Preprint   Cambridge Centre for Computational Chemical Engineering   ISSN 1473 – 4273

A chemical species ontology for data integration and knowledge discovery

Laura Pascazio[1], Simon Rihm[1,2], Ali Naseri[2], Sebastian Mosbach[1,2,3], Jethro Akroyd[1,2,3], Markus Kraft[1,2,3,4,5]

released: May 16, 2023

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

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

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

[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. 306

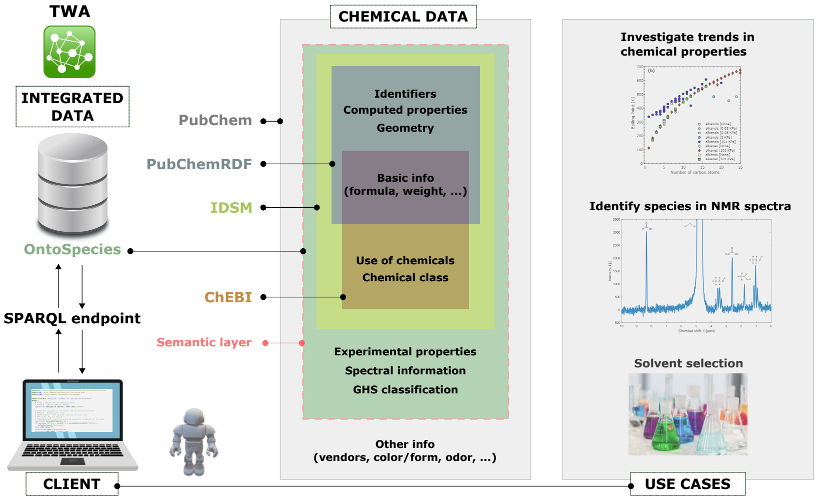
Keywords: General Chemistry Data Set, Ontology, Knowledge Graph, The World Avatar, OntoSpecies

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

Web ontologies are important tools in modern scientific research, as they provide a standardized way to represent and manage large amounts of complex data. In the chemistry field, the need for a comprehensive and reliable semantic database of chemical species is essential for accurate analysis and prediction of chemical behavior. This paper presents OntoSpecies, a web ontology designed to semantically represent chemical species and their properties. The ontology serves as a core component of the The World Avatar knowledge-graph chemistry domain and includes a wide range of identifiers, chemical and physical properties, chemical classifications and applications, as well as spectral information associated with each species. The ontology also includes provenance and attribution metadata, ensuring the reliability and traceability of the data. Most of the information about a chemical species is sourced from PubChem and ChEBI data on the respective compound webpages using a software agent, making OntoSpecies the most comprehensive semantic database on chemical species. Access to this reliable source of chemical data is provided through a SPARQL endpoint. The paper presents several use cases to demonstrate the usefulness of OntoSpecies in solving complex tasks that require information at a deep level of knowledge, making it an invaluable tool for scientific research. Overall, the approach presented in this paper is a significant advancement in the field of chemical data management, offering a powerful tool for representing, navigating and analyzing chemical information.



Highlights

- OntoSpecies ontology is developed for representing chemical species and their properties.
- OntoSpecies serves as core component of The World Avatar and it is linked to the existing ontologies in The World Avatar chemistry domain.
- A software agent is developed to dynamically collect data from PubChem and ChEBI.
- The ontological format permits advanced queries, and easy data analysis and visualization of chemical information.

Contents

1 Introduction

The role of data in chemical engineering has been steadily increasing over the years [23, 71]. Chemical engineers are now using data to gain deeper insights into complex systems and to develop more efficient and sustainable processes [19]. Data-based applications are becoming increasingly important in addressing the critical global challenges related to food, water, health, energy, and environment. With the increasing volume and complexity of chemical data, digital tools have facilitated the sharing and dissemination of scientific data and results, allowing for greater transparency and collaboration within the scientific community [33].

A large amount of data on chemical compounds has been continuously collected over time and is publicly available across many chemistry databases. One of the most comprehensive general public chemistry databases is PubChem [6, 38, 52]. It hosts information on more than 60 million unique chemical structures and it serves as a key chemical information resource for researchers in many biomedical science areas, including cheminformatics, chemical biology, and medicinal chemistry. PubChem is a data aggregator: it collects data from different data sources, including government agencies, university labs, pharmaceutical companies and substance vendors. For efficient use of this vast amount of data, PubChem provides various tools that fulfill criteria for simple and effective searching, where the chemical formula or name of the chemical species can be used as query string to obtain the information stored in the database [6, 52]. While these tools can sift through enormous amounts of chemical information, they are insufficient if the search needs to fulfill complex criteria or if new information needs to be derived from existing data. For example, the ability to find compounds with a defined set of desirable attributes would help researchers in tasks such as the selection of suitable solvents or the identification of species in a mixture.

Additionally, recent advances in machine learning illustrate very clearly why chemistry would benefit from data collected in an interoperable and reusable form: machines can read vast amounts of data, then autonomously discover the subsets that are most relevant and in turn make informed decisions. This requires machine to not only parse the data but also understand the data and its context [33]. The urgent need to make the data machine-actionable led to the design of a set of principles that we refer to as the FAIR data principles. FAIR data are data which meet principles of findability, accessibility, interoperability, and reusability (FAIR) [18, 70].

In this regard, semantic web ontologies aim to explicitly describe and relate objects using formal, logic-based representations that a machine can understand and process [5]. This will facilitate knowledge representation and integration that better meet the FAIR principles and will improve question answering in areas of critical importance to the life science [33]. This is done by present authors as part of The World Avatar (TWA) project, which aims to create an all-encompassing digital twin of the real world using knowledge graph (KG) technologies. A KG is a knowledge representation that uses a graph data structure to represent entities and their relationships. In TWA KG, ontologies are used to define the schema or structure of the KG, providing a consistent and formal way to model concepts and their relationships. An ontology to represent chemical species and their properties, OntoSpecies, was designed to serve as core component of TWA chemistry domain and to capture basic information about species, such as empirical formula, molecular weight and few other concepts [20]. However, in order to bridge the molecular-scale chemistry level to real world macro-scale phenomena and enable cross-domain applications, it is crucial to have a rich chemistry domain. Making PubChem data available to OntoSpecies would have a direct application in this regard, allowing powerful search capabilities that can be used for navigation and filtering of the large chemical space.

Integrating PubChem data in semantic database is not without precedent [21, 22]. However, due to difficulties on dealing with data from different sources and varying formats, the current semantic databases that integrate PubChem data do not include all the available information that can be accessed in the web-based PubChem resource. The information exported is currently mostly limited on identifiers and basic information. Properties like boiling point, melting point, density or solubility as well as spectral information on chemical species are currently not available in any semantic database. These information are crucial for applications in reaction modeling or lab automation.

The main objective of this work is to update and extend OntoSpecies ontology in order to represent additional information on chemical species, including concepts that are currently not exported in any other semantic database. We describe our effort to integrate data from major chemical databases such as PubChem and we illustrate the value of using semantic web technologies to seamlessly integrate and query diverse chemical knowledge in a manner that opens new avenues for knowledge discovery in chemical informatics.

The paper is structured as follows: First, an overview of the current state-of-the-art semantic representation of general chemistry and our approach as part of TWA knowledge ecosystem is presented in section 2. Next, we describe the OntoSpecies ontology, its linking mechanism to other domains in TWA, and the application of agent to dynamically populate the KG (sections 3 and 4). Finally, we discuss the utility and impact of our approach in the chemical informatics field as demonstrated by several use cases (section 5).

2 Background

In this section, a description of semantic web technologies is given and the current state-of-the art of the semantic representation of general chemistry is reviewed. Finally, an overview of TWA knowledge ecosystem is presented.

