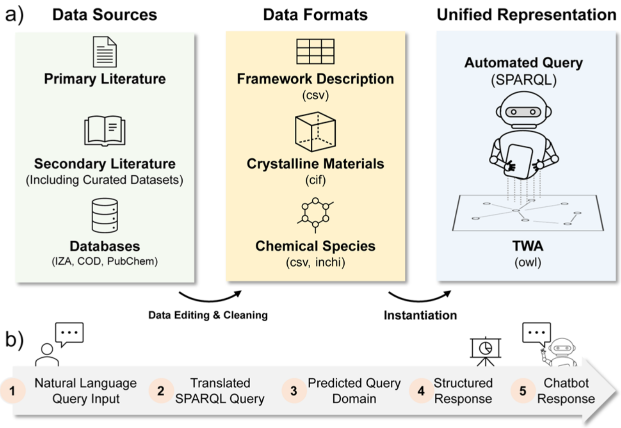


Fig. 5 Overview of (a) the data curating and processing workflow; (b) processing of natural language queries on TWA–Marie interface.

further altered but directly linked to the material instance. All data formats were augmented to produce an OWL ABox, which was subsequently uploaded to our knowledge graph. During the augmentation process, data linking is performed using the ontological designs described above. Comprehensive details on the data curation process are available in the ESI (see Section SI.1 in the ESI for more details).†

3.5 TWA integrated query interface

To facilitate the exploration of zeolite chemistry, a user interface was developed, enabling efficient interaction with data on zeolite properties (see Fig. 5b). This interface provides both field-based and natural language search options, and it is currently available through the TWA–Marie webpage, equipped with plentiful examples across various chemistry domains (see **https://theworldavatar.io/demos/marie/** for more detail).

The structured or field-based search feature related to zeolitic frameworks enables cross-structural comparison by plotting numerical data of over twenty different properties. This built-in functionality comes with the calculation of correlation coefficient and colour mapping based on a third property. Additionally, frameworks and material instances can be queried using pre-defined search fields. In the case of zeolite frameworks, users can query framework information based on X-ray diffraction (XRD) peak positions and their relative intensities, unit cell parameters, different forms of densities and building unit features describing the framework topology. Meanwhile, zeolitic materials can be retrieved based on their formula, elements that form the framework, and non-framework species/ions. As crystallographic information and academic literature are associated with the zeolitic material instances, they can also be queried using unit cell parameters and DOI numbers.

Unstructured or natural language search allows users to submit a query in natural language without locating specific input fields; users then obtain responses in both tabular and human-friendly textual formats. This is achieved by applying our previously developed method that supports our question-answering system for combustion kinetics.[90] Specifically, we performed multi-task fine-tuning on the pre-trained language model Flan-T5 for natural language-to-SPARQL translation and domain classification tasks. At test time, the model runs two inference tasks: translating natural language input into a corresponding SPARQL query and predicting TWA domain for SPARQL execution to retrieve desired information (see Section SI.2 in the ESI† for a detailed process breakdown).

4 Results and discussion

In this section, we provide an overview of the zeolite and crystallographic information within the context of TWA. We demonstrate the semantic structuring and interconnection of zeolitic, crystallographic, and species data through SPARQL queries. These queries, developed with the ontological structure in mind, enable programmatic searches. However, crafting queries may not always be straightforward. Therefore, template queries can be developed and deployed for advanced searches, either through a web interface or within a question-answering system.

4.1 Overview on TWA zeolitic instances

After instantiation of zeolite framework, material, species and crystallographic information, TWA provides coverage of 251 zeolites, over two thousand zeolite materials where the majority are supported by crystallographic information. In Fig. 6, we attempt to analyse the available data on how the different zeolitic instances are distributed across framework types and what sort of species/ions they incorporate.

The top 10 zeolitic frameworks—namely FAU, LTA, NAT, CHA, HEU, RHO, GIS, SOD, ANA, and LAU—encompass a total of 1177 instances, as demonstrated in Fig. 6a. This high instance density per framework indicates that a relatively small number of zeolitic frameworks are the focus of a significant portion of scientific reports, inquiries and analyses within the field. The FAU framework, in particular, registers the highest occurrence with 374 instances, followed by the LTA framework with 277 instances and NAT and CHA frameworks with 99 and 92 instances, respectively. Multiple reasons prompt the aggregation of these instances among the top frameworks. First, zeolitic frameworks such as FAU and CHA remain highly relevant to the industry, and thus, the number of reported material instances reflects their importance to the scientific community. On the other hand, HEU, GIS, SOD and LAU often are highly stable and competing framework materials that frequently appear in zeolite synthesis. GIS and ANA synthetically are also commonly reported in mineralogical studies, making them one of the more frequently reported zeolites. The frequency of reporting in scientific literature does not necessarily reflect the industrial relevance of a particular zeolitic framework. For instance, the MFI framework, despite being the subject of numerous industry patents,[91] illustrates this point well. Patents often cover a broad spectrum of compositional formulae to secure extensive protective rights, which complicates efforts to accurately determine the number of distinct MFI material instances developed and utilised outside academic research.

Fig. 6b presents a scatter plot that examines the correlation between the number of reported material instances of zeolite frameworks and the diversity of incorporated ions and species. While the data points predominantly cluster near the origin, indicating a prevalent trend of limited incorporation diversity across most frameworks, a few notable exceptions emerge. The frameworks of FAU, LTA, CHA, and NAT distinguish themselves not only through a higher count of material instances—374, 277, 92, and 99, respectively—but also through their considerable diversity of ions and species, with FAU, LTA, CHA, and NAT having 65, 54, 36, and 9 unique guest components, respectively. Together with HEU, RHO, and GIS frameworks, these seven types account for over 1000 material instances, demonstrating their importance and potential for structural and chemical adaptability. Limitations in terms of the diversity of incorporated species are obvious in the case of the NAT framework. This frequently studied has been found to form mainly in the presence of sodium cations, which explains the low variety of

