Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

Automated Rational Design of Metal-Organic Polyhedra

Aleksandar Kondinski[1], Angiras Menon[1], Daniel Nurkowski[2], Feroz Farazi[1], Sebastian Mosbach[1], Jethro Akroyd[1], Markus Kraft[1,2,3,4,5]

released: February 28, 2022

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

[2] CMCL Innovations
Sheraton House
Cambridge
CB3 0AX
United Kingdom

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

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

[5] The Alan Turing Institute
London
United Kingdom

Preprint No. 292

UNIVERSITY OF CAMBRIDGE

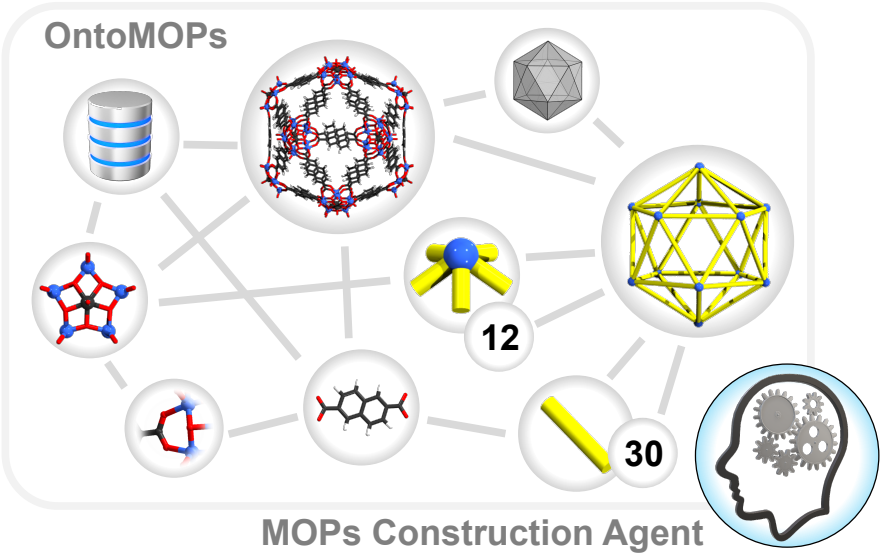
Keywords: Knowledge Engineering, Metal-Organic Polyhedra, Assembly Models, Ontology, Chemical Data

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

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

Abstract

Metal-organic polyhedra (MOPs) are hybrid organic-inorganic nanomolecules, whose rational design depends on harmonious consideration of chemical complementarity and spatial compatibility between two or more types of chemical building units (CBUs). In this work, we apply knowledge engineering technology to automate the derivation of MOP formulations based on existing knowledge. For this purpose we have: i) curated relevant MOP and CBU data; ii) developed an assembly model concept that embeds rules in the MOP construction; iii) developed an OntoMOPs ontology that defines MOPs and their key properties; iv) input agents that populate The World Avatar (TWA) knowledge graph; and v) agents that, using information from TWA, derive a list of new constructible MOPs. Our result provides rapid and automated instantiation of MOPs in TWA and unveils the immediate chemical space of known MOPs, thus shedding light on new MOP targets for future investigations.



Highlights

- Conceptualisation of metal-organic polyhedra (MOPs) by assembly models.
- Curation of MOPs and building unit data for knowledge graph development.
- Ontological representation of MOPs and their related concepts.
- Knowledge-based rational derivation of new MOP designs by a software agent.
- Mapping of the immediate MOP chemical space for future studies.

Contents

1 Introduction

Molecular engineering is an emerging study of molecular components with the aim of tailoring their programmed assembly towards new and functional materials [73]. Molecular engineering relies on a cognitive design thinking approach (*i.e.* rational design), and thus it has show a strong innovation reliability across multiple domains spanning nanotechnology [14, 74], molecular machinery [7], OLEDs [37], flexible solar cells and other technologies [38]. A special advancement to molecular engineering has been the conceptualisation of building blocks, that is molecular components that can be developed and reused across different material families. In this regard, the combination of inorganic and organic building units has subsequently led to the flourish of various molecular and functional hybrids such as supramolecular assemblies [47, 65], hybrid polyoxometalates (POMs) [4, 40], metal-organic polyhedra (MOPs) [26, 46, 70] and also extended reticular systems like metal-organic frameworks (MOFs) [48, 50].

Among the different molecular and nanoscopic hybrids, MOPs are renowned for their virtual adoption of shapes of highly symmetrical polyhedra [70]. MOPs also share similarities to other more early established hybrids, which may have contributed to their slower comprehensive recognition as a distinct material domain [26, 46]. MOPs are typically constructed from a pair of complementary organic and inorganic chemical building units (CBU). Cases when more than two CBUs form MOPs are also known [46]. The organic building units in MOPs are typically carboxylate binding which makes them very similar to many MOFs [26], but also differentiates them from other types of supramolecular assemblies where different binding functionalities may prevail [60]. The inorganic units in MOPs may be monometallic, but they are predominantly bimetallic and multimetallic [46]. Multimetallic inorganic CBUs may be metal-oxo clusters as POMs [49]. Like MOFs and other supramolecular cages, MOPs are porous and exhibit internal cavities suitable for molecular guest encapsulation [26, 77], and gas capture and separation (e.g. CO_2 [66, 76]. The high number of metal centres and nanoscopic size also makes MOPs attractive in catalysis [36, 67], for nanoscopic components for the preparation of porous soft materials [33], and porous salts [27].

Interested in the development of future AI-driven chemical scientists and laboratories capable of solving emerging real world problems [6, 34, 39], we envision a tremendous opportunity for the development of new knowledge and logic driven technologies that are capable of emulating different aspects of the expert's decision making process. Knowledge engineering (KE) is one technology [69], that efficiently couples ontological representation of key concepts, relational data in a knowledge graph (KG) and logic execution software agents towards a particular goal. (see Figure 1.b). In comparison to the widely used database approaches for storage and exploration of chemical data, KGs are based on semantics depicting a complex network of concepts and information, thus they are relatively uncharted territory in chemistry [53]. Over the past decade, KGs have been aiding the elucidation of the relationship between chemical structures and biological responses [13], which has an obvious relevance in the development of new pharmaceuticals [11]. KGs can be highly modular and dynamic; and as such their application has become popular cross many different industries [1, 44]. Synergistic use of KGs can be established by interconnecting KGs in an interoperable manner, towards solving a com-

plex goal. This has enabled the creation of a world model called *The World Avatar* (TWA) which potentially comprises any concept, instances of these concepts and agents that operate on both concepts and instances. Hence, TWA can be viewed as an universal digital twin (UDT) [1, 16]. The chemistry space in TWA so far contains information on quatum chemistry, chemical species, reaction networks and experimental observations including agents capable of model calibration and cross domain linkage [5, 19, 20, 45].

The purpose of this work is to expand on the capabilities of *The World Avatar* by developing knowledge graph technologies for the representation and rational design of MOP and projection of their immediate chemical space (Section 2). To achieve this, we first develop a concept of assembly models to represent the geometric features of a MOP and how it is constructed from its constituent CBUs (Section 3). These relations between chemical and topological features are encoded via the newly developed "OntoMOPs" ontology representing MOPs in TWA. MOP data has been systematically curated, cleaned and organised with consideration of their composition and structure. TWA is populated with 151 MOP and 137 CBU instances with a set of custom built software tools (Section 4). Finally a MOP Discovery agent has been developed and used to perform a series of queries and set operations (Section 5.1) from which it identifies new MOPs formulations by considering chemical and spatial compatibility of different CBUs known to build MOPs (Section 5.2).

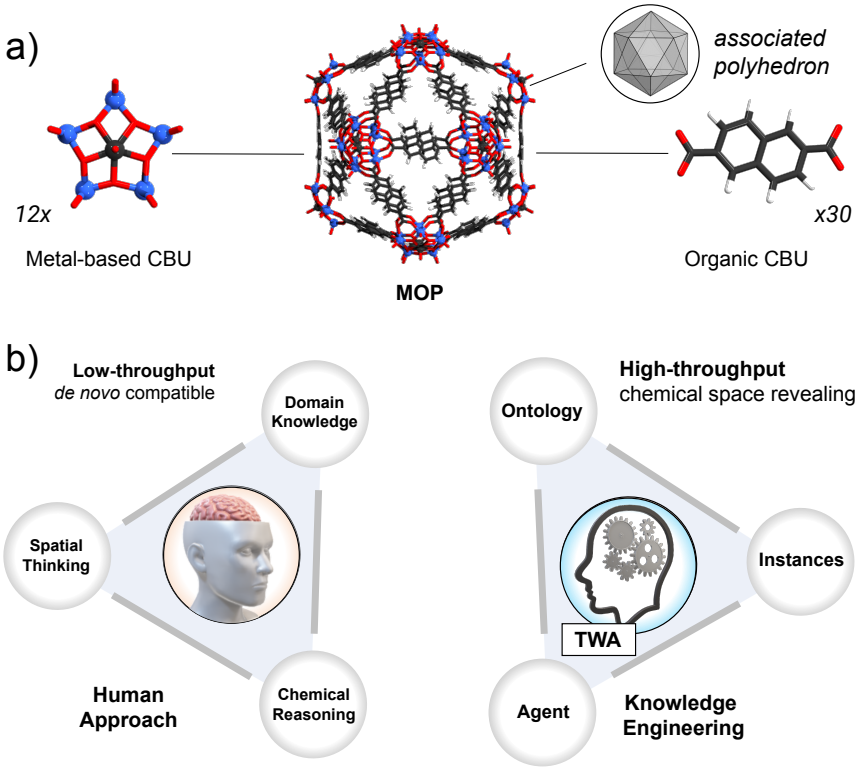


Figure 1: *a) Ball and stick representation of a MOP, its components and perceived shape. b) Schematic representation of the human approach vs. the knowledge engineering approach when rationally designing MOPs.*

2 Immediate Chemical Space and its Uncertainties

"How can one design a structure if its "blueprint" is unknown?" is a question that Yaghi and coworkers raise in their recent perspective defining the digital reticular chemistry covering 1-/2-/3-dimensional metal-organic materials [52]. This overview provides a perspective on how to merge machine learning (ML), database technology and mechatronics for the automated discovery and development of MOFs. In the work, the authors acknowledge the vastness of chemical space that emerges as a result of building block, topological and isomeric variability; however, they also emphasise the value of being able to pre-select and recognise viable material targets with promising pre-calculated properties. This is in contrast to the more common material development followed by property description.

In the article material construction is described as the linking of different building units based on "empirical" knowledge of what the structural outcome might be [52]. The authors see this approach as having "a heavily reliance on experience" and circumventing this represents an open challenge. However, this empirical knowledge approach also comes with uncertainties, some of which may derive from the synthetic complexity where the reagents likely includes additional chemical species not considered in the conceptual modelling, but also due to uncertainties in the expected outcome. Secondary building units "SBUs" that appear compatible with a particular symmetric framework, when actually reacting in a synthetic pathway may form another unanticipated structure at the end. This can occur because the SBUs may adopt different modularities [52] during different reactive processes. These uncertainties arising from different modularities are genuine and they are not unique to MOFs and COFs, but also to MOPs [46].

From a viewpoint of molecular engineering, a key question is how many and what variety of new structures can be constructed based on known building units? Answering this complex question, provides: i) a better overview on what new materials are in the immediate vicinity of our current knowledge; ii) the possibility to estimate the structural uncertainties occurring when a pair of building units can construct more than one structure. An automated approach to this problem suggests potential formulation targets. Molecular modelling and calculations can then be used to predict material properties. This in turn is useful for future targeted synthesis. Consequently, the "immediate chemical space" (ICS) can be unearthed in this way. The ICS is thus predominantly focused on "constructible" topologies without further explicit concern of how many additional constructed derivatives can be combinatorially derived as a function of conformational and configurational variances in the redox, pronation and chiral nature of the building units. In this view, the ICS is an instance-based projection that at the same time is restrictive, but also pragmatic in terms of molecular engineering.

In contrast to the ML and database approach which essentially relies on learning from vast amounts of data [52], the KE builds on the knowledge and experience of a domain expert and thus new predictions can also be made for domains where data is not vast. The KE approach also provides the possibility to formulate new concepts and assess their value in terms of algorithmic output quality. In this context, we have effectively differentiated between the chemical and geometric nature of Yaghi's SBU concept [71], thus developing a new representation *via* a chemical building unit "CBU" that functions as a generic (*i.e.* geometric) building unit "GBU". Topologically complementary GBUs act as the key com-

ponents in the construction of assembly models (AMs) that then provide the "blueprints" for the formulation of MOPs based on complementary CBUs related to the starting GBUs (see section 3 for more details). By studying the relationship between CBUs, GBUs, AMs and MOPs we can project the ICS of MOPs. As more than one outcome may be formed when two CBUs interact, we obtain awareness of the uncertainty which is useful when designing a synthetic approach. When the outcome is a new and an unanticipated MOP then this structure and its AM are added to the knowledge graph, followed by an update of the ICS in an instance-based manner.