2.1 Semantic web

Since the landmark publication by Berners-Lee et al. [5], semantic web has been emerging as an increasingly important approach for better scientific data sharing and faster data processing using computers. Numerous high-tech companies such as Google, IBM, Microsoft, Facebook, and eBay have been developing commercial products using semantic web technologies [53]. In the context of the pharmaceutical industry, AstraZeneca leads the way on the application of semantic knowledge as a means to drug discovery [26]. Semantic web technologies and standards include the trio of the Resource Description Framework (RDF) [65], Web Ontology Language (OWL) [64], and SPARQL Protocol and RDF Query Language (SPARQL) [67]. RDF is a standard model that uses machine understandable metadata to describe the type and relation of any web resource. Each small piece of information is represented as an RDF statement, also called a "triple", of subject-predicate-object. Internationalized Resource Identifiers (IRIs) are used to name each part of the "subject-predicate-object" triple [61]. The semantics and syntax in a given RDF model are defined in controlled vocabularies or ontologies. An ontology normally consists of two components: a terminology component (TBox) and an assertion component (ABox) [68]. TBoxes refer to the description at a conceptual level, while ABoxes store the data that is a realization of the concepts defined by the TBox. SPARQL serves as an RDF query language and data access protocol for the semantic web [67].

2.2 Semantic representation in the general chemistry domain

Chemical informatics has a long history of utilizing semantic web technologies. One of the first attempts is the Chemical Semantic Web by Murray-Rust and co-workers [9, 24, 51] where Chemical Markup Language (CML) was employed to host the data. CML is an XML language designed to hold most of chemistry's central concepts. It was later extended with the development of CompChem by adding computational chemistry semantics on top of the CML schema [56]. Although CML schema represents chemical data, covering concepts related to atoms, molecules, reactions, computational chemistry and spectroscopy, it is not capable of encoding any desired knowledge in such a way that the meaning is wholly preserved. In contrast, semantic web ontologies aim to explicitly describe and relate objects using formal logic-based representation.

Since OWL became more and more popular in modeling ontologies, more activities of ontology development have been demonstrated in the scientific domain. In the general chemistry and biology domain, the Chemical Entities of Biological Interest (ChEBI) ontology [11, 30, 31] from the European Bioinformatics Institute (EMBL-EBI) is probably one of the most widely used. ChEBI ontology provide knowledge at both a terminology and an assertion level. It is a publicly available and manually annotated ontology, containing approximately 60,000 fully-annotated entities, and over 100,000 preliminary (partially-annotated) entities, as of the last release [17]. It provides a comprehensive and well-documented classification of chemical entities. However, ChEBI is manually maintained and as such cannot easily scale to the full scope of publicly available chemical data. Other ontologies with a similar scope in the general chemistry domain are the Chemical Information Ontology (CHEMINF) [29] and the Chemical Methods Ontology (CHMO) [16]. CHEMINF represents chemical structure and richly describes chemical descriptors and properties, whether intrinsic or computed. It also includes the definition of commonly used software and algorithms, like the PubChem software library, as well as format specifications for chemical data, such as the MOLfile format specification. Complementary to CHEMINF that covers the computational and theoretical methods, CHMO intends to describe the physical and practical ones, such as mass spectrometry and electron microscopy. These ontologies aim to encode the terms, definitions, and logical axioms of chemical information entities. Contrary to ChEBI, CHEMINF and CHMO are just terminological ontologies and do not include any assertion components.

Ontologizing existing databases in RDF format is also demonstrated in the community.

ChEMBL RDF [72] converts data from the ChEMBL database [15, 48] into RDF triples. ChEMBL is a database of bioactive molecules with drug-like properties. The EBI RDF platform encompasses six public life science databases including ChEMBL, UniProt, Reactome, BioModels, BioSamples, and Expression Atlas [34]. Bio2RDF serves as a mash-up system that integrates publicly available bioinformatics databases to provide interlinked life science data [4]. PubChemRDF [21] is the semantic version of the current largest open-source chemical information repository, PubChem [38]. The PubChemRDF content includes the core chemical information archived in the PubChem compound and substance databases, the semantic relationships between compounds and substances, and the provenance and attribution metadata of substances.

Table 1: *Summary of key ontologies and semantic databases in the general chemistry domain. Text in blue indicates novel content and features of OntoSpecies, which distinguish it from the other existing semantic databases.*

Name	TBox	ABox	Info on chemical species	SPARQL endpoint	Description
CHEMINF	✓	×	identifiers ✓ computed properties ✓ experimental properties ✓ classification and uses × spectral information ×	×	ontology (OWL)
CHMO	✓	×	identifiers × computed properties × experimental properties × classification and uses × spectral information ✓	×	ontology (OWL)
ChEBI	✓	✓	identifiers ✓ computed properties × experimental properties × classification and uses ✓ spectral information ×	×	semantic database (OWL, RDF) (manually annotated)
PubChemRDF	✓	✓	identifiers ✓ computed properties ✓ experimental properties × classification and uses × spectral information ×	×	semantic database (RDF)
ChEMBL RDF	✓	✓	identifiers ✓ computed properties ✓ experimental properties × classification and uses × spectral information ×	×	semantic database (RDF)
IDSM	✓	✓	identifiers ✓ computed properties ✓ experimental properties × classification and uses ✓ spectral information ×	✓	semantic database (RDF)
OntoSpecies (this work)	✓	✓	identifiers ✓ computed properties ✓ experimental properties ✓ classification and uses ✓ spectral information ✓	✓	dynamic KG (self-growing and cross-domain connections)

One of the limitations of the RDF version of these databases is that they do not natively support SPARQL endpoints and protocols. Galgonek and Vondrasek recently addressed this issue by integrating PubChem, ChEMBL, and ChEBI data sets into one database called the Integrated Database of Small Molecules (IDSM) that is accessible through a SPARQL endpoint that provide an access point to chemical data from different sources [22]. The IDSM database was created by loading the RDF data downloaded from

ChEMBL, ChEBI and PubChem servers. However, another major limitation is that the RDF version of these datasets includes only basic information on chemical species, i.e. identifiers, geometry and computed properties like mass and charges. They are generally used for the annotation of chemicals and navigation of search results. Data on experimental properties or spectral information are currently not reported in any of the mentioned databases. These information are crucial for applications in reaction modeling or lab automation as described later in the paper.

Tab. 1 shows the main characteristics of the current developed semantic databases and ontologies in general chemistry. In summary, to our knowledge there is no semantic database that represents data on chemical species at a deep level of knowledge and that is accessible through a SPARQL endpoint. In this work, with the development of OntoSpecies, we aim to address all the current limitations, with an ontology where the FAIR principles are better realized and a deeper level of knowledge is represented.

2.3 The World Avatar

The World Avatar project aims to create a digital representation of the real world. The digital world is composed of a dynamic knowledge graph (KG) that contains concepts and the data that describe the world, and an ecosystem of autonomous computational agents. The agents simulate the behaviour of the world and continuously update the concepts and data. A KG is a network of data expressed as a directed graph, where the nodes of the graph are concepts or their instances (data items) and the edges of the graph are links between related concepts or instances. KGs are often built using the principles of Linked Data. They provide a powerful means to host, query and traverse data, and to find and retrieve related information. The autonomous computational agents are the key aspect of the dynamic nature of the KG. They continuously and independently act on the KG performing various tasks. Such tasks include performing calculations using data in the KG, passing information to other software/users outside of the KG and then taking these results to create new instances in the KG, updating existing instances in the KG with improved information where appropriate. Agents perform these tasks with the aim of producing a self-growing, self-updating, and self-improving KG. Additionally, an agent can also consist of sub-agents to answer more complex queries or perform sophisticated tasks.