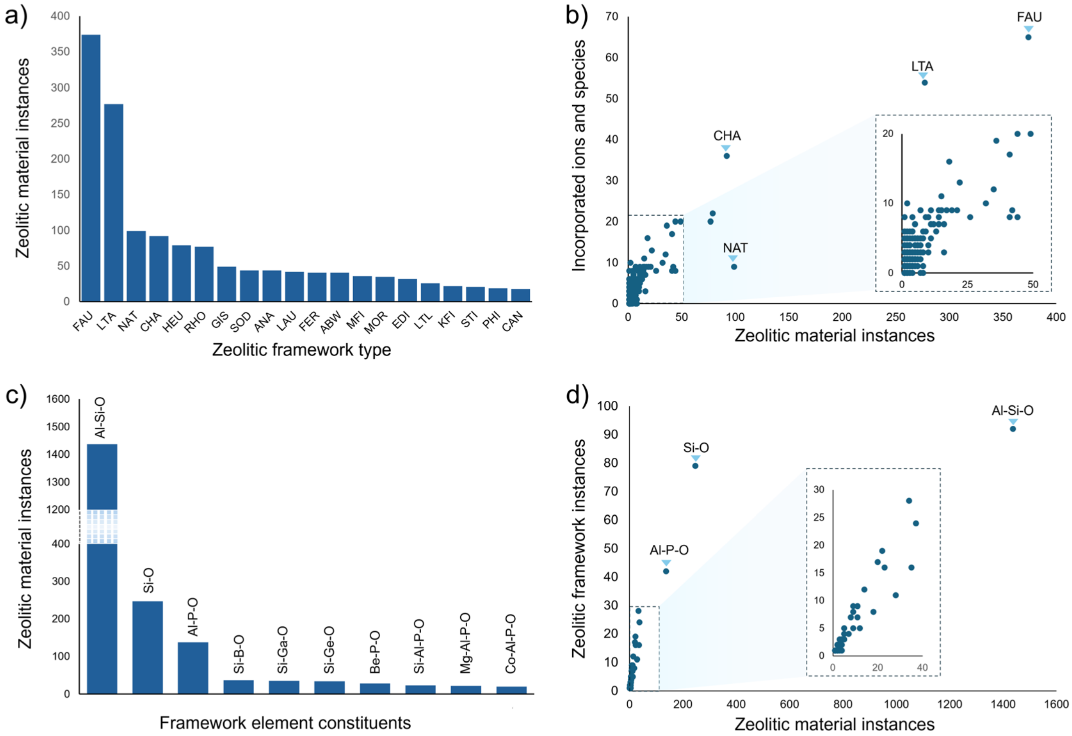


Fig. 6 Overview of the zeolitic instances from the TWA: (a) bar chart displaying zeolitic framework types alongside their quantity of material instances. (b) A scatter plot shows the number of material instances for each framework type *versus* the diversity of incorporated species. (c) Bar chart detailing the frequency of the 10 most prevalent framework-building elements and their various combinations. (d) Scatter plot presenting the number of zeolitic material instances in relation to the range of elemental combinations within framework types.

incorporated species. From the collected information, an overwhelming majority of zeolite framework types—approximately 92.3% have been associated with less than 25 material instances and fewer than ten different ions or species. This stark contrast indicates that a small minority of zeolite frameworks are associated with most of the reported zeolite materials and incorporated species.

Within our knowledge graph, there are 73 distinct sets of framework-building elements. A significant majority of these, comprising 1437 zeolitic material instances, consist of aluminium and silicon, as illustrated in Fig. 6c. This prevalence aligns with the common definition of zeolites as hydrated aluminosilicates often containing sodium, potassium, calcium, and other cations. Correspondingly, aluminosilicates dominate within the largest set of framework topologies (92 topologies), as depicted in Fig. 6d. Following aluminosilicate zeolites, purely silicate-based frameworks are the second most represented, with 247 instances across 79 topologies. Aluminophosphates also feature prominently, with 137 material instances spread over 42 topologies. Beyond these three prevalent material types, our knowledge graph encompasses a variety of structures composed of different elemental combinations.

4.2 Custom SPARQL-based requests

SPARQL is an RDF query language designed to retrieve semantically structured data. It is paired with Blazegraph,[92] an open-source triplestore that serves as the main graph database infrastructure of TWA. Understanding the ontological structure enables the crafting and execution of customised SPARQL queries over Blazegraph for programmatic data retrieval. Examples of such queries, specifically for accessing crystallographic information *via* our web-hosted TWA – Blazegraph, are documented in the ESI† of this work (Section SI.4†). Fig. 7

```
PREFIX zeo: <http://www.theworldavatar.com/kg/ontozeolite/>
PREFIX os:
↪ <http://www.theworldavatar.com/ontology/ontospecies/OntoSpecies.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?ZeoliteName (SAMPLE(?IUPAC) AS ?GuestIUPACName) ?GuestFormula
↪ ?MolecularWeight (SAMPLE(?BPValue) AS ?BoilingPoint_value)
↪ (SAMPLE(?BPUnit) AS ?BoilingPoint_unit)
WHERE {
  ?Framework zeo:hasFrameworkCode "LAU" ;
             zeo:hasZeoliticMaterial ?Material .
  ?Material  rdf:type zeo:ZeoliticMaterial ;
             os:name ?ZeoliteName ;
             zeo:hasGuestCompound ?Guest .
  SERVICE
  ↪ <https://theworldavatar.io/chemistry/blazegraph/namespace/ontospecies> {
    ?Guest rdfs:label ?GuestFormula ;
           os:hasIUPACName/os:value ?IUPAC ;
           os:hasMolecularWeight/os:value ?MolecularWeight .
    OPTIONAL {
      ?Guest  os:hasBoilingPoint ?BPNode .
      ?BPNode os:value ?BPValue ;
              os:unit/rdfs:label ?BPUnit ;
              os:hasReferenceState/os:value ?value .
    }
  }
} GROUP BY ?ZeoliteName ?GuestFormula ?MolecularWeight ORDER BY ?ZeoliteName
```

Fig. 7 Example of a SPARQL query that retrieves information cross zeolitic framework, zeolitic material to molecule species. Example output is Fig. S6 in the ESI.†

illustrates a query retrieving chemical information about species within zeolites. To extract the required data, the system navigates the graph, starting from the specified zeolite framework to associated material instances and then to the interconnected species IRI, before retrieving details about those chemical species (*e.g.*, molecular weight). Typically, chemists with domain expertise extract such information manually through cognitive processes. This case demonstrates how TWA can perform cognitive-like tasks, by navigating its knowledge graph.[29]

4.3 Web-assisted data exploration

TWA is the first instance of semantically-assisted machine-to-machine communication,[93] however, to enable humans to interact with TWA, tools to overcome the human–machine barrier are needed. In this section, we showcase examples where SPARQL queries are automatically drafted based on user input. We first show a query of properties across zeolitic frameworks, which are then illustrated using the built-in plotting and correlation tools property correlations (Subsection 4.3.1). Next, we show a query of frameworks and material instances based on pre-drafted inputs 4.3.2 and this type of search is also extended to finding structural models based on powder XRD peak positioning (subsection 4.3.3). Finally, we cover examples of querying TWA with the help of natural language processing (Subsection 4.3.4).