The ICS is part of the overall chemical space, and it connects the known domain (i.e. experimentally verified MOPs) with the uncharted or deep chemical space (see Figure 2). The MOP instances of the ICS are rationally designed constructs based on known CBUs. The automated rational proposal of constructible MOPs is not only of synthetic interest, but also in terms of molecular modelling and calculations. Unlike the modelling and calculation of organic cages [18], accurate calculations on multi-metallic MOPs normally cannot be obtained by forcefield calculation [62], and thus more computationally demanding DFT approaches are needed [35]. The latter approach can be very informative in terms of structure and electronic properties, and when a particular target fulfills criteria to be regarded as realistic or "viable" [32], the predictions of its properties can be suitable for further selection of technologically relevant targets [30].

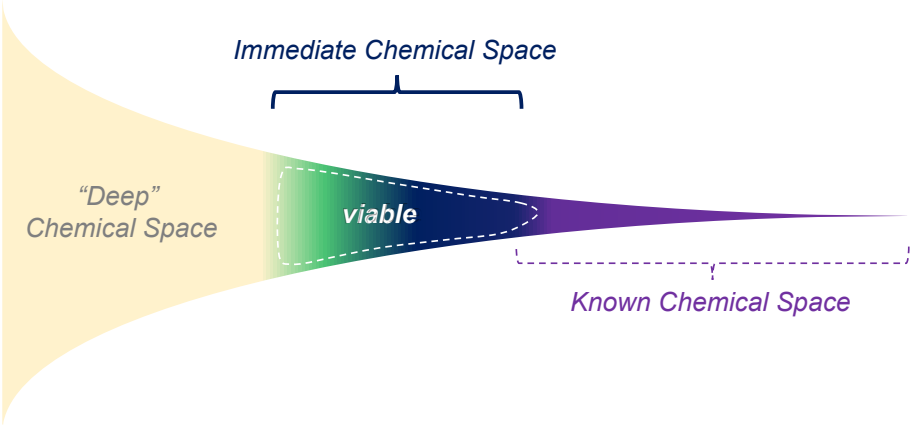


Figure 2: *Schematic illustration of the three regions of the chemical space of MOPs: known domain, its immediate chemcial space that can be logically constructed and the uncharted (i.e. deep) chemical space normally "unlocked" by new AM and CBU development.*

3 Assembly Models

3.1 Polyhedra Modelling during Early Cognitive Development

In contrast to adults, children learn how to think abstractly through sensory input [15]. Construction of polyhedral and reticular assemblies is an abstract and intellectually chal-

lenging topic. However, research with didactic toy-based hands-on manipulatives points to the contrary. Using a generic set of interlocking disks and only the restriction to build symmetrically, children have been shown to be able to construct sub-components and to assemble them into larger high-symmetry assemblies resembling reticular and polyhedral structures [42, 43]. Children are able to achieve this in the absence of prior mathematical knowledge (e.g. dihedral angles) through playful experimentation with the different sub-components, leading them to discover assemblies of reticular and polyhedral materials. This motivates the concept of an assembly model (AM) for MOPs, by which a larger structure is assembled from smaller sub-components, in this case generic building units (GBUs). The assembly model concept also provides a framework of meta-rules for algorithmic discovery of new MOPs, analogous to how children intuitively derive new structures from sub-components without explicit instruction.

3.2 Chemical Complementarity

Whether two CBUs are chemically complementary depends on the features of their "binding sites". In MOPs, the interaction is typically between cationic metal-based CBUs and anionic organic CBUs acting as Lewis acids and bases respectively. The organic ligands typically are bidentate (carboxylate) ligands but other modularities may be observed as well. For successful integration in highly symmetrical assemblies, the metal sites also need to connect to the organic ligands in an orderly manner. Finally the local stereochemistry between the binding sites is another important feature. Within MOPs, the binding sites of a pair of complementary CBUs are well aligned with the virtual line connecting the central points of each CBU. This is normally different for many other supramolecular coordination cages (e.g. pyridyl-imine that bind sideways) and *mer-*/*fac*-isomerism at each site can occur [9]. The basic aspects of chemical complementarity need to be taken into consideration when structures are being algorithmically assembled.

3.3 Topological Compatibility

Coordination cages comprising single metal nodes (M) and organic bridging ligands (L) are typically noted as M_xL_y (*e.g.* $M_{12}L_{24}$ [29]. However, the latter notation does not explicitly describe the overall arrangement and may cause ambiguity when describing isomeric topologies such as cuboctahedral and anticuboctahedral $M_{12}L_{24}$ [46]. The ambiguity can be eliminated when describing MOPs as polyhedral shapes [70]. In the latter approach, a particular atom or a moiety is aligned with an element of a polyhedral shape (*e.g.* corner, edge or face). However, MOPs are highly symmetrical molecules (*i.e.* "Keplerates") [56], and so differences in prioritisation of one molecular fragment over the other may lead to envisioning more than one single shape, leading to correct but inconsistent shape descriptions.

To solve problems with ambiguities and shape inconsistencies we derived an "assembly model" based approach. In our approach, a MOP is envisioned as a highly symmetrical assembly comprised of a pair of chemical building units (CBUs) appearing in strictly defined numbers. Each CBU shows particular modularity and shape features similar to

that of a coordination complex which we refer to as "planarity". The combination of modularity and planarity provides a foundation to define a virtual "generic building unit" (*i.e.* GBU). Similarly, to the CBUs, GBUs appearing in strictly defined numbers can interconnect into larger and virtual Assembly Models (AMs), which in the case of MOPs are polyhedral and cage-like. The AMs come with an ideal symmetry point group and in terms of interconnectivity resemble the MOP. In this way, AMs act as a "construction template" for MOPs. Considering that one needs at least two GBUs to construct an AM, the AM has the advantage to relate to a single shape. An illustration of this is the MOP $[WV_5O_{11}]_{12}[C_{10}H_6(CO_2)_2]_{30}^{12-}$ which is comprised of twelve inorganic $[WV_5O_{11}]^{4+}$ CBUs functioning as "5-pyramidal" GBUs and thirty organic $[C_{10}H_6(CO_2))_2]_{30}^{2-}$ CBUs functioning as "2-linear" GBUs. The latter MOP has an assembly model $(5\text{-pyramidal})_{12}(2\text{-linear})_{30}$ with I_h symmetry (see Figure 3.a).

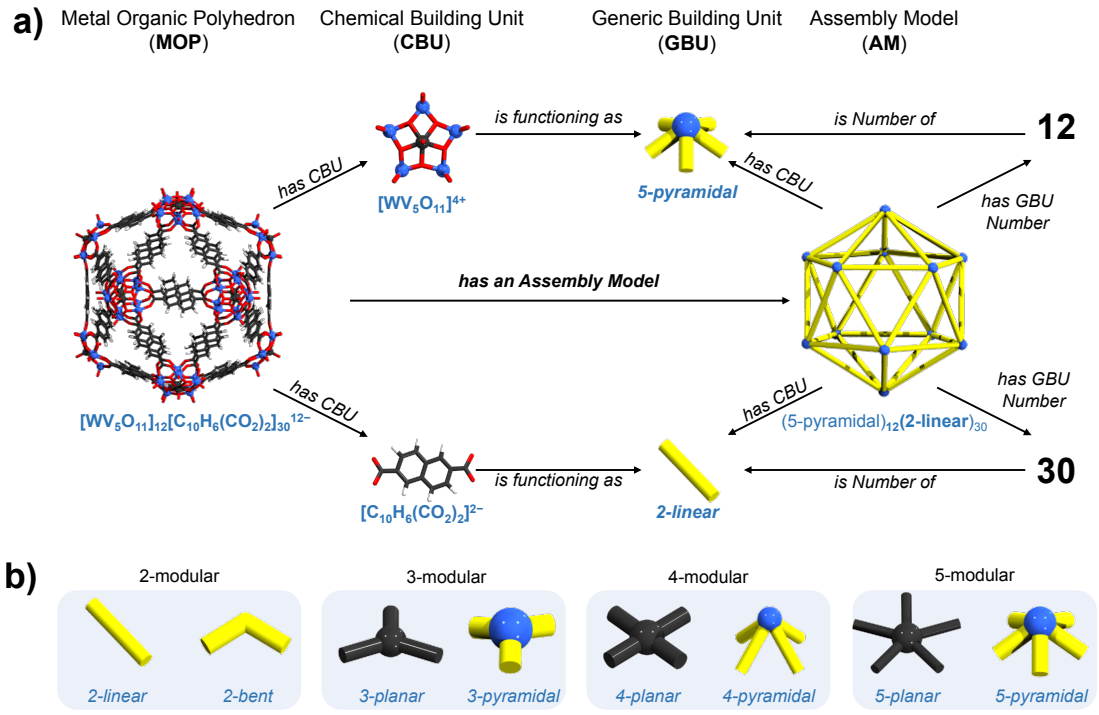


Figure 3: *a) Relations between MOPs, CBUs, GBUs and assembly models. b) Four general types of GBUs.*

3.4 Derivation of Assembly Models

Based solely on planarity and modularity one can derive a set of GBUs (see Figure 3.b). This set of GBUs is sufficient to build many different AMs resembling different shapes. This is because the GBUs can be abstractly compared to elements of a polyhedron. For example, 2-linear building units derive from edges, while 3-, 4- and 5-pyramidal GBUs typically act as vertices. On the other hand, the 3-, 4- and 5-planar GBUs align well with the centre of the trigonal, square and pentagonal faces respectively. The 2-bent GBUs can be seen as edge-based cross-points connecting planar GBUs from different faces of

the polyhedron (see Figure 4). The derivation of assembly models from the platonic

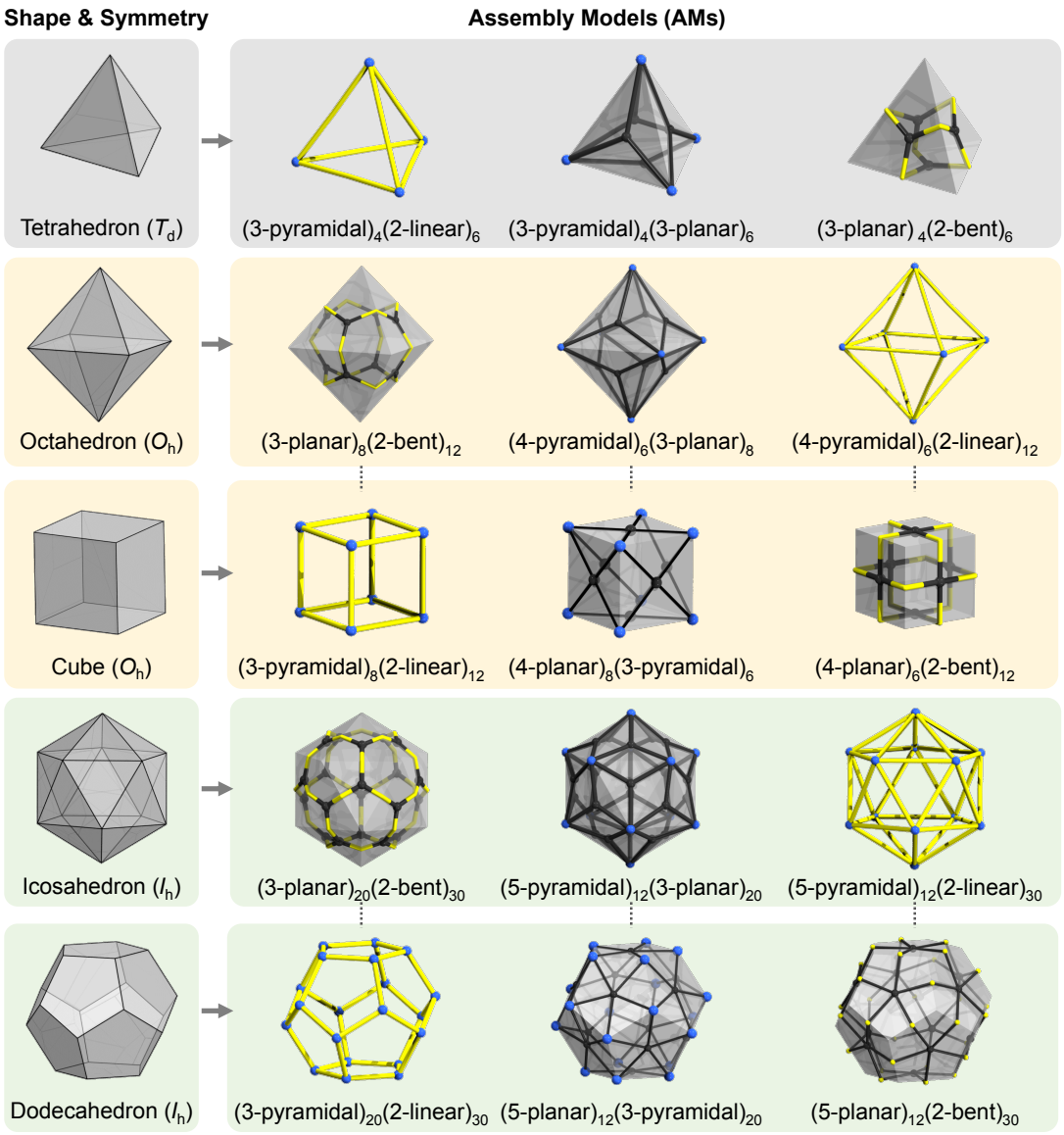


Figure 4: *Derivation of assembly models from the shape of the well-known platonic solids.*