TWA KG currently includes several ontologies that span a variety of domains. In the process engineering and industrial domains, this includes the well-established OntoCAPE, an ontology for computer aided process engineering that has been integrated into TWA [46] as well as OntoEIP for describing the functions and interactions underlying Eco-Industrial parks [75–77]. In energy and power systems, OntoPowerSys was developed to describe electrical power systems that support industrial plans [12]. This has also been coupled with OntoTwin, an ontology that allows cross-domain coupling [13]. Ontologies for semantic smart city planning by utilizing 3D models include OntoCityGML [14] and the Weather Ontology [60]. Finally, several ontologies have been developed in the chemistry domain to describe different types of chemical data. It provides ontologies for representation of chemical species (OntoSpecies) [20], chemical reaction kinetics (OntoKin) [20], quantum chemistry (OntoCompChem and OntoPESScan) [42, 49], and reaction experiments (OntoReaction - under development).

3 OntoSpecies

OntoSpecies serves as the core ontology in TWA chemistry domain. It is an ontology that describes unique chemical species and their chemical properties. In its previous implementation, OntoSpecies was designed to capture basic information about species, such as empirical formula and molecular weight. It was designed to be linked with other ontologies like OntoKin [20] and OntoCompChem [42]. In this work, OntoSpecies has been extended to include a wide range of identifiers, chemical and physical properties, chemical classifications and applications, plus spectral information associated with each species, and the provenance and attribution metadata. Most of the information about a species are collected from the info archived in the respective PubChem compound web source as explained in section 4.

3.1 Terminology component

OntoSpecies TBox containing the full class and relational definitions is available at `http://159.223.42.53:5003/ontospecies`. Each entity is uniquely represented by an IRI. These IRIs are lexically meaningful to help to query information from the KG. A set of standardized ontologies for enhanced data integration and interoperability were collected to define the domain specific knowledge, including CHEMINF [29], CHMO [16], Units of Measurement (om) [25], Gainesville Core (gc) [54] and Simple Knowledge Organization System (skos) [66]. Adoption of these core ontologies helps to ensure that the mapping of chemical information is compatible across multiple semantic web resources. For convenience, prefixed names are used instead of the full IRIs in the following sections that describe the schema and its use cases. The definitions of namespace prefixes used in the paper are summarised in Tab. 2.

Table 2: *The prefixes and corresponding namespaces of the ontologies used in this work.*

Prefix	Namespace
os	<http://www.theworldavatar.com/ontology/ontospecies/OntoSpecies.owl#>
okin	<http://www.theworldavatar.com/ontology/ontokin/OntoKin.owl#>
occ	<http://www.theworldavatar.com/ontology/ontocompchem/OntoCompChem.owl#>
rdf	<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs	<http://www.w3.org/2000/01/rdf-schema#>
xsd	<http://www.w3.org/2001/XMLSchema#>
owl	<http://www.w3.org/2002/07/owl#>
skos	<http://www.w3.org/2004/02/skos/core#>
om	<http://www.ontology-of-units-of-measure.org/resource/om-2/>
pt	<http://www.daml.org/2003/01/periodictable/PeriodicTable#>
gc	<http://purl.org/gc/>
CHEMINF	<http://semanticscience.org/resource/>
CHMO	<http://purl.obolibrary.org/obo/>

The main class of the ontology is `os:Species` (Fig. 1). A species can be defined as "an ensemble of chemically identical molecular entities" [47] and can be seen as the equivalent of a PubChem "compound". Because of the different nomenclature systems that are used by different organizations to refer to a species, we decided to assign its molecular formula as label to the species (predicate `rdfs:label`). However, all the synonyms given to a species are also stored as alternative labels (predicate `skos:altLabel`). This helps querying for a chemical species.

A species is composed of atoms (class `gc:Atom`). An atom is associated to an element (class `pt:Element`) in the periodic table through the object property `os:isElement`. Atoms are connected together by chemical bonds (class `os:AtomicBond`). A chemical bond has a bond order (predicate `os:hasBondOrder`) that is a data property of type `xsd:integer` and is identified by (predicate `os:definedBy`) the two atoms that participate in the bond. The atom coordinates in space (classes `os:XCoordinate`, `os:YCoordinate` and `os:ZCoordinate`) define the geometry (class `os:Geometry`) of the species. Atom coordinates are linked to the geometry by the predicate `os:fromGeometry`. Functional groups (class `os:FunctionalGroup`), groups of atoms within a molecule that have similar chemical properties whenever it appears in various compounds, are associated to a species through the predicate `os:hasFunctionalGroup`. Species and elements are also linked to a range of other concepts that are subclasses of one of the following classes: `os:Identifier`, `os:Property`, `os:Classification`, `os:Use` or `os:SpectralInformation`.

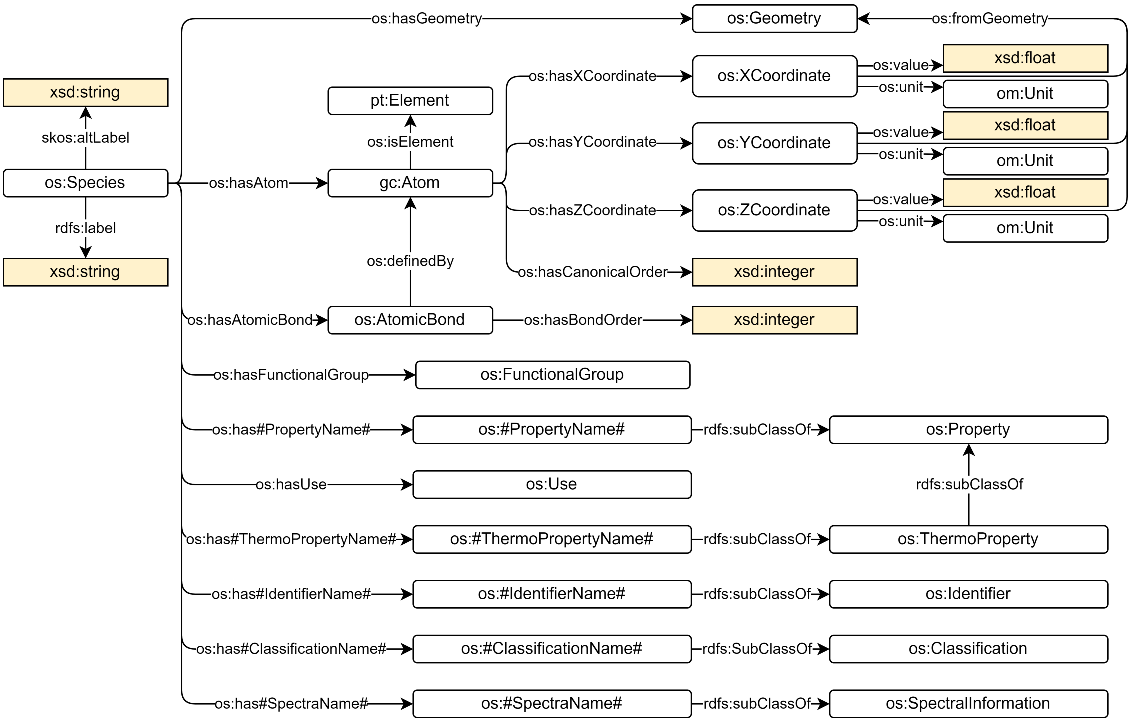


Figure 1: *`Species` concept in the OntoSpecies TBox.*

An identifier (class `os:Identifier`) provides a way for the identification of a species or element, which in most cases should aim to be unique and easy to use as an unambiguous reference for the chemical entity. The identifier representation consists of triples specifying the value (predicate `os:value`) and the provenance (predicate `os:hasProvenance`) of the identifier (Fig. 2). Examples of subclasses of the class `os:Identifier` (defined as `os:#IdentifierName#` in Fig. 1 and 2) are the International Chemical Identifier -

InChI (class `os:InChI`), the InChI Key (class `os:InChIKey`) or the International Union of Pure and Applied Chemistry (IUPAC) Name (class `os:IUPACName`).