4.3.1 Queries for cross-framework comparisons. The *Zeolite Explorer* tool facilitates the identification of overarching trends in zeolite frameworks by allowing users to input specific framework parameters. Upon specification, the system populates a predefined SPARQL query, which retrieves the corresponding data. This data is then visually represented in a color-coded two-dimensional plot. SPARQL queries allow flexible addition of filters that can narrow down subsets of data for closer inspection; however, for demonstration purposes of all the zeolitic instances in TWA, that option is committed in the follow-up discussion. Thus, this section discusses the range of parameters users can explore, highlighting that not all parameters exhibit significant correlations. Conversely, properties such as the number of tile edges and ring member sizes generally demonstrate strong correlations due to their shared structural roles in defining framework characteristics.

An interesting aspect of zeolite materials is the distinction between accessible and occupiable areas. Although zeolites can have large cavity cages, their accessibility is often limited by the small size of the channels leading to them. Channels defined by six-membered rings are largely considered inaccessible for diffusion. When we plot all zeolitic frameworks in the TWA, regardless of the ring size of the involved channels, we observe that a large subset of them correlates linearly with the accessible area; however, due to the subset a structure having narrow channel sizes, the overall correlation coefficient of drops to 0.83, as shown in Fig. 8a and b. These plots employ different colour schemes to highlight the largest and smallest ring sizes, respectively, but both illustrate the same underlying data relationship. This pattern is exemplified by the sodalite framework (SOD), which does not show any accessible area in Fig. 8a, despite its large cavity sizes. In contrast, zeolites with highly accessible areas over 2500 m^3 g^{-1} often have ring sizes exceeding 10 members but also include some of the smallest rings, as seen in Fig. 8b. This variability can be attributed to the

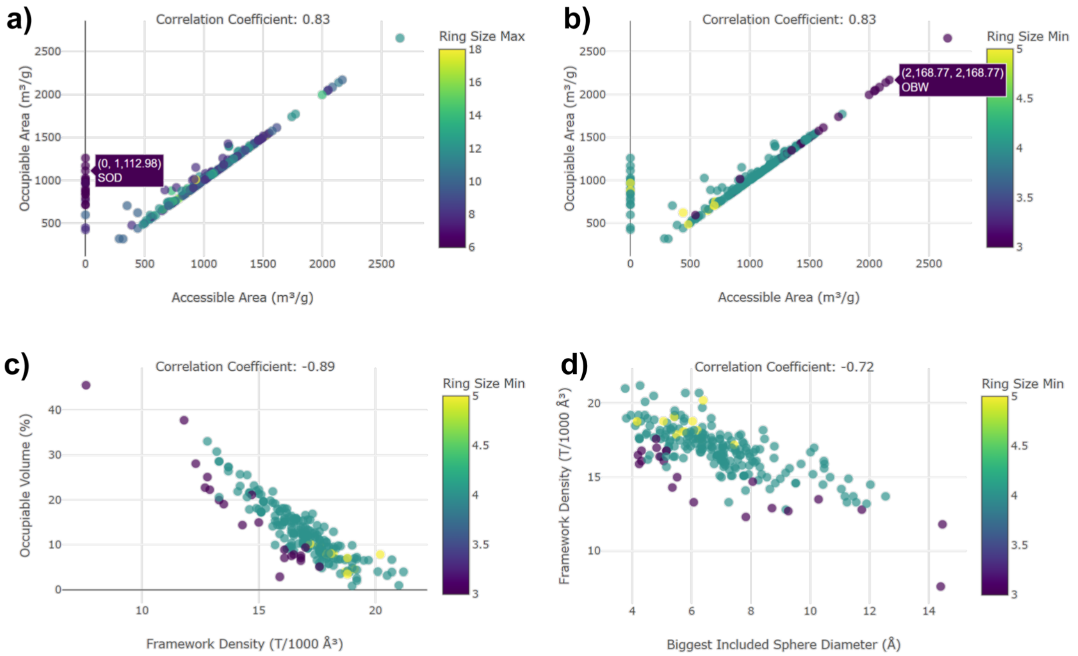


Fig. 8 Correlations and properties of zeolite frameworks. (a) Accessible area *vs.* occupiable area, colouring: ring sizes max; (b) accessible area *vs.* occupiable area, colouring: ring sizes min; (c) framework density *vs.* occupiable volume, colouring ring size min; (d) biggest included sphere diameter *vs.* framework density, colouring: ring size min.

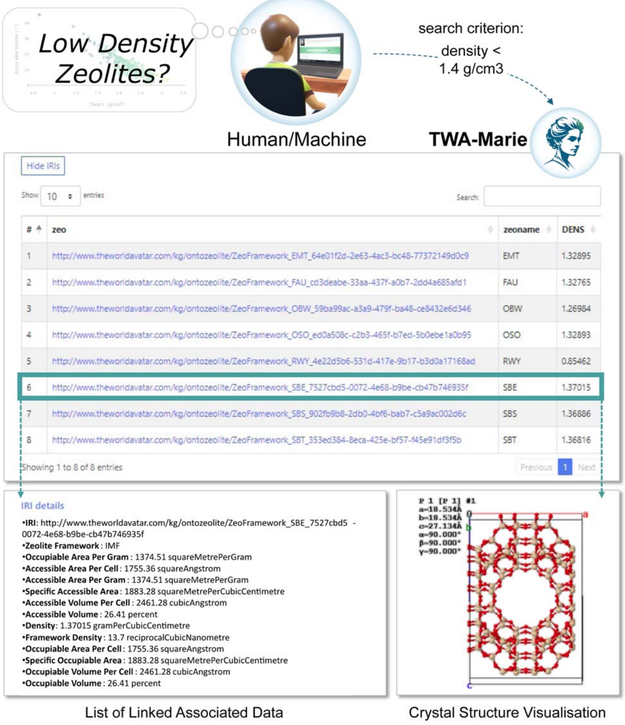
cage structures resembling truncated polyhedra, where truncation forms openings of various sizes, enhancing internal accessibility.

In Fig. 8c, a clear trend is observed where the occupiable volume decreases as the framework density increases (correlation coefficient = −0.89), indicating a strong inverse relationship. Framework density, defined as the number of tetrahedral atoms per 1000 cubic angstroms (Å^3), is inversely correlated with pore size. This relationship illustrates that denser structures, characterised by smaller pores, offer less available cavity space. Further analysis highlights a correlation between ring size and framework density: structures with ring sizes of 3 are generally less dense, featuring more expansive cavities, whereas zeolites with a minimum ring size of 5 are among the most densely packed, leading to significantly lower occupiable densities (correlation coefficient = −0.72). This pattern is also supported by the largest included sphere diameter, indicating that the least framework density is typically found in zeolites with a minimum ring size of 3, as depicted in Fig. 8d.

The knowledge graph structure allows for the dynamic addition of new zeolitic frameworks, ensuring that the information remains current. This structure supports flexible data exploration and updates, in contrast to static tables, which cannot be easily modified with new data. Fig. 8, illustrating data retrieval *via* SPARQL from the TWA, highlights the advantages of this system, enabling users to effectively explore and compare properties of different zeolitic frameworks. This approach offers a significant improvement over traditional manuscript-based information retrieval[94] by facilitating better interaction with and updates to the data.