solids provides two additional insights. First, the close interconnection of an AM with a single shape is essential because, most fundamentally, it is not only the building units that define the MOP. In return, the symmetry and shape of the assembly model "softly encode" particular properties of the building units, such as differences in dihedral angles. In this regard, a "3-pyramidal" GBU involved in the construction of a tetrahedral (3-pyramidal)$_4$(2-linear)$_6$ assembly model is not the same as the "3-pyramidal" GBU involved in the construction of dodecahedral (3-pyramidal)$_{20}$(2-linear)$_{30}$ (i.e. the dihedral increase from 70.52° to 116.56°). Further on, pairs of shapes sharing the same symmetries derive pairs of "inverse" assembly models where the GBU retains its modularity. Still, there is an inversion in terms of planarity (*i.e.* planar becomes pyramidal, linear

becomes bent and *vice versa*). One example may be the O_h-symmetric (4-pyramidal)$_6$(3-planar)$_8$ and (4-planar)$_8$(3-pyramidal)$_6$ models that derive from an octahedron and cube respectively. A virtual transformation from such a pair of assembly models goes through yet another (4-pyramidal)$_6$(3-pyramidal)$_8$ assembly model, whose shape may be traced to the Catalan-type rhombic dodecahedron (*vide infra*).

4 The World Avatar – OntoMOPs

4.1 MOP discovery as part of a digital ecosystem

Pragmatic multi-scale material development connecting lab-scale to industrial-scale production relies on accurate life cycle assessment [59]. In the context of digital transformation, the latter is a real cross-domain world problem that can be virtually represented by a universal digital twin. The universal digital twin receives an influx of knowledge, operates through a complex network of concepts, relationships, and synergetic software agents that simulate and analyse different what-if scenarios, based on which decisions are made and implemented [16, 17].

The World Avatar (www.theworldavatar.com) is a universal digital twin, implemented using Semantic Web technology (see Figure 5) [8]. The choice of the technology is based on the FAIR Guiding Principles for scientific data, that is: findable, accessible, interoperable and reusable [75]. In the context of chemistry, TWA hosts a federation of chemical and process development ontologies combining experimental, modelling and theoretical aspects [5, 19, 20, 45]. The chemical ontologies including the herein developed OntoMOPs

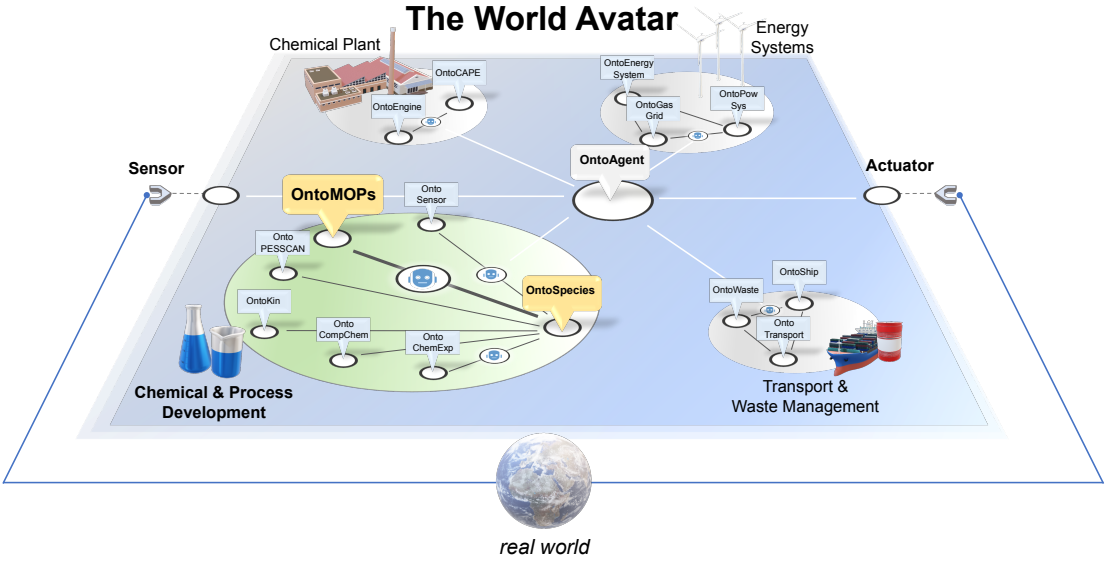


Figure 5: *A selection of ontologies and their connectivity which have been integrated in TWA. OntoMOPs and OntoSpecies are part of the Chemical and Process knowledge representation.*

can share concepts with other ontologies, while software agents can enable interoperability, allowing for complex queries and model phenomena.

The World Avatar platform is cross-domain and multiscale operational digital twin [54]. Considering the urgency and interest in industrialisation of metal organic material hybrids [12], *The World Avatar* has the potential to connect material development [6] with scaled-up process implementation in chemical plants, with further optimisation of the energy consumption, material logistics, and waste minimisation in the overall process.

4.2 Ontological Modelling

To apply the knowledge engineering approach [69], we developed the OntoMOPs ontology iteratively, following standard ontology development practices [21, 23, 24, 28, 57, 61, 72]. The primary goal of the OntoMOPs ontology is to provide semantics to the relationship between MOPs, CBUs and assembly models, ultimately laying the foundation for the development of a knowledge graph that is comprehensible to agents that can be integrated in TWA. The second goal of the OntoMOPs ontology is to provide a semantics-enabled complex query answering system that can inform professionals working on the modelling and preparation of MOPs. The former targets offer a way to define the scope of the ontology. The scope, in this case, is to answer problems regarding the construction of MOPs by providing information that can be used for informed decisions.

Our work depends on developing a Terminological Component that essentially defines classes and properties and a domain vocabulary (*i.e.* TBox). The assertion component (*i.e.* ABox) brings facts associated with the concepts of the TBox (*i.e.* information about MOPs, CBUs and AMs). The combination of TBox and ABoxes can then be used to answer the following competency questions:

- List all MOPs having a particular CBU.
- List all MOPs having a particular AM.
- What type of AM have been constructed using a particular CBU?
- Show all MOPs having tetrahedral shape.
- Show all GBUs required to form a particular shape/AM.
- Show the substituting functionality of a particular CBU.
- What is the associated modularity of a particular species acting as a CBU in MOPs?

To answer these questions, we structure our ontology into three main components (see Figure 6). These components and concepts are created and interconnected using is-a, has-a and is-functioning-as relations. In the MOP component, the main concept is a Metal-Organic Polyhedron which "is-a" Coordination Cage pointing out of our ontology. The Metal-Organic Polyhedron "has-a" Chemical Building Unit and "has-an" Assembly Model, representing the two central concepts in the second and third components, respectively. The Chemical building unit is interconnected to the Assembly Model component through "isFunctioningAs" relation pointing to the Generic building Unit concept.

In the MOP component, we see connections of the MOPs class with other concepts such as MOPcharge, MOPformula, and molecular mass. The concept of MOP also connects to the concept of Provenance, which contains data properties such as the DOI number of the article where a particular MOP is being reported. As many MOPs are related to motifs in crystalline materials, we also connected the concept of MOP to a CCDC number that can help locate the structure of the MOP in the Cambridge Crystallographic Data Centre [3]. The MOP component also provides opportunities for future developments. One example is the presence of the "Cavity" and "CavityVolume" which are intended to be populated in near future with calculated void data, relevant for porosity applications.

In the Assembly Model component, the concept Assembly Model is connected to GBU and a GBU Number via has-a relations. The assembly model is also related to a symmetry point group and polyhedral shapes. Here, polyhedral such as Tetrahedron, Octahedron, Cube, Dodecahedron, Icosahedron, Rhombicuboctahedron and Cuboctahedron are encoded. The polyhedral shape also has a data property - a shape symbol that uses the letter nomenclature for polyhedra reported in the reticular chemistry resource [58]. The planarity and the modularity are encoded as data properties of the GBU.

The CBU component provides a connection between the OntoMOPs ontology and the OntoSpecies ontology. OntoSpecies is an ontology currently consisting of nearly 11000 instances of chemical species for which there are a number of properties. This includes geometry, charge, spin multiplicity and InChI. The OntoSpecies ontology, has been primarily introduced to help with identifying chemical species uniquely [20]. This identification occurs via Internationalized Resource Identifiers (IRIs) that help to connect chemical species with CBUs of MOPs, labeled using arbitrary strings. The CBU component in OntoMOPs does not aim to store these properties again; however, it models what chemical functionalities relate to the particular species in the context of the larger MOP assembly. These functionalities may be related to the (stereo)chemical nature of the binding site and thus used to model information suitable for distinguishing chemical complementarity between two CBUs. The CBU component also models information related to the central component, namely the presence of substituents and spacer groups, which can provide help when querying MOP for a specific substituent or functionality. Using IRIs, the CBU component is connected to one or more GBUs, which models in how many different ways the CBU can connect and build a structure.

The OntoMOPs Ontology consists of 32 classes, 25 object properties and 18 Data properties. The concepts are consistently arranged when exploring using the HermiT reasoner [25, 55].

4.3 MOP Information and Geometry Data Curation

When collecting information and geometry data on MOPs and their CBUs, we kept in mind that although synthetic chemists may benefit from the projections of our work, our work in the first line is intended to aid directly future high-throughput computations of MOPs. According to the reviewed literature, the latter domain of MOP research is currently lacking in pace compared to experimental developments [26, 46]. Computations, especially DFT-based ones, can provide further information on optimised geometry,