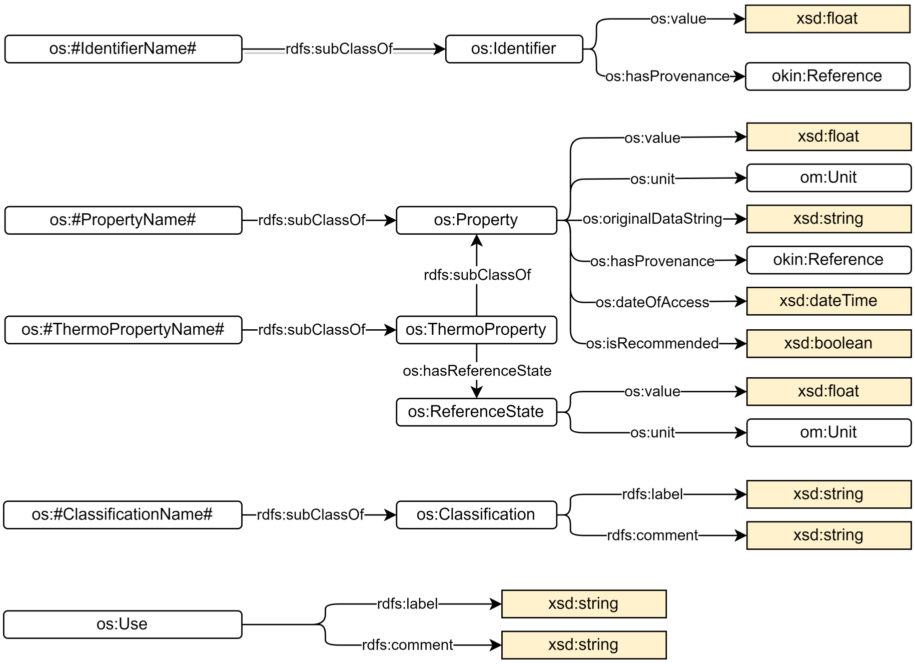


Figure 2: *`Identifier`, `Property`, `Classification` and `Use` concepts in the OntoSpecies TBox.*

The `os:Property` class represents a chemical or physical property of a species or element. The property representation consists of triples specifying the value (predicate `os:value`), the unit (predicate `os:unit`), and the provenance (predicate `os:hasProvenance`) of the property (Fig. 2). Because most of the properties are collected from the PubChem database where they are stored in PubChem as strings, the original string (class `os:originalDataString`) and the date of acquisition (data property of type `xsd:dateTime`) of the property are also linked to the property. A `os:ThermoProperty` is a subclass of `os:Property` that is evaluated at a reference state (class `os:ReferenceState`). Examples of subclasses of the class `os:Property` (defined as `os:#PropertyName#` in Fig. 1 and Fig. 2) are the molecular weight (class `os:MolecularWeight`) or the charge (class `os:Charge`) of the species. Examples of subclasses of the class `os:ThermoProperty` (defined as `os:#ThermoPropertyName#` in Fig. 1 and Fig. 2) are the boiling point (class `os:BoilingPoint`), density (class `os:Density`) or solubility (class `os:Solubility`) of the species.

Chemical species can be grouped by their structure or similar features, by their application and role, or by other factors. They are also classified by organizations around the world. As an example, GHS (Globally Harmonized System of Classification and Labelling of Chemicals) is a United Nations system to identify hazardous chemicals and to inform users about these hazards [62]. GHS has been adopted by many countries around the world and is now also used as the basis for international and national transport regulations for dangerous goods. In OntoSpecies we include `os:ChemicalClass` and `os:GHSHazardStatment` as subclasses `os:Classification` and the class `os:Use` to represent the application and role of a species (Fig. 2).

Spectral data (`os:SpectralInformation`) is also linked to the species. At the moment, 1-D and 2-D nuclear magnetic resonance (NMR) (classes `os:1DNMRSpectra` and `os:2DNMRSpectra`) and mass spectrometry (MS) (class `os:MassSpectrometry`) are included in the OntoSpecies TBox. Every spectrum is associated to a graph (class `os:SpectraGraph`) that shows the peaks (class `os:Peak`) of the species. Additional information are also recorded like the ionization mode (class `os:IonizationMode`) for spectra of type `os:MassSpectrometry` or solvent and frequency (classes `os:Solvent` and `os:Frequency`) for spectra of type `os:NMRSpectra`.

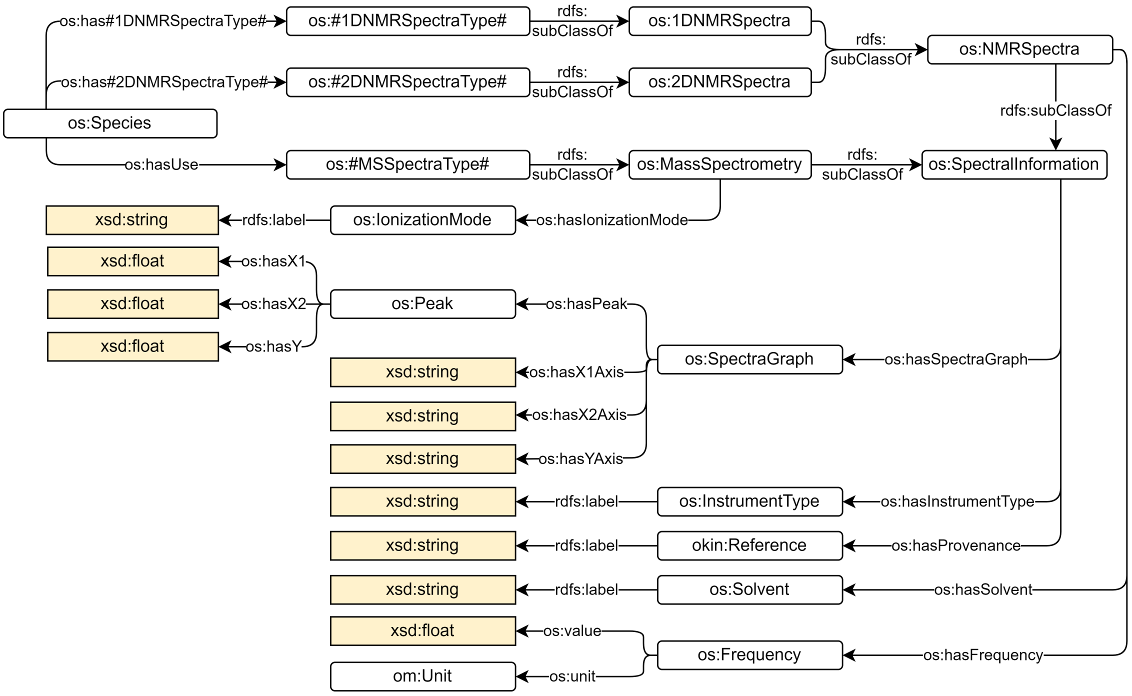


Figure 3: *`SpectralInformation` concept in the OntoSpecies TBox*

PubChem, as an archive, takes care to preserve the provenance of information. Identifiers and properties are linked to the respective sources (class `okin:Reference`) through the predicate `os:hasProvenance` (see Fig. 2). The provenance can be given as a URL of a web page or the DOI of a scientific paper.

A full list of classes in OntoSpecies and their description is reported in appendix A.1. Classes that represent concepts which are already defined in the CHEMINF or CHMO vocabularies are linked to their corresponding class in those vocabularies through the use of the predicate `owl:equivalentClass`. The association with central vocabularies like CHEMINF or CHMO enables comparison between chemical properties arising from different databases in a standardized fashion thus helping interoperability. The choice to use a lexically meaningful IRI to represent the property instead of directly using its equivalent CHEMINF or CHMO class enhances the reading and querying of data by humans.