4.3.2 Structured query of framework properties. The graph-based structuring of knowledge in zeolitic materials promotes efficient navigation through interconnected information, utilizing common edges and nodes to link related entities. Our ontology structures, illustrated in Fig. 4, facilitate straightforward transitions from species-level data to zeolite frameworks and subsequently to their crystalline properties. This architecture supports queries such as: *"find property X of zeolite(s) Z belonging to a given framework F"*. Further, as shown in Fig. 9, users can query: *"Find all Frameworks F that have properties X satisfying given conditions"*. These queries return detailed information about zeolite frameworks, including crystalline structure and porosity. Access to these properties is provided through the *Advanced Search* function of the web interface, which offers input fields for specifying values, value ranges, or strings related to various material properties. Upon input submission, the backend processes the data into one or several SPARQL queries, each producing a list of results. The final output, an intersection of these lists, ensures comprehensive and accurate retrieval of data.

4.3.3 Structured query of reference XRD powder patterns. Reference powder X-ray diffraction (XRD) spectra comprise published spectral data of materials that have undergone rigorous verification processes. In the field of zeolite chemistry, reference spectra are available for the majority of framework compounds.[95] These spectra are instrumental in spectral blueprinting analysis, a methodology employed by researchers to ascertain whether a newly synthesized zeolite material matches an existing framework. This blueprinting process involves a detailed comparison of characteristic diffraction peaks. Blueprinting of XRD spectra has the opportunity to be automated, thus enhancing the accuracy and efficiency of the material verification process.



#	zeo	zeoname	DENS
1	http://www.theworldavatar.com/kg/ontozeolite/ZeoFramework_EMT_64e01f2d-2e63-4ac3-bc48-77372149d0c9	EMT	1.32895
2	http://www.theworldavatar.com/kg/ontozeolite/ZeoFramework_FAU_cd3deabe-33aa-437f-a0b7-2dd4a685afd1	FAU	1.32765
3	http://www.theworldavatar.com/kg/ontozeolite/ZeoFramework_OBW_59ba99ac-a3a9-479f-ba48-ce8432e6d346	OBW	1.26984
4	http://www.theworldavatar.com/kg/ontozeolite/ZeoFramework_OSO_ed0a508c-c2b3-465f-b7ed-5b0ebe1a0b95	OSO	1.32893
5	http://www.theworldavatar.com/kg/ontozeolite/ZeoFramework_RWY_4e22d5b6-531d-417e-9b17-b3d0a17168ad	RWY	0.85462
6	http://www.theworldavatar.com/kg/ontozeolite/ZeoFramework_SBE_7527cbd5-0072-4e68-b9be-cb47b746935f	SBE	1.37015
7	http://www.theworldavatar.com/kg/ontozeolite/ZeoFramework_SBS_902fb9b8-2db0-4bf6-bab7-c5a9ac002d6c	SBS	1.36886
8	http://www.theworldavatar.com/kg/ontozeolite/ZeoFramework_SBT_353ed384-8eca-425e-bf57-f45e91df3f5b	SBT	1.36816

Showing 1 to 8 of 8 entries

IRI details

- **IRI**: http://www.theworldavatar.com/kg/ontozeolite/ZeoFramework_SBE_7527cbd5 - 0072-4e68-b9be-cb47b746935f
- **Zeolite Framework** : IMF
- **Occupiable Area Per Gram** : 1374.51 squareMetrePerGram
- **Accessible Area Per Cell** : 1755.36 squareAngstrom
- **Accessible Area Per Gram** : 1374.51 squareMetrePerGram
- **Specific Accessible Area** : 1883.28 squareMetrePerCubicCentimetre
- **Accessible Volume Per Cell** : 2461.28 cubicAngstrom
- **Accessible Volume** : 26.41 percent
- **Density**: 1.37015 gramPerCubicCentimetre
- **Framework Density** : 13.7 reciprocalCubicNanometre
- **Occupiable Area Per Cell** : 1755.36 squareAngstrom
- **Specific Occupiable Area** : 1883.28 squareMetrePerCubicCentimetre
- **Occupiable Volume Per Cell** : 2461.28 cubicAngstrom
- **Occupiable Volume** : 26.41 percent

List of Linked Associated Data

Crystal Structure Visualisation

Fig. 9 An example of web-assisted SPARQL query where a user specifies a parameter *e.g.* low-density zeolitic frameworks and only through the query of TWA complete information obtained as well as a structural projection of the zeolitic material.

In our knowledge model, the XRD powder data is linked to zeolitic frameworks. However, signal positions and relative intensities are crucial for the fingerprint identification of structures, and thus, we effectively use them to query and predict XRD plots based on user input. The whole operation involves SPARQL queries, which retrieve this data, compare it with the user's input, and suggest a framework type that has been identified through the fingerprinting method. The templated SPARQL queries are adjusted to essentially respond to the question "find frameworks F that have peaks of relative height at least P_I near a given position $P_{2\theta}$". In our SPARQL template, we have provided the opportunity for up to three characteristic peaks given a position and intensity. The default width and the cut-off intensity used in the templated queries on the backend are 0.5° 2θ and 50%, respectively. Examples of this query can be when a user inputs three 2θ positions: 18, 27, and 29. The query system might suggest that the closest match is with the LOV framework, where the positions are 17.82, 27.04, and 28.92 (Fig. 10). In a hypothetical scenario, when different or less characteristic 2θ positions are provided as input, TWA will provide a list of zeolitic frameworks that meet the user's input criteria.