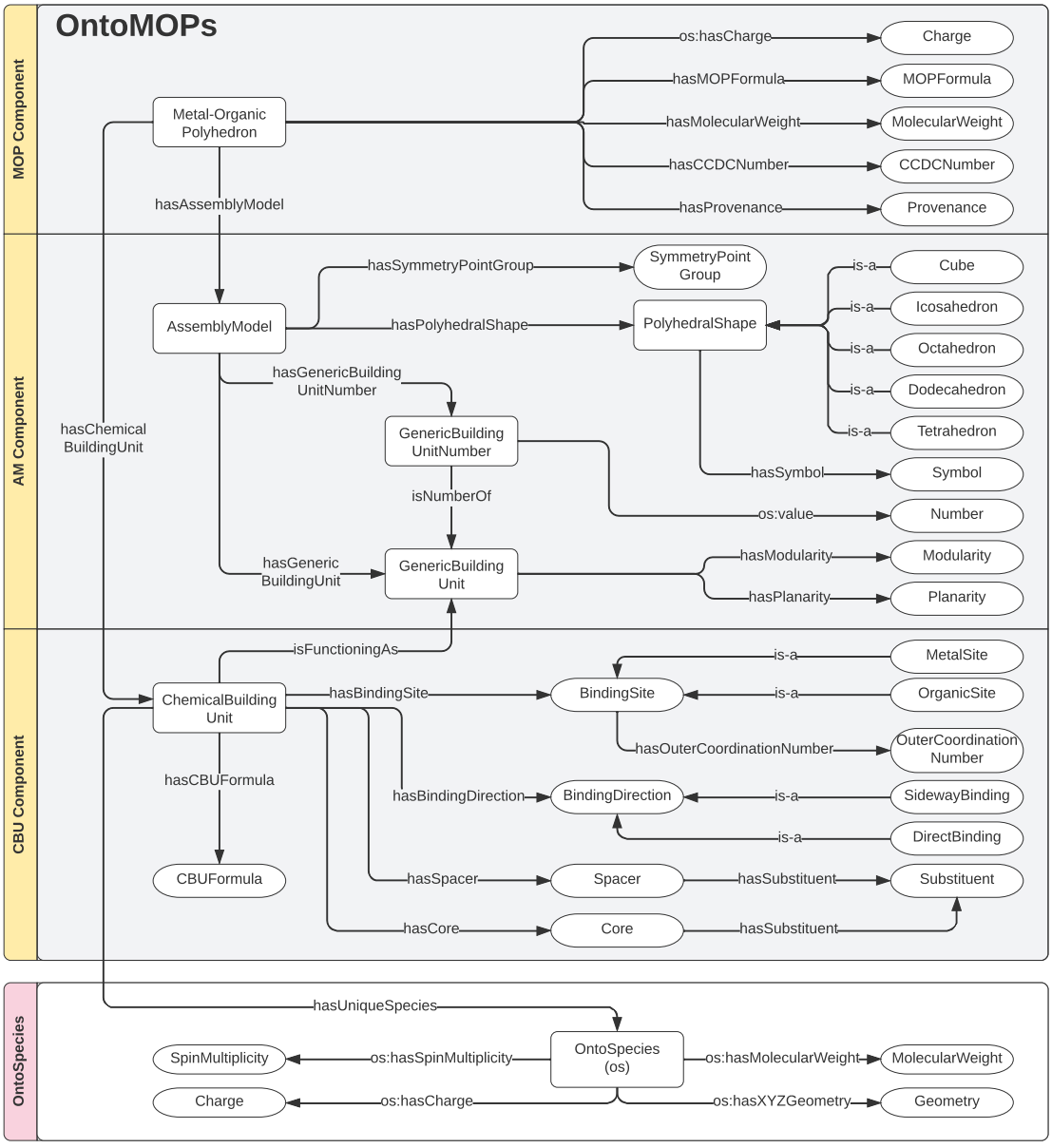


Figure 6: *Core concepts and properties of the OntoMOPs ontology.*

molecular viability and electronic insights and speed-up innovation [30, 32]. However, one has to acknowledge that MOPs, like with many POMs, represent relatively heavy molecules that are often computationally expensive for DFT approaches [41, 51]. Further on, differences in training and qualitative thinking [31] may also be present in the communication between synthetic MOP experts and computational chemists. Collaborative workflows where formulation proposals by synthetic experts are modelled and calculated by computational chemists remain low-throughput. At the same time, direct computational modelling without consideration of synthetically accessible building units can also lead to proposals that have little chance for experimental realisation. In this regard, our data collection and output are intended to close this existing gap in knowledge and communication.

When considering molecular modelling of heavy inorganic and hybrid molecules such as MOPs or POMs, typically, the structure of interest is modelled with only a simple approximation of the surrounding environment with a conductor like screening model [41, 51]. Analogous to MOF research, to start computations on existing MOPs, one would need computation-ready geometries [10]. To systematically model new MOPs, one needs geometries of building units and assembly models as templates for the rational design of MOP targets. Our data collection starts by consultation of two recently reported MOP

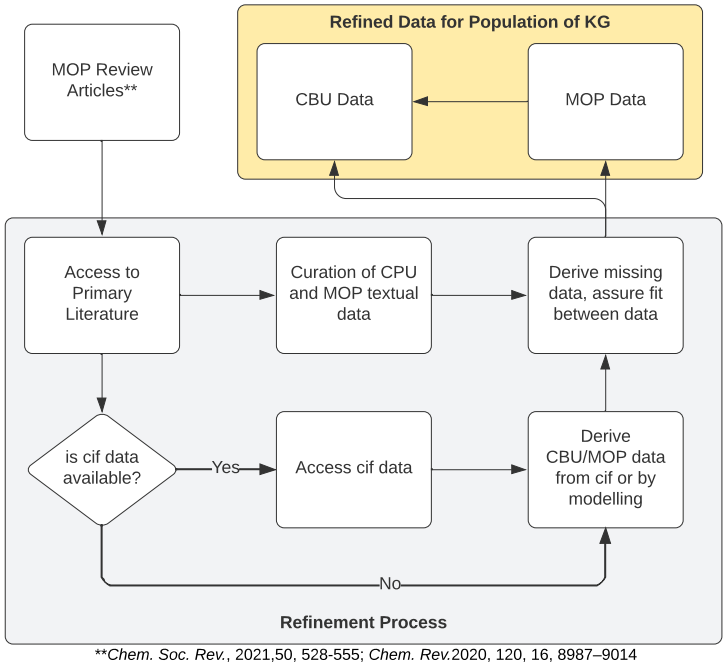


Figure 7: *Schematic representation of the different steps applied to derive and structure the MOP and CBU data.*

milestone reviews (see Figure 7) [26, 46]. These reviews also have a strong tutorial-like character, targeting predominantly synthetic and applied chemist readers. The review articles are thoroughly illustrated and provide sufficient visual aids in allocating information through the literature. However, at the same time, most of the presented information is

not practical for the direct extraction of data, but serves as a guiding overview of the primary literature. Following this, we consulted the primary literature from which we obtained information on the CBUs, MOPs and MOPs' crystallographic information files. The crystallographic information files were further used to extract *xyz* structures for the MOPs and parse them to obtain the *xyz* coordinates of the constituent CBUs. This was done in a way where solvent units, and other co-crystallized molecules or labile units binding to the metal sites were manually removed.

For MOPs where the crystallographic structure was not reported or the structure showed some anomalies for direct *xyz* export (*e.g.* disorder, atoms missing *etc.*), we used the graphical user interface of the Amsterdam Modelling Suite (www.scm.com) software for structure modelling [68]. For most of the MOPs for which crystallographic structure was not reported, their structure could be derived from other previously known MOPs or through modelling of some peripheral organic substituents resulting from post-functionalisation. For addition of those organic functionalities and for the optimisation of the organic CBUs, the universal force field was used [63, 68]. In this way, the geometries of 151 MOPs and 137 CBUs suitable for further DFT calculations (*i.e.* computation-ready) were obtained. The preparation of the working geometries was also a useful strategy that allowed us to cross-check the simplified MOP and CBU formulas and also to ensure that additional data based on the CBU geometries (*i.e.* molecular mass and InChI) is cleanly and correctly calculated.

The two review articles [26, 46] provide insights into the MOP construction based on the shape construct [58, 70]. However, the overall charge of individual MOPs is not mentioned. With consideration that MOP and CBU structures may undergo DFT calculations in future, we manually derived the overall charge for some of the structures. Considering that many building blocks are metal-based, the charge also may affect their spin multiplicity. Although molecular magnetism is not part of our current KE studies, for data completeness, we systematically assigned the maximum possible spin multiplicity to all non-diamagnetic CBUs (*i.e.* approximating all spin-up). The topic of magnetism is not systematically discussed in the literature [26, 46], although we acknowledge that many different magnetic scenarios may be possible.

4.4 Population of the KG

The data on MOPs and their chemical building units collected from the literature is stored in two CSV files (see SI). These are then instantiated in OntoMOPs using an input agent consisting of a collection of written python scripts, which take the data from the CSV files and process them to produce JSON and then OWL files which are then stored in the knowledge graph. This process results in each unique MOP being its own instance in OntoMOPs, with each chemical building unit also being a unique instance in OntoSpecies.

The developed software is freely accessible online through this link.

5 Prediction of new MOPs structures

5.1 Algorithms and Implementation

If one attempts to assemble a MOP directly by allocating chemically complementary CBUs to the corresponding GBUs of its particular assembly model, there is a high risk that irrational MOP structures will be proposed. The reason is that in this approach it is difficult to account for differences in dihedrals. An alternative strategy is to first locate all possible MOPs for a given AM. The next step is to derive the associated CBUs of those MOPs. Finally, the CBUs can be separated into "sets" based on their GBU characteristics. Using the AM as a template, MOPs can be combinatorially constructed by finding chemically complementary CBUs from these two sets. Some of the constructed MOPs will correspond to instances already present in TWA, while other will be completely new (Figure 13 in SI). However, this approach is highly restrictive and thus, if a small number of MOPs are represented by a certain AM (*i.e.* low versatility), the number of new structures that can be derived will be also highly limited. To derive a higher versatility of new rationally constructed MOP structures, one has to expand the CBU basis beyond just a single AM. To able to achieve the latter without compromising the accuracy of the rational construction, the original set of CBUs is updated with CBUs from other sets for other assembly models with which it has a CBU instance in common (Figure 13 in SI).

In this line, we developed two algorithmic approaches. Algorithm 1, represents the direct application of the AMs method and thus restricts the construction of MOPs without CBU share between sets corresponding to different AMs. When applying Algorithm 1, the sets populated with many MOPs are expected to have many different CBUs and thus project a higher potential for new instantiation. In Algorithm 2, exchanges between sets are allowed, providing an opportunity for an increase in the number of MOPs with assembly models that were originally sparsely populated.

Algorithm 1: MOP assembly – Method I.

Input: KG representation of MOPs, including CBU-GBU relationships and AMs.
Output: Candidate MOPs not currently represented in TWA.

begin
  Query the set A of all AMs from the KG.
  for $a_i \in A$ **do**
    Query the set P_i of MOPs in TWA with a_i as AM.
    Query the set G_i of all GBUs belonging to a_i from the KG.
    for $g_j \in G_i$ **do**
      Query the set C_j of all CBUs that function as GBU g_j in any MOP in P_i from the KG.
    end
    Form a candidate set $\hat{P}_i$ of MOPs by enumerating all possible combinations of CBUs that do not already occur in TWA: $\hat{P}_i := \{p \notin P_i | \forall_j c_j(p) \in C_j\}$, where $c_j(p)$ is a CBU of a MOP p that functions as the j^{th} GBU used in the definition of C_j.
  end
  return $\bigcup_i \hat{P}_i$.
end

Algorithm 2: MOP assembly – Method II.

Input: KG representation of MOPs, including CBU-GBU relationships and AMs.
Output: Candidate MOPs not currently represented in TWA.

begin
  Query the set A of all AMs from the KG.
  for $a_i \in A$ **do**
    Query the set P_i of MOPs in TWA with a_i as AM.
    Query the set G_i of all GBUs belonging to a_i from the KG.
    for $g_j \in G_i$ **do**
      Query the set $C_{j,i}$ of all CBUs that function as GBU g_j in any MOP in P_i from the KG.
    end
  end
  for $a_i \in A$ **do**
    for $g_j \in G_i$ **do**
      $C_j \leftarrow C_{j,i}$
      for $a_k \in A \setminus \{a_i\}$ **do**
        if $C_{j,i} \cap C_{j,k} \neq \emptyset$ **then**
          $C_j \leftarrow C_j \cup C_{j,k}$
        end
      end
    end
    Form a candidate set $\hat{P}_i$ of MOPs by enumerating all possible combinations of CBUs that do not already occur in TWA: $\hat{P}_i := \{p \notin P_i | \forall_j c_j(p) \in C_j\}$, where $c_j(p)$ is a CBU of a MOP p that functions as the j^{th} GBU used in the definition of C_j.
  end
  return $\bigcup_i \hat{P}_i$.
end

5.2 Algorithmic Output

In the OntoMOPs KG there are 18 different AMs (see Figure 8). All AMs are based on two different types of GBUs. The smallest AM is built using 5 GBUs and it is the diadic (3-pyramidal)$_2$(2-bent)$_3$ with D_{3h} symmetry point group. The largest AM is built using 42 GBUs and it is the (5-pyramidal)$_{12}$(2-linear)$_{30}$ with I_h symmetry point group. The remaining AMs span the range between these two extremes. All 18 AMs consist of pairs of seven different GBUs, namely 2-linear/bent 3-/4-/5-pyramidal and 3-/4-planar. The 5-planar CBUs are rare in chemistry (probably due to unusual coordination and strain), and thus the 5-planar GBU is not found among the GBUs currently in TWA. This implies that certain AMs such as the formally derived (5-planar)$_{12}$(2-bent)$_{30}$ have not been "discovered" among MOPs yet (Figure 4). However, other AMs reminiscent of Archimedean, Catalan, and Johnson solids are present in the TWA. In addition, non-polyhedral AMs such as a polygon, a prism, and a diad AM are also present in TWA. The latter three AMs may appear as "outliers". However, they are purposely present as their associated CBUs participate in the construction of other MOPs with different AMs. All AMs adopt one of the five symmetries T_d, O_h, I_h, C_s, D_{3h} and T_h. There are also two pairs of isomeric AMs, namely the(anti)cuboctahedral (4-planar)$_{12}$(2-bent)$_{20}$, and the cuboidal (3-pyramidal)$_8$(2-bent)$_{12}$ where the isomerism originates from the configurational orientation of the 2-bent GBUs. The cuboidal (3-pyramidal)$_8$(2-linear)$_{12}$ is absent from TWA, as well as the icosahedral (3-pyramidal)$_2$(2-linear)$_{30}$. The reason is that, to the best of our knowledge, there is an absence of reported inorganic CBUs that can exhibit the wide angles suitable for the construction of those AMs.