3.2 Assertion component

In OntoSpecies ABoxes of the ontology, each entity is uniquely represented by an IRI. The ABox IRIs have the prefix:

```
PREFIX oskg: <http://www.theworldavatar.com/kg/ontospecies/>
```

The IRI for a species is assigned as:

```
oskg:Species_UUID
```

where UUID (Universal Unique Identifier) is a 128-bit value used to uniquely identify an object or entity on the internet generated at the moment of the species instantiation.

The IRI for instances directly related to a chemical species (e.g. boiling point of a species, molecular weight a species or IUPAC name of a species) were constructed based on a combination of their class type name and the species UUID. The IRI also includes an index entry right after the class type name because multiple instances of the same class type might be connected to the same species. For example, the IRI for the boiling point of a species is represented as:

```
oskg:BoilingPoint_#index#_Species_UUID
```

where `#index#` goes from 1 to the number of boiling point instances connected to `oskg:Species_UUID`.

IRIs of other concepts that are not directly related to a chemical species (e.g., use, unit or provenance) have the following syntax:

```
oskg:#ClassTypeName#_UUID
```

Including the class type name in the IRI makes easier for a human user to understand what the IRI represents, reducing the errors in the querying stage.

3.3 Links to other ontologies in TWA

OntoSpecies plays a central role in TWA, enabling the linking of species to instances and concepts deriving from other ontologies in TWA KG chemistry domain (see Fig. 4). The other ontologies included in TWA chemistry domain currently are OntoKin [20], OntoCompChem [42], OntoPESScan [49], OntoMOPs [41] and OntoReaction (under development).

OntoKin is an ontology that represents reaction mechanisms [20]. In a chemical process, a reaction mechanism constitutes a set of stoichiometric reactions involving different chemical species. A reaction is described through products and reactants that are further

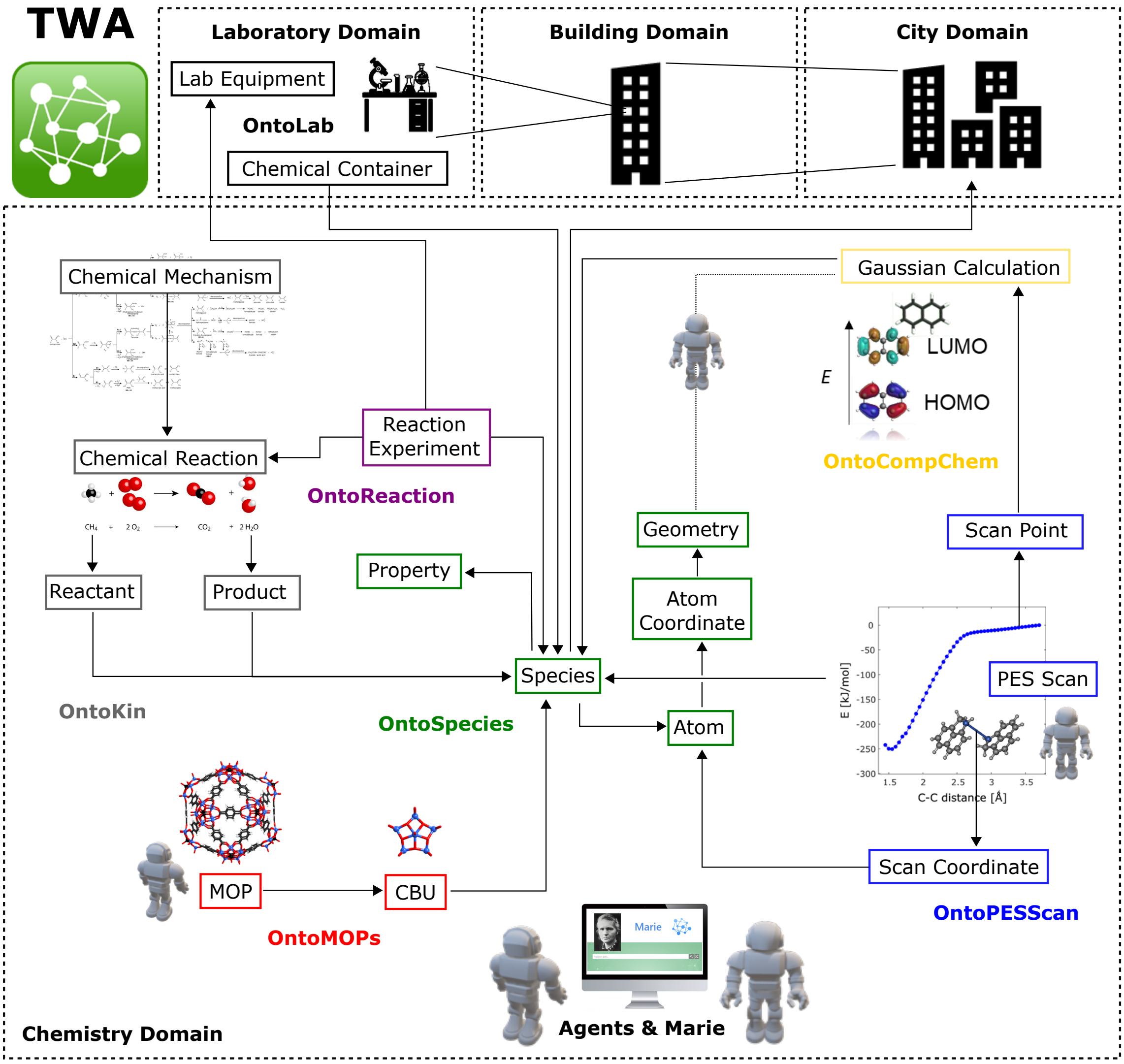


Figure 4: *Connection of OntoSpecies to other segments of TWA KG.*

described through different thermodynamic and transport model concepts and identified via OntoSpecies IRIs. OntoKin, in conjunction with OntoSpecies, can provide a facile and unambiguous comparison between other kinetic, thermodynamic, or transport models reported in the literature.

The OntoCompChem ontology represents the input and output of density functional theory (DFT) calculations, currently mainly focused on molecular systems [42]. OntoCompChem currently represents single point calculations, geometry optimizations and frequency calculations. Information like the final converged self-consistent field (SCF) energy and the calculated frontier orbitals are also stored for each calculation. For geometry optimizations, the final optimized geometry is represented, while for frequency calculations, it stores the zero-point energy correction and a full list of the computed vibrational frequencies. For the representation of potential energy surface (PES) scans, a different ontology has been developed (OntoPESScan) [49]. A PES scan calculation is linked to a number of single point calculations in OntoCompChem as every scan can be seen as a collection of those. The unique identification of bonds and atoms in OntoSpecies is used for the identification of geometry changes between calculations.

OntoReaction semantically describes reaction experiments. Species that take part in the reaction experiment are identified via OntoSpecies IRIs. By assigning different IRIs to species that are based on different isotopes, charges, or spin states, OntoSpecies becomes relevant also for the digital representation of isotope labeling experiments, redox and electrochemically driven processes, and photochemistry.

OntoMOPs ontology describes metal-organic polyhedra (MOPs) [41]. MOPs are assemblies of organic and metal-based chemical building units (CBUs) resembling the shape of regular polyhedra. The CBUs are instantiated as species in OntoSpecies.

OntoSpecies is also directly or indirectly linked with other domains in TWA. OntoLab (under development) and related ontologies semantically describe equipment such as reactors, autosamplers, and chromatography devices, in order to build complete digital twins of (chemical) laboratories. A direct link between OntoSpecies and OntoLab is the link between a chemical container and the species that it contains, while an indirect link is through OntoReaction. The reaction experiment is linked to the species involved in the reaction and it is also linked to the lab equipment where the reaction experiment is carried out. In the context of digital urban planning, data on chemical species in conjunction with real-time weather data and data on the physical infrastructure can be used to predict the dispersion of pollutants in the atmosphere of urban areas [50].