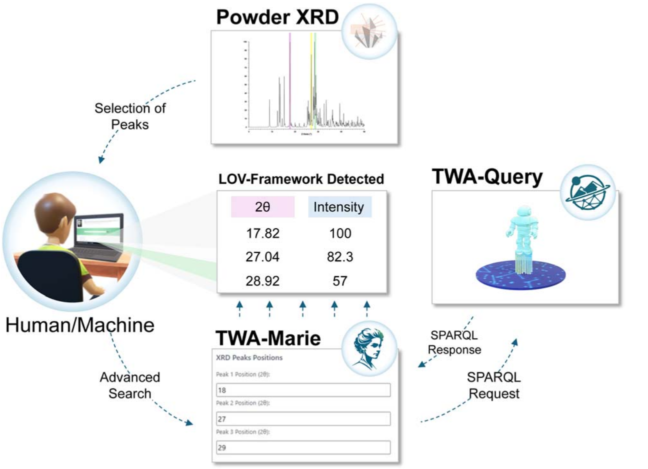
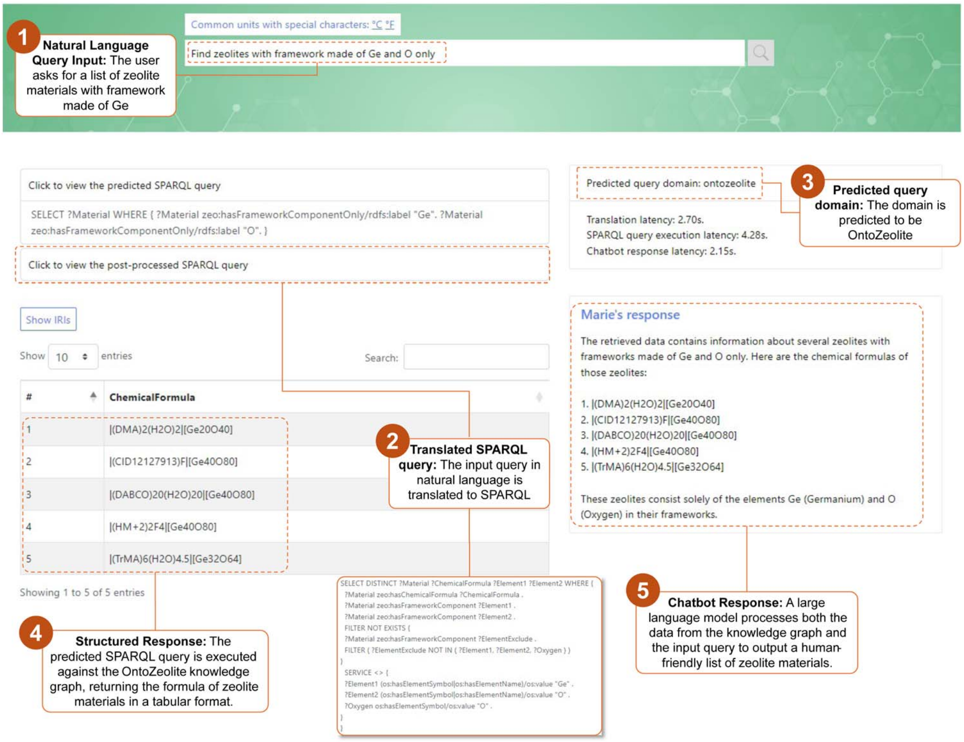


Fig. 10 Schematic illustration of search of matching framework types based on XRD diffraction characteristic peaks.

In contrast to recent studies employing machine learning for the comparison of XRD powder spectra in Metal–Organic Frameworks (MOFs),[96] the current approach offers an expandable knowledge base and relies on high-quality reference data.

4.3.4 Question-answering based using the Marie system. The natural language interface of Marie enables the retrieval of chemical information across networks of interlinked data *via* a single search entry point. We demonstrate the capability of this feature with a run-through of the steps involved in the natural language processing pipeline, as seen in Fig. 11. The user first inputs a natural language query asking for a list of zeolite materials made of Ge and O only. When the user presses *Enter* or clicks the button with the magnifying glass icon, the system performs two operations: translating the user input into a SPARQL query and identifying which knowledge domain to query; the results of this step are shown below the input field, with the SPARQL query displayed on the left and predicted domain, which is OntoZeolite, on the right. The SPARQL query is then executed against the target knowledge graph to obtain the requested information in a tabular format, showing the chemical formulae of the zeolitic materials asked for by the user. To further enhance the user experience, this structured data, together with the input question, is passed to OpenAI's chat completions API to formulate a concise, human-friendly chatbot response that directly answers the user query.[97]

In evaluating the performance of TWA–Marie, commercial ChatGPT 4, and Gemini Advanced within zeolite chemistry, notable differences in accuracy, detail, and reliability are observed (see Section SI.4 in ESI† for more details). TWA–Marie combines knowledge graph information with a large language model to deliver precise and reliable information substantiated by direct IRI and DOI links. For instance, inquiries regarding the reported unit cell parameters of specific zeolite framework types such as ABW, AHT, and LAU consistently receive accurate responses. In contrast, ChatGPT demonstrates inconsistent accuracy, occasionally providing incorrect or hallucinated data, including misidentifying the crystal system of zeolite ABW or conflating the LAU framework with LTA. Similarly, Gemini Advanced's responses often contain inaccuracies or information irrelevant to the queries posed, like in cases where it is asked about zeolites reported to include pyridine within their


1 **Natural Language Query Input:** The user asks for a list of zeolite materials with framework made of Ge

Common units with special characters: °C °F

Find zeolites with framework made of Ge and O only

Click to view the predicted SPARQL query

```
SELECT ?Material WHERE { ?Material zeo:hasFrameworkComponentOnly/rdfs:label "Ge". ?Material zeo:hasFrameworkComponentOnly/rdfs:label "O". }
```

Click to view the post-processed SPARQL query

Predicted query domain: ontozeolite

3 **Predicted query domain:** The domain is predicted to be OntoZeolite

Translation latency: 2.70s.
SPARQL query execution latency: 4.28s.
Chatbot response latency: 2.15s.

Show IRIs

Show 10 entries Search:

#	ChemicalFormula
1	|(DMA)2(H2O)2|[Ge20O40]
2	|(CID12127913)F|[Ge40O80]
3	|(DABCO)20(H2O)20|[Ge40O80]
4	|(HM+2)2F4|[Ge40O80]
5	|(TrMA)6(H2O)4.5|[Ge32O64]

Showing 1 to 5 of 5 entries

2 **Translated SPARQL query:** The input query in natural language is translated to SPARQL

```
SELECT DISTINCT ?Material ?ChemicalFormula ?Element1 ?Element2 WHERE {
  ?Material zeo:hasChemicalFormula ?ChemicalFormula .
  ?Material zeo:hasFrameworkComponent ?Element1 .
  ?Material zeo:hasFrameworkComponent ?Element2 .
  FILTER NOT EXISTS {
  ?Material zeo:hasFrameworkComponent ?ElementExclude .
  FILTER ( ?ElementExclude NOT IN ( ?Element1, ?Element2, ?Oxygen ) )
}
  SERVICE <> {
  ?Element1 (os:hasElementSymbol|os:hasElementName)/os:value "Ge" .
  ?Element2 (os:hasElementSymbol|os:hasElementName)/os:value "O" .
  ?Oxygen os:hasElementSymbol/os:value "O" .
}
}
```

4 **Structured Response:** The predicted SPARQL query is executed against the OntoZeolite knowledge graph, returning the formula of zeolite materials in a tabular format.

Marie's response

The retrieved data contains information about several zeolites with frameworks made of Ge and O only. Here are the chemical formulas of those zeolites:

1. |(DMA)2(H2O)2|[Ge20O40]
2. |(CID12127913)F|[Ge40O80]
3. |(DABCO)20(H2O)20|[Ge40O80]
4. |(HM+2)2F4|[Ge40O80]
5. |(TrMA)6(H2O)4.5|[Ge32O64]

These zeolites consist solely of the elements Ge (Germanium) and O (Oxygen) in their frameworks.