In OntoMOPs there are seven general GBUs. If placed as nodes on a graph, the general

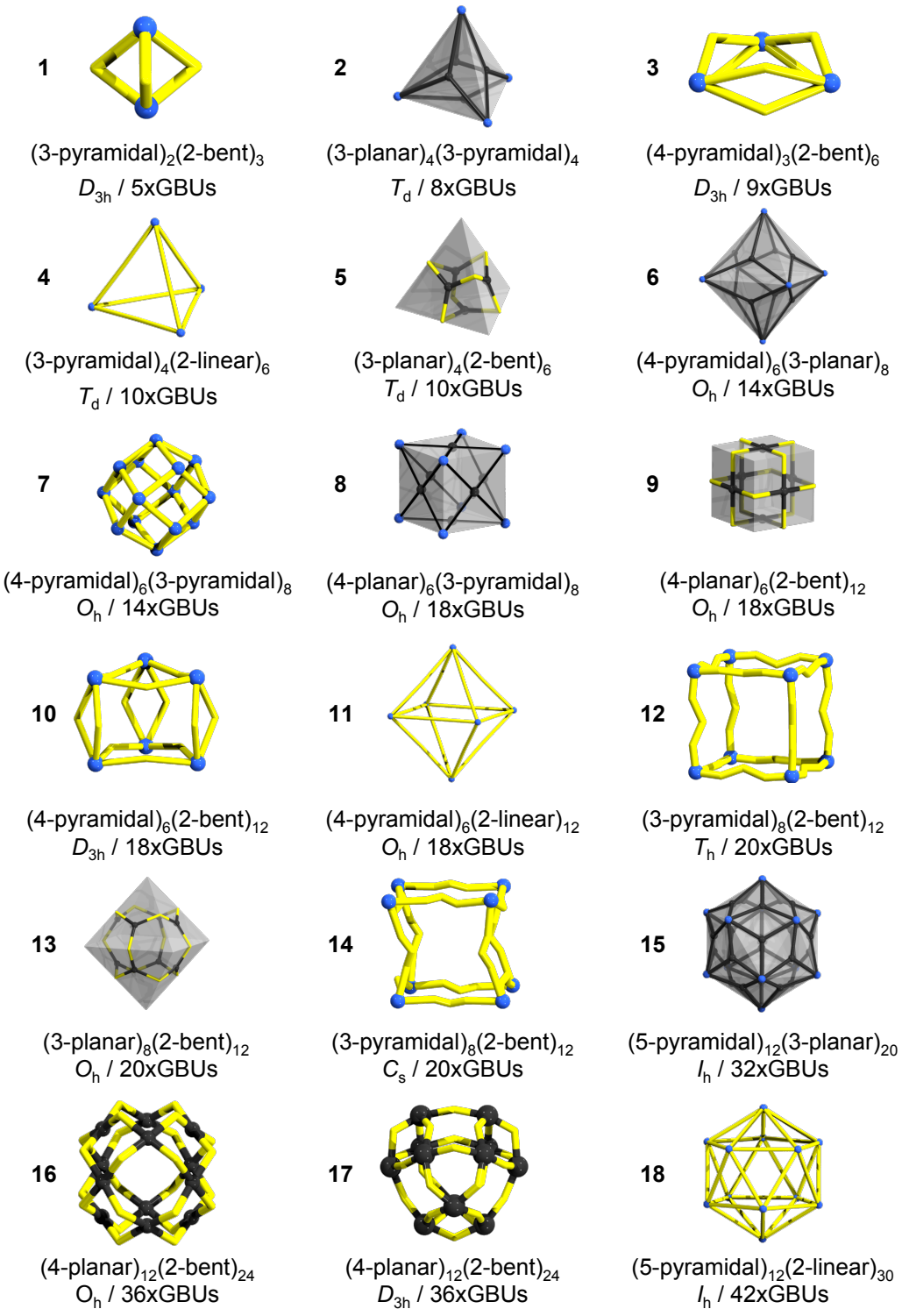


Figure 8: *Assembly Models present in the OntoMOPs cage, representing the construction principles of 151 reported MOP instances.*

GBUs are interconnected via 18 assembly models (see Figure 9.a). From the GBU nodes, the most interconnected is the one referring to the 2-bent unit, which as discussed earlier (see section 3.4) may be represented by CBUs with different dihedral angles. Therefore further differentiation between 2-bent GBUs is crucial. One of the discoveries of Algorithm 2 is that there are in total 37 related sets that have at least one CBU in common and thus they can exchange CBUs. One of the most interconnected sets is the one referring

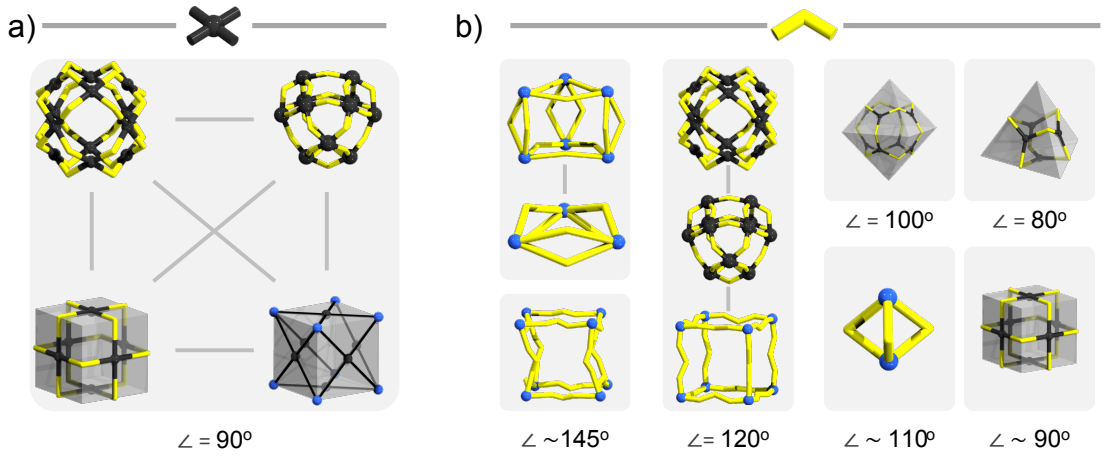


Figure 9: *Highly interrelated sets relating to different AMs containing a) 4-planar CBUs; b) 2-bent CBUs.*

to 4-planar GBUs (see Figure 9.b). This is the case because from a coordination chemistry viewpoint, most transition-metal based complexes can function as 4-planar CBUs, and thus there is no strong dihedral differentiation. However, in the case of the 2-bent GBU, our algorithm has found some common ligands between some sets, while other sets of 2-bent ligands have not been altered (Figure 9.c). This implies that even without hard-coding, the algorithm can successfully deduce that certain differences in dihedrals are acceptable when exchanging CBUs, but not all.

In order to have a perspective on the obtained number of instances from the application of the algorithms, one may considering a rough estimation of the exploratory chemical space. The exploratory chemical space associated with high-throughput synthetic explorations and such space may emerge my multiplying the combinations to be studied across number of changed parameters. If 91-organic and 46-inorganic CBUs are reacted across 18 different scenarios, then the total exploratory space would be 75348 unique chemical environments. In stark contrast to the exploratory space, algorithms 1 and 2 project an immediate chemical space of 506 and 1418 constructible MOP instances respectively (see Figure 10). This implies that the algorithms can effectively narrow down exploratory spaces and thus make automated synthetic explorations more focused. In comparison to the MOP instances currently present in TWA, where the $(4\text{-planar})_{12}(2\text{-bent})_{24}$ (O_h) archetype counts for approximately 37% of all structures, Algorithm 1 projects that, assembly model $(4\text{-planar})_{12}(2\text{-bent})_{24}$ (O_h) accounts for approximately 66% of the newly derived structures. The reason for this is that there can be many combinations between metal nodes (e.g. $[Pd_2]$, $[Cu_2]$, $[Rh_2]$, etc.) and other 2-bent organic CBUs in this AM. By contrast, in algorithm 2, it is deduced that MOPs represented by the anticuboctahe-

dral derivative of (4-planar)$_{12}$(2-bent)$_{24}$ (O_h) (i.e. (4-planar)$_{12}$(2-bent)$_{24}$ (D_{3h})) can also be constructed in large numbers. As the anticuboctahedral derivative appears to find suitable CBUs in the (3-pyramidal)$_8$(2-bent)$_{12}$ (T_h) set, the number of new predicted anticuboctahedral MOPs amounts to 397, the largest number for any of the AMs. However, this could change if additional MOPs instances that have CBUs that connect previously unconnected AMs are introduced into the KG.

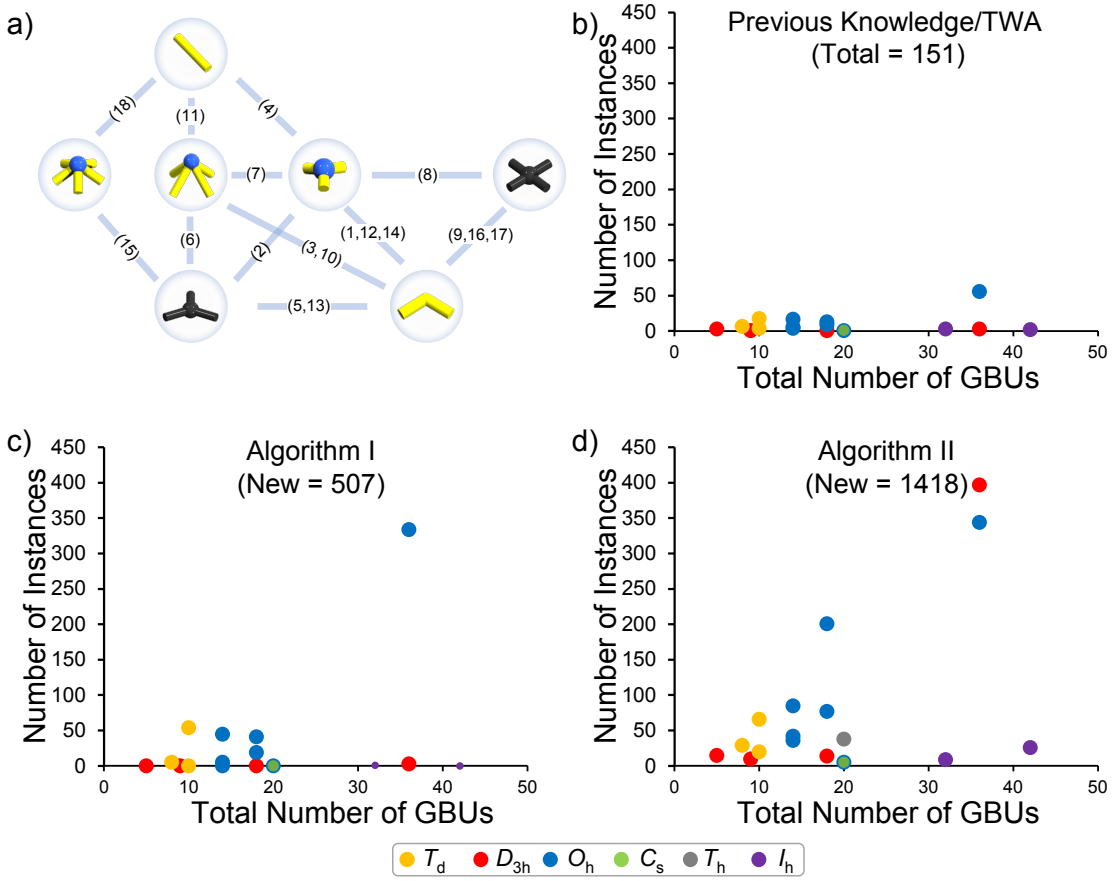


Figure 10: *a) Graph depicting the GBUs and the Assembly Models as nodes and links respectively. Number of MOP instances as a function of the total GBU sum present in TWA (b) and obtained following Algorithm 1 (c) and Algorithm 2 (d).*

Our algorithmic implementation allows us to query the molecular mass of the CBUs, and using the respective GBU numbers associated with the respective AM, one can derive the mass of the new MOPs. The molecular mass between most of the MOP instances differs except for the cases when isomers can be constructed. A histogram projection allows convenient analysis of the mass distributions in separate ranges of 1 kDa. Most of the starting MOP structures found in the literature show distribution maxima at 4 and 6 kDa with an overall median at 6584.55 g·mol^{-1}. In comparison, the new MOPs derived using algorithm I and algorithm II show maxima at 7 and 8 kDa, and median molecular mass values of 7586.83 g·mol^{-1} and 7875.685 g·mol^{-1} respectively. The shift in median is due to the fact that the newly derived MOP sets are predominantly represented by