Software agents are semantic web services that act upon the knowledge graph, collecting, updating and creating knowledge. Several examples of agents and cross-domain application exists in TWA chemistry domain. To cite some, for existing DFT calculations (OntoCompChem), an agent instantiates thermal properties (enthalpy, heat capacity and entropy) back to the involved species (OntoSpecies) as well as 7-coefficient NASA polynomials to the reaction (OntoKin) [42]. The species IRI (OntoSpecies) serves as a link between the reaction mechanisms (OntoKin) and the DFT calculation (OntoCompChem). This enables reaction modeling that requires kinetic and thermodynamic factors. The information on molecular geometry (OntoSpecies) can be retrieved and used by an agent as an initial guess of the geometry for quantum chemical calculations (OntoCompChem), and the obtained geometry can be stored as new geometry connected to the species with the appropriate provenance. A reaction experiment (OntoReaction) can be linked to a reaction mechanism (OntoKin) through the chemical species that take part in the experiment and reaction (OntoSpecies) and agents can do sensitivity analysis and calibration, as demonstrated for combustion experiments [1]. The MOP discovery agents shows that MOPs can be rationally designed, revealing which CBUs can be meaningfully combined without causing undesired strains [41].

A user-friendly question answering interface, “Marie”, enables users to retrieve information from TWA KG chemistry domain using their natural language, which is then translated behind the scenes into machine readable query. “Marie” makes our infrastructure available to users who are not aware of the knowledge structure in the KG and its different domains [78, 79].

In summary, the interconnection of the ontologies in different domains of TWA in which OntoSpecies plays a central role, as well as the integration with agents and “Marie” has three main outcomes:

1. Facilitate the search for information on a chemical species, reactions, quantum chemistry calculations and experiments through federated queries (see section 5.2) or asking questions in natural language to “Marie”.
2. Create new knowledge using information from different subdomains.
3. Bridge the gap between molecular-scale chemistry and real-world macro-scale phenomena.

4 Population and querying

To help with creating and uploading entries to the KG, a software agent has been developed to request data from chemistry databases and create the necessary OWL files for the OntoSpecies instances, streamlining the population process. The agent is freely accessible at `https://github.com/cambridge-cares/TheWorldAvatar`. The current version requests information from PubChem and ChEBI, but can be easily extended to other chemical databases. The agent can be used locally or as a simple web application. Its UML diagram is reported in Fig. 5.

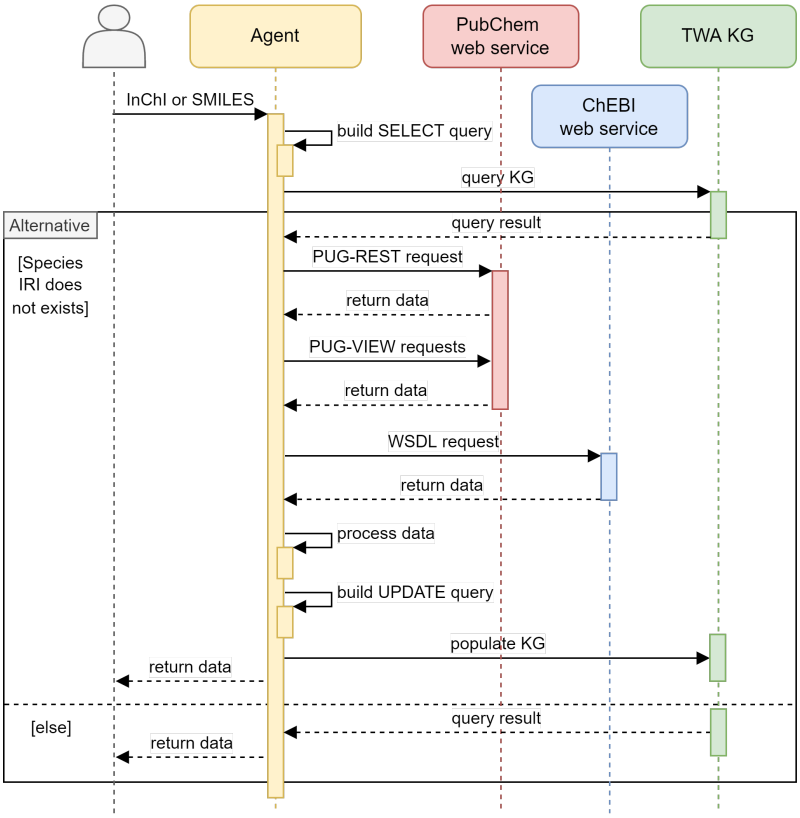


Figure 5: *UML sequence diagram of the software agent.*

The agent receives InChI or SMILES identifier of a species as input and builds its queries accordingly. In a first step, it checks if the species is already instantiated. If this is not the case, the agent requests relevant information from PubChem and ChEBI.

PubChem provides two Representational State Transfer (REST) interfaces as programmatic access routes to PubChem data: PUG-REST [35–37] and PUG-View [39]. PUG-REST retrieves information computed by PubChem, while PUG-View gives information collected from other sources, including annotations. We decided to use PUG-REST to obtain information on identifiers and computed properties by PubChem and PUG-View to obtain experimental properties, chemical classes, uses, spectral information, and synonyms of species that are not accessible by PUG-REST request. While computed properties are reported using the same format (value plus unit), the experimental properties are reported as strings, from different sources, and with different syntax and units. For example, if we look at the solubility entries for ethanol in PubChem, the following strings are listed on the compound webpage:

- greater than or equal to 100 mg/mL at 73 °F (NTP, 1992)
- 1000000 mg/L (at 25 °C)
- In water, miscible /1X10+6 mg/L/ at 25 °C
- Miscible with ethyl ether, acetone, chloroform; soluble in benzene
- Miscible with many organic solvents
- 1000.0 mg/mL
- Solubility in water: miscible
- Soluble in water
- Miscible

In order to turn these strings into quantitative values that can be assigned to the data property in OntoSpecies, we used the unit-parse python package [69]. For each string, a parser translates the string in cleaned and parsed quantity and unit that are then converted into SI units. This enables easy comparability of properties between different species. If a numerical value is not detected, the property string is discarded by the agent. If multiple values for the same property are collected from different strings, all the values are saved in different instances of the property class. A preferred value is also selected by first grouping approximate equivalent quantities and then discarding entries with faulty units or far-off values. The preferred value is then stored in a new instance of the property class and the value "True" is assigned to the data property `os:isRecommended` (see Fig. 2) and the label "PubChem agent" is assigned to its provenance.

Classification of species is taken from ChEBI database through Web Services Description Language (WSDL) request. Chemical classes in ChEBI are organized in a tree structure, also known as a hierarchy, based on the chemical structure of the entities. The lower levels of the hierarchy become more specific, with each category describing a smaller subset of chemical entities. This hierarchical structure allows for easy navigation and categorization of chemical entities based on their properties and characteristics [30]. When a new chemical class from ChEBI is instantiated in OntoSpecies, the agents checks if all the parent classes (upper level classes) are already instantiated. In case they are not, the agent creates an instance of the parent class and checks for its parent classes. The process is repeated until the full ChEBI classification hierarchy is imported into OntoSpecies. Data on functional groups along with their hierarchy is also collected from ChEBI.

The collected data from PubChem and ChEBI for a chemical species is then put in the form of a SPARQL update and instantiated in the KG. As of April 2023, the OntoSpecies KG includes more than 35,000 species and new species along with their properties are continuously instantiated.

Information on how to access OntoSpecies can be found at (`http://159.223.42.53:5003/ontospecies`).

5 Utility and discussion

In this section, we want to showcase and emphasise the wide range of possible use cases across multiple domains of our approach.