5 **Chatbot Response:** A large language model processes both the data from the knowledge graph and the input query to output a human-friendly list of zeolite materials.

Fig. 11 The user interface for natural language search, with a breakdown of the processing steps involved.

frameworks. These discrepancies highlight the superiority of TWA–Marie's approach, integrating a knowledge graph with a large language model to provide data-driven and verifiable responses.

5 Summary and conclusions

In this paper, we have detailed a semantic integration of concepts from zeolite chemistry with those of crystalline materials, alongside a focus on chemical species. This integration has been achieved through the curation of chemical, crystallographic, and zeolite data formatted according to an established ontological framework. We populated a comprehensive knowledge graph within the broader TWA model, covering frameworks associated with over two thousand composition-dependent zeolite materials and more than one thousand crystallographic structures linked to over 200 chemical species. This integration ensures that the chemical information becomes machine-actionable, enhancing the efficiency and precision of data queries and retrieval processes. This compatibility enhances the accessibility and actionability of complex chemical data by facilitating its delivery in precise, natural language. Moreover, the combination of the knowledge graph approach with LLM showed a distinct advantage over systems that depend solely on LLMs, which are prone to inaccuracies and data "hallucinations".

In this work, we have further demonstrated the interoperability of zeolitic, species, and crystalline information in a single knowledge graph. Considering the availability of further calculated and machine learning-derived insights, the current implementation has the potential to grow in the near future, encompassing much new information on a variety of chemical species potentially adaptable in different existing and hypothetical zeolites,[83] mechanical properties,[98] and adsorption properties.[99] In addition to this, the ontological description can be extended to other porous materials such as metal–organic frameworks, covalent organic frameworks, and even hydrogen-bonded organic frameworks, linking framework information with crystalline data.

Considering the relevance and need for programmatic study of crystalline information in drug design and materials engineering,[100,101] the presently reported ontological approach provides a promising alternative for crystallographic queries in the near future. This will be realized by further expansion of the OntoCrystal ontology that will be enriched with open crystallographic data,[102] enabling links from molecular instances (individual species) to their crystallographic structures. This extended implementation, in addition to programmatic exploration, will facilitate the study of polymorphs through natural language queries. In the context of zeolite chemistry, semantically representing extensive crystallographic information is expected to enable a more complete programmatic assortment and linking of zeolitic materials, thereby enhancing data completeness. The open availability of our data will likely offer further advantages for educational, fundamental and applied chemistry research. The integration of diverse but interrelated chemical concepts enables tackling complex multicomponent chemical systems such as surface chemistry, reticular chemistry, and supramolecular chemistry,[29] but potentially also composite material systems involving zeolites.[103] This approach offers significant potential for interoperability within complex chemical material systems, thereby motivating continued exploration and detailed characterisation of these systems.

Data availability

All data covered in this research can be automatically queried through TWA-Marie interface (**https://theworldavatar.io/demos/marie/**). Further, comprehensive research data supporting this publication will be made available upon acceptance *via* the University of Cambridge data repository (**https://doi.org/10.17863/CAM.108233**). All codes are accessible *via* the TWA git repository (**https://github.com/cambridge-cares/TheWorldAvatar/tree/main/Agents/ZeoliteAgent**).

Author contributions

AK and MK conceptualized the study. AK authored the original manuscript, which was edited and revised by all authors. PR developed the ontologies with contributions from FF, LP, and AK. DT trained the question-answering system. LP developed the web interface for queries. SG validated the agent instantiation. Data curation was managed by A. K., LP, and PR. MK acquired research funding.

Conflicts of interest

There are no conflicts to declare.

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 authors have applied a Creative Commons Attribution (CC BY) licence to any arising Author Accepted Manuscript version.

Notes and references

1 K. Möller and T. Bein, *Chem. Soc. Rev.*, 2013, **42**, 3689–3707.
2 E. M. Flanigen, *Introduction to Zeolite Science and Practice*, Elsevier, 1991, vol. 58, pp. 13–34.
3 E. M. Flanigen, *Stud. Surf. Sci. Catal.*, 2001, **137**, 11–15.
4 K. B. Tankersley, N. P. Dunning, C. Carr, D. L. Lentz and V. L. Scarborough, *Sci. Rep.*, 2020, **10**, 18021.
5 C. Chizallet, C. Bouchy, K. Larmier and G. Pirngruber, *Chem. Rev.*, 2023, **123**, 6107–6196.
6 H. Xu and P. Wu, *Natl. Sci. Rev.*, 2022, **9**, nwac045.
7 E. Pérez-Botella, S. Valencia and F. Rey, *Chem. Rev.*, 2022, **122**, 17647–17695.
8 B. Yue, S. Liu, Y. Chai, G. Wu, N. Guan and L. Li, *J. Energy Chem.*, 2022, **71**, 288–303.

9 A. Primo and H. Garcia, *Chem. Soc. Rev.*, 2014, **43**, 7548–7561.

10 Y. Wu and B. M. Weckhuysen, *Angew. Chem., Int. Ed.*, 2021, **60**, 18930–18949.

11 S. Kumar, R. Srivastava and J. Koh, *J. CO_2 Util.*, 2020, **41**, 101251.

12 A. Hedström, *J. Environ. Eng.*, 2001, **127**, 673–681.

13 Y. Li, L. Li and J. Yu, *Chem*, 2017, **3**, 928–949.

14 N. F. Himma, A. K. Wardani, N. Prasetya, P. T. Aryanti and I. G. Wenten, *Rev. Chem. Eng.*, 2019, **35**, 591–625.

15 N. E. R. Zimmermann and M. Haranczyk, *Cryst. Growth Des.*, 2016, **16**, 3043–3048.

16 N. Zheng, X. Bu, B. Wang and P. Feng, *Science*, 2002, **298**, 2366–2369.

17 O. M. Yaghi, *ACS Cent. Sci.*, 2019, **5**, 1295–1300.

18 B. Smit and T. L. Maesen, *Chem. Rev.*, 2008, **108**, 4125–4184.

19 V. Van Speybroeck, K. Hemelsoet, L. Joos, M. Waroquier, R. G. Bell and C. R. A. Catlow, *Chem. Soc. Rev.*, 2015, **44**, 7044–7111.

20 W. Chaikittisilp, in *Data-Driven Approach for Rational Synthesis of Zeolites and Other Nanoporous Materials*, John Wiley & Sons, Ltd, 2023, ch. 9, pp. 233–250.

21 M. Moliner, Y. Román-Leshkov and A. Corma, *Acc. Chem. Res.*, 2019, **52**, 2971–2980.

22 D. Schwalbe-Koda and R. Gómez-Bombarelli, *AI-Guided Design and Property Prediction for Zeolites and Nanoporous Materials*, 2023, pp. 81–111.