MOPs that associate with (anti)cuboctahedral AMs employing 36 GBUs. In addition, when turning from reported to algorithmically derived MOPs one also observes a rise in the number of very heavy MOP structures, which are those that span the region of 23-26 kDa (Figure 11.a). The reasons for this rise are that there are new (anti)cuboctahedral MOP constructions that employ heavy organic CBUs (*e.g.* those with long alkyl chains) as well the general rise with of MOPs employing heavy POM-based inorganic nodes. The

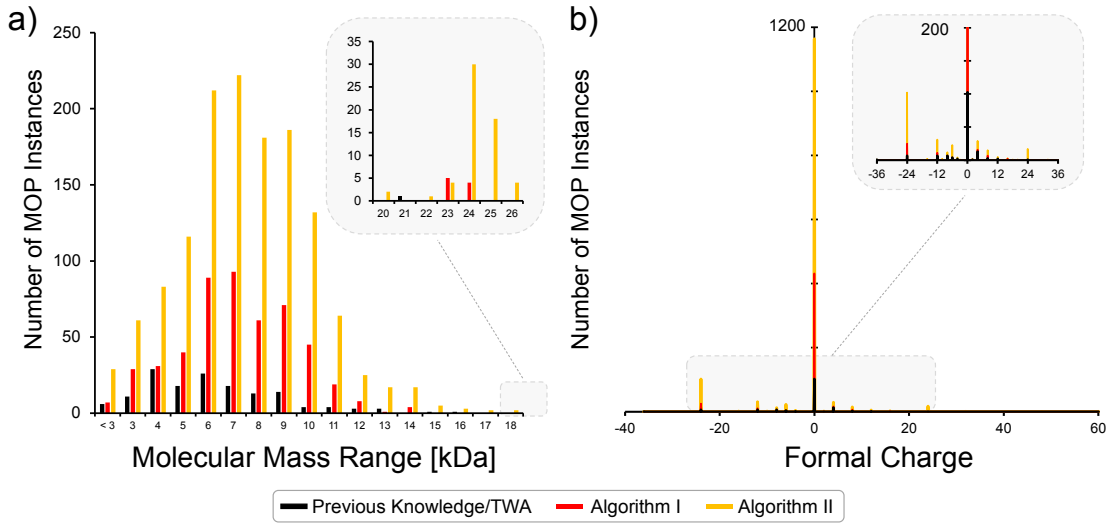


Figure 11: *Distribution of reported MOPs instances and the newly algorithmically derived MOPs instances as a function of a) their molecular mass ranges; b) their overall charge.*

overall MOP charge is highly relevant when devising new porous ionic solid combinations that rely on both positively and negatively charged MOPs. However, one in general needs to be careful with this interpretation as charged MOPs may be able to co-exist in a set of different charge states. The different charge states may be associated with different oxidation numbers of protonation states of the CBUs. Our algorithm is currently exploring the constructability problem, where the protonation and the oxidation state may be less relevant unless they block the binding site of the CBUs.

The distribution of the overall MOP charges show that most instances, from literature and those algorithmically derived are in the range of -36 up to +24 (Figure 11.b). To have a complete and saturated assembly model, the number of binding units from the organic and inorganic units should match. As the number of binding units typically "mirror" the magnitude of the absolute charge, the net charge outcome of the MOP ends up being neutral. Indeed some 64% of all MOP instances in the OntoMOPs KG are neutral. However, when there is a deviation from this scenario, the overall MOP structure may appear as charged. For instance, positively charged MOPs result from the use of neutral organic linkers (*e.g.* $[C_6H_4(C_3H_2N_2)_2]$) and positively charged inorganic CBUs. On the other hand, negatively charged MOPs typically derive from the combination of highly negative POM based CBUs (e.g. $[PW_9O_{37}Ni_6NH_2C_4H_3]$) and negatively charged carboxylate ligands, or use of 4-pyramidal organic ligands (*e.g.* $[(C_6HO_3)_4(C_4H_8)_4]_6)]$ and low charged metal cations (e.g. $[M_3]^{6+}$). Although not fully arbitrary, negative charges may derive

from the use of benzene-1,3,5-tricarboxylate ligand (*i.e.* BTC = $[(C_6H_3)(CO_2)_3]^{2-}$) as 2-bent units. The BTC is well known as a 3-planar organic CBU. When employed as 2-bent CBU, one site remains unsaturated, making the structures interesting in post-synthetic functionalisation [2]. When modelling, one may consider a scenario where the free carboxylate binding site is protonated, de-protonated or combination of both. As we were interested in obtaining the maximum outcome on constructable MOPs, BTC was considered to be a deprotonated CBU.

As mentioned earlier, the data curation has been based on information presented in the two most recent and most influential review articles, both covering reported MOPs until mid-2020 [26, 46]. By not adding newly reported MOP instances after that period, one can observe if the algorithm predicts instances that experts would also envision and attempt to prepare. In this line, one general trend is to substitute a smaller with a larger orgaic unit. Considering that the octahedral MOP $[V_5O_9]_6[(C_6H_3)(CO_2)_3]_8^{6-}$ is present in TWA [79], the algorithm has derived a new larger structure with formula $[V_5O_9]_6[L]_8^{6-}$ where L = $[(C_3N_3)(C_6H_4)_3(CO_2)_3]$, $[(C_6H_3)(C_6H_4)_3(CO_2)_3]$, $[(C_6H_3)(C_2C_6H_4)_3(CO_2)_3]$, and $[(C_6H_3)((C_6H_4)_2)_3(CO_2)_3]$. Among the different ligands, the use of of 1,3,5-tris(4-carboxyphenyl)-benzene to form has been reported by Su group in August 2020 [22]. The obtained structure was not covered in the review articles, however, its prediction suggest that our algorithm to significant level can replicate rational designs of experts (see Figure 12.a). $[V_5O_9]_6[(C_6H_3)(C_6H_4)_3(CO_2)_3]_8^{6-}$ has been reported by Su group in August 2020 [22]. The latter structure was not covered in the review articles, however, its prediction suggest that the algorithm to significant level can replicate the rational design by experts (see Figure 12.a). Considering the icosahedral $[WV_5O_{11}]_{12}[C_6H_4(CO_2)_2]_{30}^{12-}$ as reported in [78], the algorithm proposed a derivative structure in which one hydrogen atom of the organic CBU is formally substituted by a halogen atom. One proposed formulation is $[WV_5O_{11}]_{12}[C_6H_3Br(CO_2)_2]_{30}^{12-}$. This structure would be the subject of rich configurational isomerism. Thus in addition to the presented model, many other structures may be derived.(see Figure 12.b).

6 Summary and Outlook

The classical concept of secondary building units has been an important concept over the past two decades leading to the development of MOPs, MOFs and COFs. In this work, we differentiated between the chemical and structural nature of the SBU, and derived a conceptual description of MOPs based on assembly models. The key concepts were then used to extend TWA with the OntoMOPs ontology connecting to existing concepts from OntoSpecies. The TWA was populated with MOP data which we curated from the literature and structured in a systematic way to facilitate its further use in the exploration of the immediate chemical space. Algorithms were constructed for the discovery of new MOPs that makes use of information in OntoMOPs. Based on the available 137 CBUs and 151 experimentally verified MOPs, this MOP Discovery agent rationally proposed 1418 new MOPs structures that were previously not recorded in the literature. The overall study also shows that semantically driven and instance-based approaches can function simply based on meta-rules. In such a system, "outliers" do not break the meta-rules, but only update

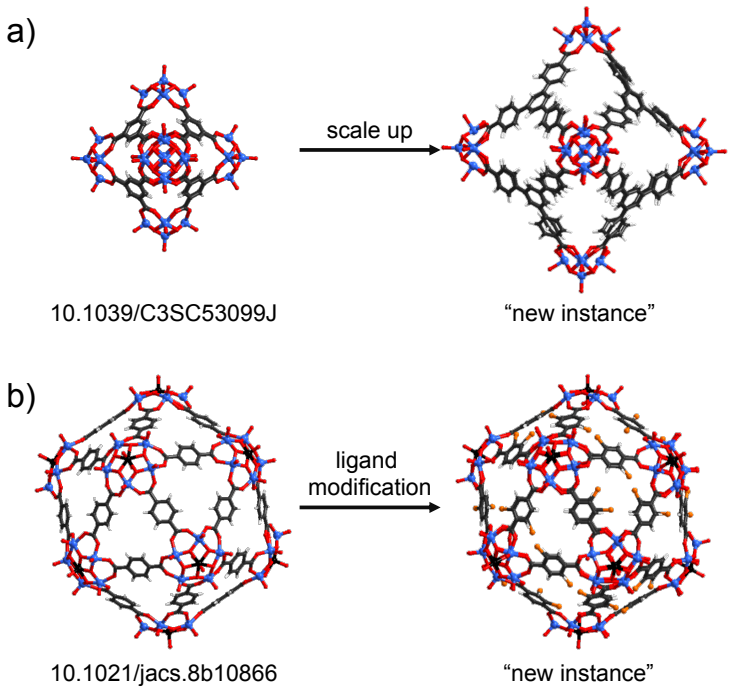


Figure 12: *Models of MOPs based on output from Algorithm 2: a) size increase based on utilisation of a spacer moieties; b) Br-substituted derivatives.*

the set of assembly "blueprints", thus the next iteration is more refined and potential uncertainties are predicted. Our approach can be combined with other developments such as Waller's algorithm that discovers chemical reactivity [64]. This can identify species that can potentially function as new CBUs and thus enable for more rapid exploration of the deep (i.e. uncharted) chemical space of MOPs in conjunction with existing data in our knowledge graph.

The semantically-based, ontology-driven discover algorithms successfully undertook rational structural proposals for MOPs, and we are currently extending this approach to related polyhedral and reticular materials. Using natural language processing for chemistry, our group has currently developed the "Marie" platform [80] that is able to interact with chemists and provide feedback. It is planned to extend Marie to make complex queries for MOPs and other reticular and polyhedral materials possible. This will make it more natural for MOP chemists to interact with The World Avatar, with the aim to improve the quality and quantity of data in TWA,which will in turn allow for increased potential of new discoveries in the MOPs field.

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. The authors are grateful to the UK Engineering and Physical

Sciences Research Council (EPSRC, grant number: EP/R029369/1) and ARCHER for financial and computational support as a part of their funding to the UK Consortium on Turbulent Reacting Flows (www.ukctrf.com). AK and MK thanks the Humboldt Foundation (Berlin, Germany) and the Isaac Newtwon Fund (Cambridge, UK) for Feodor Lynen Fellowship. Mr. Jiaru Bai (University of Cambridge) is thanked for discussions and feedback.