5.1 Data access via complex queries

The ontological format permits complex queries, easy data analysis, and visualization. This can be used to compare chemical properties of similar compounds, find compounds with required characteristics as well as automate laborious data gathering from research activities. To demonstrate the strengths and capabilities of the extended OntoSpecies, several use cases are presented in this section along with their corresponding SPARQL queries. The successful application of software agents postprocessing the obtained data and dynamically expanding the KG is also showcased.

USE CASE 1: Reproduce, monitor, and investigate trends in chemical properties.

Different classes of molecules and compounds follow specific trends in their physical properties - e.g., boiling temperature. These trends can be monitored against a well defined parameter relevant to the structure of the molecules such as the carbon chain length in aliphatic hydrocarbons. Classical examples are homologous series of organic molecules such as straight-chain alkanes, alkenes, and alcohols [63]. The OntoSpecies ontology provides the infrastructure for time-efficient queries of compounds' physical properties and structural information to illustrate trends.

Fig. 6 shows a SPARQL query that selects the boiling point temperature of all species classified as alcohols. The query returns the number of carbon atoms, boiling point temperature in K and the boiling point reference pressure in kPa for 214 different alcohol species. Fig. 7 shows the boiling point temperature versus the number of carbon atoms for alcohols (circles).

```
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 ?speciesIRI ?formula ?nC ?BPvalue ?RefStateValue
WHERE {
   # select a species and its chemical formula
   # ?formula - species chemical formula
   ?speciesIRI rdf:type os:Species ; rdfs:label ?formula .

   # select boiling point in unit kelvin and its reference pressure
   # ?BPValue - boiling point value
   # ?RefStateValue - boiling point reference pressure value
   # "K" - label for unit Kelvin
   # "PubChem agent" - label for accessing properties recommended by the agent
   ?speciesIRI os:hasBoilingPoint ?BP .
   ?BP os:value ?BPValue ; os:unit ?u ; os:hasReferenceState ?RefState ;
      os:hasProvenance ?provenance .
   ?unit rdfs:label "K".
   ?RefState os:value ?RefStateValue .
   ?provenance rdfs:label "PubChem agent" .

   # select species classified as "alcohol" navigating the classification tree
   ?speciesIRI (rdf:|!rdf:)* ?x .
   ?x ?y ?z .
   ?z (rdf:|!rdf:)* ?classIRI .
   ?classIRI rdf:type os:ChemicalClass ; rdfs:label "alcohol" .

   # select the number of carbons from the chemical formula
   # and store the result in the variable ?nC
   BIND(strbefore(strafter(str(?formula),'C'),'H') AS ?nC)
}
```

Figure 6: *SPARQL query that returns the number of carbons, boiling point and the boiling point reference pressure of alcohols.*

Filtering species only based on broad chemical classes such as alcohols might not lead to a well defined trend as shown in Fig. 7a. This is caused by differences among species in the same chemical class that may have additional functional groups or a very different structure (e.g., primary alcohol, secondary alcohol or presence of double bonds). Additional restrictions can then be applied. As an example, alkanols species are alcohols that have a chemical formula of the type $C_xH_{2x+2}O$. This subclass of alcohols can be selected by adding a filter on the chemical formula to the query reported in Fig. 6. The results are plotted in Fig. 7b (circles) and we can see that they follow a more defined trend than their parent class (alcohols). Fig. 7b also show the boiling point trend for alkanes (diamonds) and alkenes (squares).

The application of the chemical class taken from ChEBI in combination with boiling points provided by PubChem demonstrates the interoperability of the TWA approach. Moreover, the use of additional filters such as sum formulas shows the capability of adding knowledge that is not included in the data sources, such as the concept of an alkanol that does not exist in ChEBI (see section 5.2). Trend charts can also be helpful in inferring unrecorded properties like the reference state which is missing in some of the data. For example, in Fig. 7b, white diamonds and squares (alkanes and alkenes with none reference pressure) are well aligned with green squares (alkenes at 101 kPa) and red diamonds (alkanes at 101 kPa) which suggests that the none-recorded reference pressures are highly probable to be 101 kPa. Similar analysis for white circle (alkanols with no given reference pressure) suggests that this data is collected at the reference pressure of 101 kPa, too.

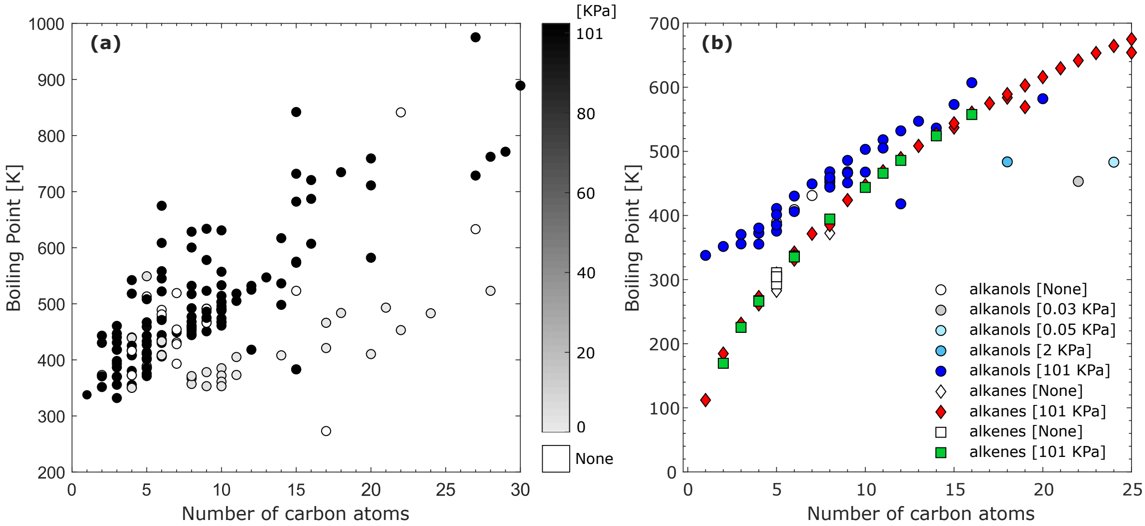


Figure 7: *Illustration of boiling point trends for different classes of aliphatic organic molecules in comparison with increasing carbon chain length generated by OntoSpecies queries. (a) alcohols (b) alkanes, alkenes, alkanols. Different colors correspond to different reference pressures.*

More complex and lesser-known trends can also be investigated. For example, the size-dependence of alkanes' ionisation energies shown in Fig. 8 is still a topic of ongoing research [3]. The trend observed via experimental results by Bakulin et al. [3] to develop an underlying model were qualitatively reproduced by data available via PubChem.

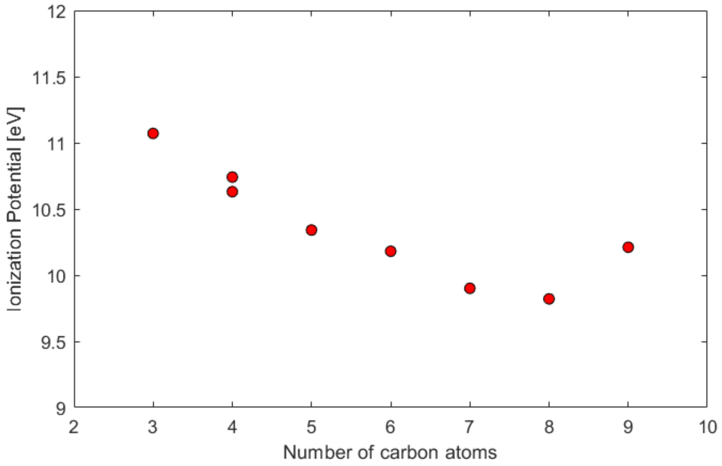


Figure 8: *Trend of ionisation energies of saturated hydrocarbons with increasing carbon chain length.*

Finally, the ability to easily investigate properties of classes also enables prediction of unknown values. In the simplest case, unknown values of a molecule that is part of a group with a well-known trend along a single variable (e.g., boiling points of large alkanes) can be extrapolated (see section 5.2). But also multivariate models based on machine learning or group contributions can be developed.