23 A. Gandhi and M. F. Hasan, *Curr. Opin. Chem. Eng.*, 2022, **35**, 100739.

24 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, *Nat. Commun.*, 2022, **13**, 1667.

25 M. Glauer, F. Neuhaus, S. Flügel, M. Wosny, T. Mossakowski, A. Memariani, J. Schwerdt and J. Hastings, *Digital Discovery*, 2024, **3**, 896–907.

26 Y. Li and J. Yu, *Chem. Rev.*, 2014, **114**, 7268–7316.

27 J. Shin, D. Jo and S. B. Hong, *Acc. Chem. Res.*, 2019, **52**, 1419–1427.

28 A. W. Burton and S. I. Zones, *Stud. Surf. Sci. Catal.*, 2007, **168**, 137–179.

29 A. Kondinski, J. Bai, S. Mosbach, J. Akroyd and M. Kraft, *Acc. Chem. Res.*, 2023, **56**, 128–139.

30 P. Judson, *Knowledge-based Expert Systems in Chemistry: Artificial Intelligence in Decision Making*, Royal Society of Chemistry, 2019, vol. 15.

31 A. Kondinski, A. Menon, D. Nurkowski, F. Farazi, S. Mosbach, J. Akroyd and M. Kraft, *J. Am. Chem. Soc.*, 2022, **144**, 11713–11728.

32 E. J. Corey, *Angew. Chem., Int. Ed.*, 1991, **30**, 455–465.

33 D. A. Pensak and E. J. Corey, *Computer-Assisted Organic Synthesis*, ACS Publications, 1977, ch. 1, pp. 1–32.

34 G. Tecuci, D. Marcu, M. Boicu and D. A. Schum, *Knowledge Engineering: Building Cognitive Assistants for Evidence-Based Reasoning*, Cambridge University Press, 2016.

35 T. Berners-Lee, J. Hendler and O. Lassila, *Sci. Am.*, 2001, **284**, 34–43.

36 A. Kondinski, S. Mosbach, J. Akroyd, A. Breeson, Y. R. Tan, S. Rihm, J. Bai and M. Kraft, *Chem*, 2024, **10**(4), 1071–1083.

37 L. Pascazio, S. Rihm, A. Naseri, S. Mosbach, J. Akroyd and M. Kraft, *J. Chem. Inf. Model.*, 2023, **63**, 6569–6586.

38 F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca and M. Kraft, *J. Chem. Inf. Model.*, 2020, **60**, 108–120.

39 F. Farazi, N. B. Krdzavac, J. Akroyd, S. Mosbach, A. Menon, D. Nurkowski and M. Kraft, *Comput. Chem. Eng.*, 2020, **137**, 106813.

40 J. Bai, R. Geeson, F. Farazi, S. Mosbach, J. Akroyd, E. Bringley and M. Kraft, *J. Chem. Inf. Model.*, 2021, **61**, 1701–1717.

41 J. Bai, S. Mosbach, C. J. Taylor, D. Karan, K. F. Lee, S. D. Rihm, J. Akroyd, A. A. Lapkin and M. Kraft, *Nat. Commun.*, 2024, **15**, 462.

42 J. Deb, L. Saikia, K. D. Dihingia and G. N. Sastry, *J. Chem. Inf. Model.*, 2024, **64**(3), 799–811.

43 K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, S. Cox, W. A. de Jong, M. L. Evans, N. Gastellu, J. Genzling, M. V. Gil, A. K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lála, T. Liu, S. Ma, S. Majumdar, G. W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M. C. Ramos, B. Ranković, S. G. Rodriques, J. N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B. E. Smith, J. Van Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G. A. Zia, A. Scourtas, K. J. Schmidt, I. Foster, A. D. White and B. Blaiszik, *Digital Discovery*, 2023, **2**, 1233–1250.

44 S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang and X. Wu, *IEEE Trans. Knowl. Data Eng.*, 2024, 1–20.

45 S. V. Krivovichev, *Angew. Chem., Int. Ed.*, 2014, **53**, 654–661.

46 S. V. Krivovichev, *Microporous Mesoporous Mater.*, 2013, **171**, 223–229.

47 International Zeolite Association (IZA), https://www.iza-online.org/, accessed: April 25, 2024.

48 B. W. Boal, J. E. Schmidt, M. A. Deimund, M. W. Deem, L. M. Henling, S. K. Brand, S. I. Zones and M. E. Davis, *Chem. Mater.*, 2015, **27**, 7774–7779.

49 F. Liebau, *Microporous Mesoporous Mater.*, 2003, **58**, 15–72.

50 N. A. Anurova, V. A. Blatov, G. D. Ilyushin and D. M. Proserpio, *J. Phys. Chem. C*, 2010, **114**, 10160–10170.

51 V. A. Blatov, O. Delgado-Friedrichs, M. O'Keeffe and D. M. Proserpio, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 2007, **63**, 418–425.

52 V. A. Blatov, A. P. Shevchenko and D. M. Proserpio, *Cryst. Growth Des.*, 2014, **14**, 3576–3586.

53 V. A. Blatov, G. D. Ilyushin and D. M. Proserpio, *Chem. Mater.*, 2013, **25**, 412–424.

54 D.-K. Nguyen, V.-P. Dinh, H. Q. Nguyen and N. T. Hung, *J. Chem. Technol. Biotechnol.*, 2023, **98**, 1339–1355.

55 A. Kondinski and K. Y. Monakhov, *Chem.–Eur. J.*, 2017, **23**, 7841–7852.

56 S. R. Hall, F. H. Allen and I. D. Brown, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 1991, **47**, 655–685.

57 H. J. Bernstein, J. C. Bollinger, I. D. Brown, S. Gražulis, J. R. Hester, B. McMahon, N. Spadaccini, J. D. Westbrook and S. P. Westrip, *J. Appl. Crystallogr.*, 2016, **49**, 277–284.

58 S. van Smaalen, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1991, **43**, 11330–11341.

59 S. V. Smaalen, *Crystallogr. Rev.*, 1995, **4**, 79–202.

60 M. J. Statt, B. A. Rohr, D. Guevarra, J. Breeden, S. K. Suram and J. M. Gregoire, *Digital Discovery*, 2023, **2**, 909–914.

61 S. D. Rihm, J. Bai, A. Kondinski, S. Mosbach, J. Akroyd and M. Kraft, *Nexus*, 2024, **1**, 100004.

62 K. Hippalgaonkar, Q. Li, X. Wang, J. W. Fisher, J. Kirkpatrick and T. Buonassisi, *Nat. Rev. Mater.*, 2023, **8**, 241–260.