Nomenclature

ABox Assertional Component (of an ontology)

AI Artificial Intelligence

AM Assembly Model

CBU Chemical Building Unit

COF Covalent Organic Framework

GBU Generic Building Unit

KG Knowledge Graph

MOF Metal-Organic Framework

MOP Metal-Organic Polyhedron

POM Polyoxometalate

TBox Terminological Component (of an ontology)

TWA The World Avatar

A Supporting Information

Supporting Files:

A.1 Algorithmic Output

Table 1: *Number of MOP Formula models derived using algorithm I and algorithm II in comparison with the MOP formulas in the KG*

Nr.	Assembly Model	In KG	Algorithm I	Algorithm II
1	(3-pyramidal)$_2$(2-bent)$_3$ (D_{3h})	3	0	15
2	(3-planar)$_4$(3-pyramidal)$_4$ (T_d)	7	5	29
3	(4-pyramidal)$_3$(2-bent)$_6$ (D_{3h})	1	0	10
4	(3-pyramidal)$_4$(2-linear)$_6$ (T_d)	18	54	66
5	(3-planar)$_4$(2-bent)$_6$ (T_d)	4	0	20
6	(4-pyramidal)$_6$(3-planar)$_8$ (O_h)	17	45	85
7	(4-pyramidal)$_6$(3-pyramidal)$_8$ (O_h)	6	5	42
8	(4-planar)$_6$(3-pyramidal)$_8$ (O_h)	4	0	36
9	(4-planar)$_6$(2-bent)$_{12}$ (O_h)	13	41	77
10	(4-pyramidal)$_6$(2-bent)$_{12}$ (D_{3h})	1	0	14
11	(4-pyramidal)$_6$(2-linear)$_{12}$ (O_h)	9	19	201
12	(3-pyramidal)$_8$(2-bent)$_{12}$ (T_h)	1	0	38
13	(3-planar)$_8$(2-bent)$_{12}$ (O_h)	1	0	5
14	(3-pyramidal)$_8$(2-bent)$_{12}$ (C_s)	1	0	5
15	(5-pyramidal)$_1$2(3-planar)$_{20}$ (I_h)	3	1	9
16	(4-planar)$_1$2(2-bent)$_{24}$ (O_h)	57	333	343
17	(4-planar)$_1$2(2-bent)$_{24}$ (D_{3h})	3	3	397
18	(5-pyramidal)$_1$2(2-linear)$_{30}$ (I_h)	2	0	26

A.2 OntoMOPs

Classes

$$
\begin{aligned}
\text{MolecularCage} &\sqsubseteq \top \\
\text{CoordinationCage} &\sqsubseteq \text{MolecularCage} \\
\text{MetalOrganicPolyhedron} &\sqsubseteq \text{CoordinationCage} \\
\text{AssemblyModel} &\sqsubseteq \top \\
\text{GenericBuildingUnit} &\sqsubseteq \top \\
\text{GenericBuildingUnitNumber} &\sqsubseteq \top \\
\text{ChemicalBuildingUnit} &\sqsubseteq \top \\
\text{Cavity} &\sqsubseteq \top \\
\text{BindingDirection} &\sqsubseteq \top \\
\text{SidewayBinding} &\sqsubseteq \text{BindingDirection} \\
\text{DirectBinding} &\sqsubseteq \text{BindingDirection} \\
\text{BindingSite} &\sqsubseteq \top \\
\text{OrganicSite} &\sqsubseteq \text{BindingSite} \\
\text{MetalSite} &\sqsubseteq \text{BindingSite} \\
\text{Spacer} &\sqsubseteq \top \\
\text{Provenance} &\sqsubseteq \top \\
\text{PolyhedralShape} &\sqsubseteq \top \\
\text{Tetrahedron} &\sqsubseteq \text{PolyhedralShape} \\
\text{Cube} &\sqsubseteq \text{PolyhedralShape} \\
\text{Octahedron} &\sqsubseteq \text{PolyhedralShape} \\
\text{Icosahedron} &\sqsubseteq \text{PolyhedralShape} \\
\text{Dodecahedron} &\sqsubseteq \text{PolyhedralShape} \\
\text{Cuboctahedron} &\sqsubseteq \text{PolyhedralShape} \\
\text{Rhombicuboctahedron} &\sqsubseteq \text{PolyhedralShape} \\
\text{RhombicDodecahedron} &\sqsubseteq \text{PolyhedralShape} \\
\text{Species} &\sqsubseteq \top \\
\text{Charge} &\sqsubseteq \top \\
\text{MolecularWeight} &\sqsubseteq \top
\end{aligned}
$$

Object Properties

$$\begin{aligned}
\text{MetalOrganicPolyhedron} &\sqsubseteq \leq 1\ \text{hasChemicalBuildingUnit.ChemicalBuildingUnit} \sqcap \\
&\quad \geq 1\ \text{hasChemicalBuildingUnit.ChemicalBuildingUnit} \\
\text{MetalOrganicPolyhedron} &\sqsubseteq \leq 1\ \text{hasAssemblyModel.AssemblyModel} \sqcap \\
&\quad \geq 1\ \text{hasAssemblyModel.AssemblyModel} \\
\text{MetalOrganicPolyhedron} &\sqsubseteq \leq 1\ \text{hasCavity.Cavity} \sqcap \\
&\quad \geq 1\ \text{hasCavity.Cavity} \\
\text{MetalOrganicPolyhedron} &\sqsubseteq \leq 1\ \text{hasProvenance.Provenance} \sqcap \\
&\quad \geq 1\ \text{hasProvenance.Provenance} \\
\text{Cavity} &\sqsubseteq \leq 1\ \text{hasMOPCavityVolume.Volume} \sqcap \\
&\quad \geq 1\ \text{hasMOPCavityVolume.Volume} \\
\text{AssemblyModel} &\sqsubseteq \leq 1\ \text{hasGenericBuildingUnit.GenericBuildingUnit} \sqcap \\
&\quad \geq 1\ \text{hasGenericBuildingUnit.GenericBuildingUnit} \\
\text{AssemblyModel} &\sqsubseteq \leq 1\ \text{hasGenericBuildingUnitNumber.GenericBuildingUnitNumber} \sqcap \\
&\quad \geq 1\ \text{hasGenericBuildingUnitNumber.GenericBuildingUnitNumber} \\
\text{AssemblyModel} &\sqsubseteq \leq 1\ \text{hasPolyhedralShape.PolyhedralShape} \sqcap \\
&\quad \geq 1\ \text{hasPolyhedralShape.PolyhedralShape} \\
\text{ChemicalBuildingUnit} &\sqsubseteq \leq 1\ \text{hasBindingDirection.BindingDirection} \sqcap \\
&\quad \geq 1\ \text{hasBindingDirection.BindingDirection} \\
\text{ChemicalBuildingUnit} &\sqsubseteq \leq 1\ \text{hasCore.Core} \sqcap \\
&\quad \geq 1\ \text{hasCore.Core} \\
\text{ChemicalBuildingUnit} &\sqsubseteq \leq 1\ \text{hasSpacer.Spacer} \sqcap \\
&\quad \geq 1\ \text{hasSpacer.Spacer} \\
\text{ChemicalBuildingUnit} &\sqsubseteq \leq 1\ \text{hasBindingSite.Binding Site} \sqcap \\
&\quad \geq 1\ \text{hasBindingSite.Binding Site} \\
\text{GenericBuildingUnitNumbert} &\sqsubseteq \leq 1\ \text{isNumberOf.GenericBuildingUnit} \sqcap \\
&\quad \geq 1\ \text{isNumberOf.GenericBuildingUnit} \\
\text{PolyhedralShape} &\sqsubseteq \leq 1\ \text{hasSymbol.Symbol} \sqcap \\
&\quad \geq 1\ \text{hasSymbol.Symbol} \\
\text{ChemicalBuildingUnit} &\sqsubseteq \leq 1\ \text{OS:hasUniqueSpecies.OS:Species} \sqcap \\
&\quad \geq 1\ \text{OS:hasUniqueSpecies.OS:Species} \\
\text{MetalOrganicPolyhedron} &\sqsubseteq \leq 1\ \text{OS:hasMolecularWeight.OS:MolecularWeight} \sqcap \\
&\quad \geq 1\ \text{OS:hasMolecularWeight.OS:MolecularWeight} \\
\text{MetalOrganicPolyhedron} &\sqsubseteq \leq 1\ \text{OS:hasCharge.OS:Charge} \sqcap \\
&\quad \geq 1\ \text{OS:hasCharge.OS:Charge}
\end{aligned}$$

Data Properties

$$\exists\ \text{hasCBUFormula}.\top \sqsubseteq \text{ChemicalBuildingUnit}$$
$$\top \sqsubseteq \forall\ \text{hasCBUFormula.String}$$
$$\exists\ \text{hasMOPFormula}.\top \sqsubseteq \text{MetalOrganicPolyhedron}$$
$$\top \sqsubseteq \forall\ \text{hasMOPFormula.String}$$
$$\exists\ \text{hasCCDCNumber}.\top \sqsubseteq \text{MetalOrganicPolyhedron}$$
$$\top \sqsubseteq \forall\ \text{hasCCDCNumber.Integer}$$
$$\exists\ \text{hasCCDCNumber}.\top \sqsubseteq \text{MetalOrganicPolyhedron}$$
$$\top \sqsubseteq \forall\ \text{hasCCDCNumber.Integer}$$
$$\exists\ \text{hasModularity}.\top \sqsubseteq \text{GenericBuildingUnit}$$
$$\top \sqsubseteq \forall\ \text{hasModularit.Integer}$$
$$\exists\ \text{hasXYZGeometry}.\top \sqsubseteq \text{MetalOrganicPolyhedron}$$
$$\top \sqsubseteq \forall\ \text{hasXYZGeometry.String}$$
$$\exists\ \text{hasPlanarity}.\top \sqsubseteq \text{GenericBuildingUnit}$$
$$\top \sqsubseteq \forall\ \text{hasPlanarity.String}$$
$$\exists\ \text{hasUnitNumberValue}.\top \sqsubseteq \text{GenericBuildingUnitNumber}$$
$$\top \sqsubseteq \forall\ \text{hasUnitNumberValue.Integer}$$
$$\exists\ \text{hasSymmetryPointGroup}.\top \sqsubseteq \text{AssemblyModel}$$
$$\top \sqsubseteq \forall\ \text{hasSymmetryPointGroup.String}$$
$$\exists\ \text{hasReferenceDOI}.\top \sqsubseteq \text{Provenance}$$
$$\top \sqsubseteq \forall\ \text{hasReferenceDOI.String}$$
$$\exists\ \text{hasSymbol}.\top \sqsubseteq \text{PolyhedralShape}$$
$$\top \sqsubseteq \forall\ \text{hasSymbol.String}$$
$$\exists\ \text{hasSymbol}.\top \sqsubseteq \text{PolyhedralShape}$$
$$\top \sqsubseteq \forall\ \text{hasSymbol.String}$$
$$\exists\ \text{hasOuterCoordinationNumber}.\top \sqsubseteq \text{BindingSite}$$
$$\top \sqsubseteq \forall\ \text{hasOuterCoordinationNumber.Integer}$$
$$\exists\ \text{hasSymbol}.\top \sqsubseteq \text{PolyhedralShape}$$
$$\top \sqsubseteq \forall\ \text{hasSymbol.String}$$
$$\exists\ \text{OS:value}.\top \sqsubseteq \text{AssemblyModel}$$
$$\top \sqsubseteq \forall\ \text{OS:value.String}$$

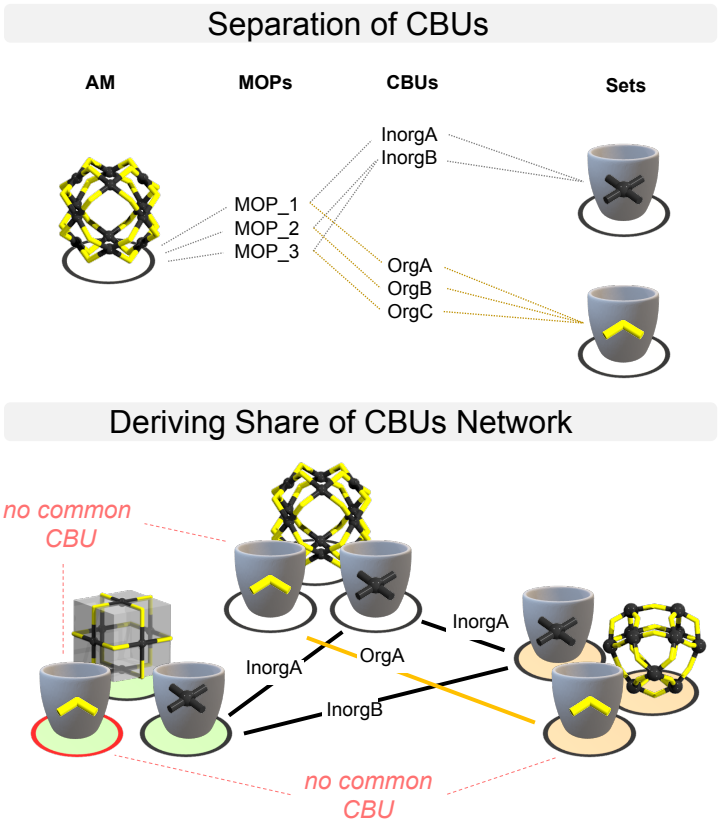


Figure 13: *Schematic representation of the creation of sets of CBUs associated with certain GBU and AM (top); creation of CBU share network between different sets of CBUs based on having CBUs in common. Illustration is only provided as an example OrgA, InorgA etc. represent organic and inorganic CBUs respectively. Between two sets of CBUs there will be no CBU common especially when stark differences in terms of dihedral angle exist (e.g.* 90° *vs* 120°*).*

References

[1] J. Akroyd, S. Mosbach, A. Bhave, and M. Kraft. Universal digital twin – a dynamic knowledge graph. *Data-centric Eng.*, 2:e14, 2021.

[2] J. Albalad, A. Carné-Sánchez, T. Grancha, L. Hernández-López, and D. Maspoch. Protection strategies for directionally-controlled synthesis of previously inaccessible metal–organic polyhedra (mops): the cases of carboxylate-and amino-functionalised rh (ii)-mops. *Chem. Commun.*, 55(85):12785–12788, 2019.