63 X. Peng and X. Wang, *MRS Bull.*, 2023, **48**, 179–185.

64 G. Deepak, Z. Gulzar and A. A. Leema, *Comput. Electr. Eng.*, 2021, **96**, 107604.

65 Digital Twin Hub, *Climate Resilience Demonstrator*, https://digitaltwinhub.co.uk/credo/credo/, 2023, accessed: March 5, 2024.

66 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, *Building a Cross-Sector Digital Twin*, Centre for Digital Built Britain, 2022.

67 N. B. Krdzavac, S. Mosbach, D. Nurkowski, P. Buerger, J. Akroyd, J. W. Martin, A. Menon and M. Kraft, *J. Chem. Inf. Model.*, 2019, **59**, 3154–3165.

68 A. Menon, L. Pascazio, D. Nurkowski, F. Farazi, S. Mosbach, J. Akroyd and M. Kraft, *ACS Omega*, 2023, **8**, 2462–2475.

69 J. Akroyd, S. Mosbach, A. Bhave and M. Kraft, *Data-Centric Eng.*, 2021, **2**, e14.

70 S. Mosbach, A. Menon, F. Farazi, N. Krdzavac, X. Zhou, J. Akroyd and M. Kraft, *J. Chem. Inf. Model.*, 2020, **60**, 6155–6166.

71 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, *ACS Omega*, 2020, **5**, 18342–18348.

72 A. C. Ghosh, A. Legrand, R. Rajapaksha, G. A. Craig, C. Sassoye, G. Balázs, D. Farrusseng, S. Furukawa, J. Canivet and F. M. Wisser, *J. Am. Chem. Soc.*, 2022, **144**, 3626–3636.

73 X. Zhou, D. Nurkowski, S. Mosbach, J. Akroyd and M. Kraft, *J. Chem. Inf. Model.*, 2021, **61**, 3868–3880.

74 X. Zhou, D. Nurkowski, A. Menon, J. Akroyd, S. Mosbach and M. Kraft, *Digital Chem. Eng.*, 2022, **3**, 100032.

75 D. Tran, L. Pascazio, J. Akroyd, S. Mosbach and M. Kraft, *ACS Omega*, 2024, **9**, 13883–13896.

76 S. Staab and R. Studer, *Handbook on Ontologies*, Springer Verlag Berlin Heidelberg, 2004.

77 B. Glimm, I. Horrocks, B. Motik, G. Stoilos and Z. Wang, *J. Autom. Reas.*, 2014, **53**, 245–269.

78 K. R. Taylor, R. J. Gledhill, J. W. Essex, J. G. Frey, S. W. Harris and D. C. De Roure, *J. Chem. Inf. Model.*, 2006, **46**, 939–952.

79 P. Murray-Rust, *Nature*, 2008, **451**, 648–651.

80 M. Kraft and S. Mosbach, *Philos. Trans. R. Soc., A*, 2010, **368**, 3633–3644.

81 K. M. Jablonka, L. Patiny and B. Smit, *Nat. Chem.*, 2022, **14**, 365–376.

82 DCMI Usage Board, *Bibliographic Ontology (BIBO) in RDF, Maintainer: DCMI Usage Board (contact: Bruce d'Arcus), 2016-05-11*, https://www.dublincore.org/specifications/bibo/bibo/bibo.rdf.xml, Creators: Bruce D'Arcus, Frédérick Giasson.

83 E. Pan, S. Kwon, Z. Jensen, M. Xie, R. Gómez-Bombarelli, M. Moliner, Y. Román-Leshkov and E. Olivetti, *ACS Cent. Sci.*, 2024, **10**(3), 729–743.

84 H. Rijgersberg, M. van Assem and J. Top, *Semant. Web.*, 2013, **4**, 3–13.

85 H. Rijgersberg, *OM – Ontology of units of Measure*, 2023, https://github.com/HajoRijgersberg/OM.

86 L. Komissarov and T. Verstraelen, *Sci. Data*, 2022, **9**, 61.

87 *Database of Zeolite Structures*, https://re3data.org, Registry of Research Data Repositories, 2024, DOI: 10.17616/R3HS6N.

88 S. Yang, M. Lach-Hab, I. I. Vaisman, E. Blaisten-Barojas, X. Li and V. L. Karen, *J. Phys. Chem. Ref. Data*, 2010, **39**, 033102.

89 C. Zheng, Y. Li and J. Yu, *Sci. Data*, 2020, **7**, 107.

90 L. Pascazio, D. Tran, S. Rihm, J. Bai, J. Akroyd, S. Mosbach and M. Kraft, *Question-Answering System for Combustion Kinetics*, c4e-Preprint Series, Cambridge Technical Report Technical Report 315, 2023.

91 Y. Li, G. Zhu, Y. Wang, Y. Chai and C. Liu, *Microporous Mesoporous Mater.*, 2021, **312**, 110790.

92 Blazegraph™ DB, 2024, https://blazegraph.com/, last accessed: 2024-04-12.

93 M. Q. Lim, X. Wang, O. Inderwildi and M. Kraft, in *The World Avatar—A World Model for Facilitating Interoperability*, ed. O. Inderwildi and M. Kraft, Springer International Publishing, Cham, 2022, pp. 39–53.

94 E. L. First, C. E. Gounaris, J. Wei and C. A. Floudas, *Phys. Chem. Chem. Phys.*, 2011, **13**, 17339–17358.

95 *Collection of Simulated XRD Powder Patterns for Zeolites*, ed. M. Treacy and J. Higgins, Elsevier Science B.V., Amsterdam, 5th edn, 2007.

96 H. Wang, Y. Xie, D. Li, H. Deng, Y. Zhao, M. Xin and J. Lin, *J. Chem. Inf. Model.*, 2020, **60**, 2004–2011.

97 OpenAI, *ChatGPT (Mar 14 version)*, https://chat.openai.com/chat, 2024, Large language model.

98 J. D. Evans and F.-X. Coudert, *Chem. Mater.*, 2017, **29**, 7833–7839.

99 Y. Sun, R. F. DeJaco, Z. Li, D. Tang, S. Glante, D. S. Sholl, C. M. Colina, R. Q. Snurr, M. Thommes, M. Hartmann and J. I. Siepmann, *Sci. Adv.*, 2021, **7**, eabg3983.

100 M. J. Bryant, S. N. Black, H. Blade, R. Docherty, A. G. Maloney and S. C. Taylor, *J. Pharm. Sci.*, 2019, **108**, 1655–1662.

101 S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder and K. A. Persson, *Comput. Mater. Sci.*, 2015, **97**, 209–215.

102 S. Gražulis, D. Chateigner, R. T. Downs, A. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck and A. Le Bail, *J. Appl. Crystallogr.*, 2009, **42**, 726–729.

103 L. R. Rad and M. Anbia, *J. Environ. Chem. Eng.*, 2021, **9**, 106088.