[3] F. H. Allen, S. Bellard, M. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B. Hummelink-Peters, O. Kennard, W. Motherwell, et al. The cambridge crystallographic data centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallogr. B*, 35(10):2331–2339, 1979.

[4] A. V. Anyushin, A. Kondinski, and T. N. Parac-Vogt. Hybrid polyoxometalates as post-functionalization platforms: from fundamentals to emerging applications. *Chem. Soc. Rev.*, 49(2):382–432, 2020.

[5] 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.

[6] 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*, pages 292–309, 2022.

[7] V. Balzani, A. Credi, F. M. Raymo, and J. F. Stoddart. Artificial molecular machines. *Angew. Chem. Int. Ed.*, 39(19):3348–3391, 2000.

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

[9] A. M. Castilla, W. J. Ramsay, and J. R. Nitschke. Stereochemistry in subcomponent self-assembly. *Acc. Chem. Res.*, 47(7):2063–2073, 2014.

[10] Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, et al. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. *J. Chem. Eng. Data*, 64(12):5985–5998, 2019.

[11] S. Cohen, M. Hershcovitch, M. Taraz, O. Kißig, A. Wood, D. Waddington, P. Chin, and T. Friedrich. Drug repurposing using link prediction on knowledge graphs with applications to non-volatile memory. In *Complex Networks & Their Applications X*, pages 742–753. Springer, 2021.

[12] A. U. Czaja, N. Trukhan, and U. Müller. Industrial applications of metal–organic frameworks. *Chem. Soc. Rev.*, 38(5):1284–1293, 2009.

[13] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, 36(suppl_1): D344–D350, 2007.

[14] K. E. Drexler. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U.S.A*, 78(9):5275–5278, 1981.

[15] I. Dumontheil. Development of abstract thinking during childhood and adolescence: The role of rostrolateral prefrontal cortex. *Dev. Cogn. Neurosci.*, 10:57–76, 2014.

[16] 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.

[17] A. Eibeck, A. Chadzynski, M. Q. Lim, K. Aditya, L. Ong, A. Devanand, G. Karmakar, S. Mosbach, R. Lau, I. A. Karimi, et al. A parallel world framework for scenario analysis in knowledge graphs. *Data-centric Eng.*, 1:e6, 2020.

[18] J. D. Evans, K. E. Jelfs, G. M. Day, and C. J. Doonan. Application of computational methods to the design and characterisation of porous molecular materials. *Chem. Soc. Rev.*, 46:3286–3301, 2017.

[19] 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.

[20] 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.

[21] M. Fernández-López, A. Gómez-Pérez, and N. Juristo. Methontology: from ontological art towards ontological engineering. *AAAI Press Technical Reports*, 1997.

[22] H. Gan, N. Xu, C. Qin, C. Sun, X. Wang, and Z. Su. Equi–size nesting of platonic and archimedean metal–organic polyhedra into a twin capsid. *Nat. Commun.*, 11(1): 1–8, 2020.

[23] F. Giunchiglia, B. Dutta, V. Maltese, and F. Farazi. A facet-based methodology for the construction of a large-scale geospatial ontology. *J. Data Semant*, 1(1):57–73, 2012.

[24] F. Giunchiglia, B. Dutta, and V. Maltese. From knowledge organization to knowledge representation. *Knowl. Organ.*, 41(1):44–56, 2014.

[25] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, and Z. Wang. Hermit: an owl 2 reasoner. *J. Automat. Reason.*, 53(3):245–269, 2014.

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

[27] A. J. Gosselin, A. M. Antonio, K. J. Korman, M. M. Deegan, G. P. Yap, and E. D. Bloch. Elaboration of porous salts. *J. Am. Chem. Soc.*, 143(37):14956–14961, 2021.

[28] M. Grüninger and M. S. Fox. Methodology for the design and evaluation of ontologies. *Int. Jt. Conf. Artif. Intell.*, 1995.

[29] K. Harris, Q.-F. Sun, S. Sato, and M. Fujita. $M_{12}L_{24}$ spheres with endo and exo coordination sites: scaffolds for non-covalent functionalization. *J. Am. Chem. Soc.*, 135(34):12497–12499, 2013.

[30] T. Heine. Grand challenges in computational materials science: from description to prediction at all scales. *Front. Mater.*, 1:7, 2014.

[31] R. Hoffmann. Qualitative thinking in the age of modern computational chemistry—or what lionel salem knows. *J. Mol. Struct. Theochem*, 424(1):1–6, 1998. ISSN 0166-1280. A Faithful Couple: Qualitative and Quantitative Understanding of Chemistry.

[32] R. Hoffmann, P. Schleyer, and H. Schaefer III. Predicting moleculesâ€"more realism, please! *Angew. Chem. Int. Ed.*, 47(38):7164–7167, 2008.

[33] N. Hosono and S. Kitagawa. Modular design of porous soft materials via self-organization of metal–organic cages. *Acc. Chem. Res.*, 51(10):2437–2446, 2018.

[34] O. Inderwildi and M. Kraft, editors. *Intelligent Decarbonisation*. Lecture Notes in Energy. Springer International Publishing, 1 edition, 2022.

[35] Y. Jiang, P. Tan, S.-C. Qi, C. Gu, S.-S. Peng, F. Wu, X.-Q. Liu, and L.-B. Sun. Breathing metal–organic polyhedra controlled by light for carbon dioxide capture and liberation. *CCS Chemistry*, 3(6):1659–1668, 2021.

[36] J. Jiao, C. Tan, Z. Li, Y. Liu, X. Han, and Y. Cui. Design and assembly of chiral coordination cages for asymmetric sequential reactions. *J. Am. Chem. Soc.*, 140(6): 2251–2259, 2018.

[37] J. Kido, M. Kimura, and K. Nagai. Multilayer white light-emitting organic electroluminescent device. *Science*, 267(5202):1332–1334, 1995.

[38] S. Kim, J. K. Lee, S. O. Kang, J. Ko, J.-H. Yum, S. Fantacci, F. De Angelis, D. Di Censo, M. K. Nazeeruddin, and M. Grätzel. Molecular engineering of organic sensitizers for solar cell applications. *J. Am. Chem. Soc.*, 128(51):16701–16707, 2006.

[39] H. Kitano. Nobel turing challenge: creating the engine for scientific discovery. *NPJ Syst. Biol. Appl.*, 7(1):1–12, 2021.

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

[41] A. Kondinski. Computational modelling of isomeric polyoxometalates. In *Chemical Modelling*, pages 39–71. RSC, 2021.

[42] 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.

[43] 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.

[44] M. Kraft and A. Eibeck. J-park simulator: Knowledge graph for industry 4.0. *Chem. Ing. Tech.*, 92(7):967–977, 2020.

[45] 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.

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

[47] J.-M. Lehn. Supramolecular chemistry. *Science*, 260(5115):1762–1764, 1993.

[48] H. Li, M. Eddaoudi, M. O'Keeffe, and O. M. Yaghi. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature*, 402(6759): 276–279, 1999.

[49] X.-X. Li, D. Zhao, and S.-T. Zheng. Recent advances in pom-organic frameworks and pom-organic polyhedra. *Coord. Chem. Rev*, 397:220–240, 2019.

[50] J. R. Long and O. M. Yaghi. The pervasive chemistry of metal–organic frameworks. *Chem. Soc. Rev.*, 38(5):1213–1214, 2009.

[51] X. López, J. J. Carbó, C. Bo, and J. M. Poblet. Structure, properties and reactivity of polyoxometalates: a theoretical perspective. *Chem. Soc. Rev.*, 41(22):7537–7571, 2012.

[52] H. Lyu, Z. Ji, S. Wuttke, and O. M. Yaghi. Digital reticular chemistry. *Chem*, 6(9): 2219–2241, 2020.

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

[54] 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.

[55] B. Motik, R. Shearer, and I. Horrocks. Hypertableau reasoning for description logics. *J. Artif. Intell. Res.*, 36:165–228, 2009.

[56] A. Müller, E. Krickemeyer, H. Bögge, M. Schmidtmann, and F. Peters. Organizational forms of matter: an inorganic super fullerene and keplerate based on molybdenum oxide. *Angew. Chem. Int. Ed.*, 37(24):3359–3363, 1998.

[57] N. F. Noy, D. L. McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

[58] M. O'Keeffe, M. A. Peskov, S. J. Ramsden, and O. M. Yaghi. The reticular chemistry structure resource (rcsr) database of, and symbols for, crystal nets. *Acc. Chem. Res.*, 41(12):1782–1789, 2008.

[59] F. Piccinno, R. Hischier, S. Seeger, and C. Som. From laboratory to industrial scale: a scale-up framework for chemical processes in life cycle assessment studies. *J. Clean. Prod*, 135:1085–1097, 2016.

[60] B. Pilgrim and N. R. Champness. Metal-organic frameworks and metal-organic cages–a perspective. *ChemPlusChem*, 85(8):1842–1856, 2020.

[61] H. S. Pinto and J. P. Martins. Ontologies: How can they be built? *Knowledge and information systems*, 6(4):441–464, 2004.

[62] D. A. Poole, E. O. Bobylev, S. Mathew, and J. N. Reek. Topological prediction of palladium coordination cages. *Chem. Sci.*, 11(45):12350–12357, 2020.

[63] A. K. Rappé, C. J. Casewit, K. Colwell, W. A. Goddard III, and W. M. Skiff. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.*, 114(25):10024–10035, 1992.

[64] M. H. Segler and M. P. Waller. Modelling chemical reasoning to predict and invent reactions. *Chem. Eur. J.*, 23(25):6118–6128, 2017.

[65] J. W. Steed and J. L. Atwood. *Supramolecular chemistry*. John Wiley & Sons, 2022.

[66] A. C. Sudik, A. R. Millward, N. W. Ockwig, A. P. Côté, J. Kim, and O. M. Yaghi. Design, synthesis, structure, and gas (N_2, Ar, CO_2, CH_4, and H_2) sorption properties of porous metal-organic tetrahedral and heterocuboidal polyhedra. *J. Am. Chem. Soc.*, 127(19):7110–7118, 2005.

[67] C. Tan, J. Jiao, Z. Li, Y. Liu, X. Han, and Y. Cui. Design and assembly of a chiral metallosalen-based octahedral coordination cage for supramolecular asymmetric catalysis. *Angew. Chem. Int. Ed.*, 57(8):2085–2090, 2018.

[68] G. t. Te Velde, F. M. Bickelhaupt, E. J. Baerends, C. Fonseca Guerra, S. J. van Gisbergen, J. G. Snijders, and T. Ziegler. Chemistry with adf. *J. Comput. Chem.*, 22 (9):931–967, 2001.

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

[70] D. J. Tranchemontagne, Z. Ni, M. O'Keeffe, and O. M. Yaghi. Reticular chemistry of metal–organic polyhedra. *Angew. Chem. Int. Ed.*, 47(28):5136–5147, 2008.

[71] D. J. Tranchemontagne, J. L. Mendoza-Cortés, M. O'Keeffe, and O. M. Yaghi. Secondary building units, nets and bonding in the chemistry of metal–organic frameworks. *Chem. Soc. Rev.*, 38(5):1257–1283, 2009.

[72] M. Uschold and M. King. *Towards a methodology for building ontologies*. Citeseer, 1995.

[73] A. Von Hippel. Molecular engineering. *Science*, 123(3191):315–317, 1956.

[74] G. M. Whitesides, J. P. Mathias, and C. T. Seto. Molecular self-assembly and nanochemistry: a chemical strategy for the synthesis of nanostructures. *Science*, 254(5036):1312–1319, 1991.

[75] 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, et al. The fair guiding principles for scientific data management and stewardship. *Sci. Data*, 3(1): 1–9, 2016.

[76] W.-H. Xing, H.-Y. Li, X.-Y. Dong, and S.-Q. Zang. Robust multifunctional Zr-based metal-organic polyhedra for high proton conductivity and selective CO_2 capture. *J. Mater. Chem. A*, 6(17):7724–7730, 2018.

[77] D. Zhang, T. K. Ronson, Y.-Q. Zou, and J. R. Nitschke. Metal–organic cages for molecular separations. *Nat. Rev. Chem.*, 5(3):168–182, 2021.

[78] Y. Zhang, H. Gan, C. Qin, X. Wang, Z. Su, and M. J. Zaworotko. Self-assembly of goldberg polyhedra from a concave $[WV_5O_{11}(RCO_2)_5(SO_4)]$ 3–building block with 5-fold symmetry. *J. Am. Chem. Soc.*, 140(50):17365–17368, 2018.

[79] Z. Zhang, L. Wojtas, and M. J. Zaworotko. Organic–inorganic hybrid polyhedra that can serve as supermolecular building blocks. *Chem. Sci.*, 5(3):927–931, 2014.

[80] 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.