USE CASE 2: Selection of suitable solvents based on multiple criteria.

The selection of suitable solvents is a key challenge in experimental chemistry. Due to the importance and complexity of this multi-objective task (especially considering heightened focus on sustainability), many guidelines and tools have been developed [7, 10, 57] to assist in the selection of solvents for a given reaction or separation. Most selection criteria can be summarised in four categories:

1. Technical considerations: stability at operating condition (e.g., boiling point), miscibility with solute (e.g., solubility in water), etc.
2. Separability for detection (e.g., different molecular weight for MS detection, different NMR peak locations) or further processing (e.g., different boiling points for crystallisation).
3. Sustainability considerations: efficient solvent utilisation, recovery and re-use (e.g., boiling point for ease of distillation and separability of mixtures) and potential environmental, health, and safety impacts (i.e. acute toxicity, flash point for risk of ignition).
4. Economic considerations (e.g. raw material cost, and cost related to disposal, recovery, abatement and liability).

Criteria of all four categories are mostly related to quantitative properties that are represented within OntoSpecies (see section 3). Complex queries with OntoSpecies therefore enable selection or pre-selection of solvents based on multiple criteria. As an example, a query can be used to assist chemists and process engineers with decisions regarding solvent separability. Solvent exchanges through distillation are quite common in pharmaceutical syntheses, and not all solvent mixtures are easily separated.

Fig. 9 shows a query that selects all species that can be used as co-solvents for propan-2-ol and are classified as alcohols. We decided to use alcohols because they are generally more desirable than solvents in other classes (e.g. bases or ethers class) [10]. We applied two selection criteria:

1. The co-solvent boiling point needs to be 15 K lower or higher than the propan-2-ol boiling point to ensure ease of separation between the two solvents by distillation.
2. Exclude co-solvents with high potential health impact. To do so, we removed all the species that have GHS safety statement related to risk of cancer and risk for unborn child [7].

```
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 ?speciesIRI ?iupacstring
WHERE {
   # select a species and its IUPAC name
   # ?iupac - species IUPAC name IRI
   # the filter on the IUPAC name IRI selects the IUPAC name indexed 1
   ?speciesIRI rdf:type os:Species ; rdfs:label ?formula ; os:hasIUPACName ?iupac .
   ?iupac os:value ?iupacstring .
   FILTER(regex(str(?iupac) ,'_1_'))

   # select species classified as "alcohol" navigating the classification tree
   ?speciesIRI (rdf:|!rdf:)* ?x .
   ?x ?y ?z .
   ?z (rdf:|!rdf:)* ?classIRI .
   ?classIRI rdfs:label "alcohol" .
   FILTER( regex(?formula, "C[0-9]{0,3}H[0-9]{0,3}O[0-9]{0,3}$") )

   # select species used as "solvent"
   ?speciesIRI os:hasUse ?use .
   ?use rdfs:label ?usestring .
   FILTER(regex(lcase(?usestring), "solvent"))

   ###################################### CRITERION 1 ######################################
   # select species with boiling point ?BPValue < Tb - 15 K and ?BPValue > Tb + 15 K
   # where Tb = 355 K (propan-2-ol boiling point)
   ?speciesIRI os:hasBoilingPoint ?BP .
   ?BP os:value ?BPValue ; os:unit ?u .
   ?u rdfs:label "K" .
   FILTER(?BPValue > 370 || ?BPValue < 340)

   ###################################### CRITERION 2 ######################################
   # select species that do not have hazard statements related to unborn child or cancer
   FILTER NOT EXISTS{
      ?speciesIRI os:hasGHSHazardStatements ?ghs .
      ?ghs rdfs:comment ?ghsstring ;
           rdfs:label ?ghscode .
      FILTER((regex(?ghsstring,'unborn child') || regex(?ghsstring,'cancer'))  &&
      ↪  regex(?ghscode,'H3'))
   }
} ORDER BY ?iupacstring
```

Figure 9: *SPARQL query to select a suitable co-solvent for propan-2-ol to enable easiest separation by distillation (criterion 1) and with high health impact (criterion 2).*

Applying the first criterion, only five alcohols are reported by the Solvent Selection Guide by Curzions et al. [10] as suitable co-solvents for propan-2-ol to enable easiest separation by distillation: 2-(2-butoxyethoxy)ethanol, 2-methoxyethanol, butan-1-ol, ethylene glycol and methanol. However, a list of 49 solvents is returned querying OntoSpecies, including the five alcohols selected by the Solvent Selection Guide by Curzions et al. [10]. If we apply both the first and second criterion, a list of 45 species is returned by the query. Undesirable species like 2-methoxyethanol were removed from the selection as they are considered dangerous for human health as also reported by GSK and Sanofi's selection guides [7]. The list of compounds are reported in appendix A.2. This indicates that our approach is evidently capable of assisting in solvent selection considering solvents that are included in existing solvent selection guides, while also considering potential new solvents that are not commonly included in these guides.

It is worth noting that some of the species might be too costly to be feasible as a solvent. Cost considerations are excluded from our selection because raw material costs are currently not present in OntoSpecies and their inclusion will be the subject of future work. However, some cost considerations can still be taken into account indirectly - e.g., by selecting solvents with a boiling point that is not too high ($T_b < 150\,^{\circ}\mathrm{C}$) to reduce the cost of solvent recovery by distillation. A list of only 13 species is returned if we include the third criterion in the selection (see column "Criteria 1-2-3" of Tab. A2 in appendix A.2).

USE CASE 3: Identification of species in unknown mixture based on NMR spectrum.

In chemical engineering and related research fields, researchers often need to use tabulated data from websites or publications, which involves tedious searches and manual copying of data into their own tables. This especially applies to the analysis of unknown compounds via spectroscopy. Nuclear magnetic resonance (NMR) spectroscopy is based on the nuclear spin exhibited by some nuclei such as ^{1}H. Based on overall molecular structure, different hydrogen atoms in a molecule exhibit different resonance frequencies at which its spin flips to align with an external magnetic field [63].

In an NMR spectrum, the detected intensity is shown on the vertical axis and the shift of nuclei's resonance frequency is indicated on the horizontal axis. In ^{1}H NMR, different protons exhibit distinguishable resonance frequencies because they are located in different areas of the molecule and exhibit different interactions with other nuclei, resulting in a "chemical shift" compared to a reference standard. As an example, Fig. 10 shows the ^{1}H NMR spectrum of a catholyte sample taken from an electrochemical reactor synthesising a variety of organic molecules from carbon dioxide and water [58]. As minor products of this reaction are partially unknown, the NMR spectrum is used to detect liquid products of electrochemical CO_2 reduction.

To identify the species, expected chemical shifts of all possible products were collected manually, summarized in a table and then compared to the experimental data for every species [43, 58]. This process can be time-consuming and introduces bias by leaving some species out of the analysis. For this reason, there have been many efforts to automate spectral deconvolution and analysis [8, 32, 44], many of which rely on a standardised procedure or manufacturer-specific software [44]. Lately, some impressive results were achieved with the help of machine learning (ML) tools, in particular deep learning techniques [8]. Training of these models requires massive data sets (e.g., derived from PubChem [8]) and often additional knowledge to predict shifts and multiplicities - for example, using an assignment algorithm as described by Howarth et al. [32]. As we have access to known NMR spectra of thousands of components from PubChem as well as structural and classification data from other sources that can be queried, OntoSpecies and TWA enable a dynamic approach to automated spectral analysis.

In an initial attempt, we developed an agent to assist analysis of the NMR spectrum in a semi-automated manner: Firstly, the possible species that can be produced in the electrochemical reaction are detected querying for all the species with chemical formula $C_xH_yO_z$

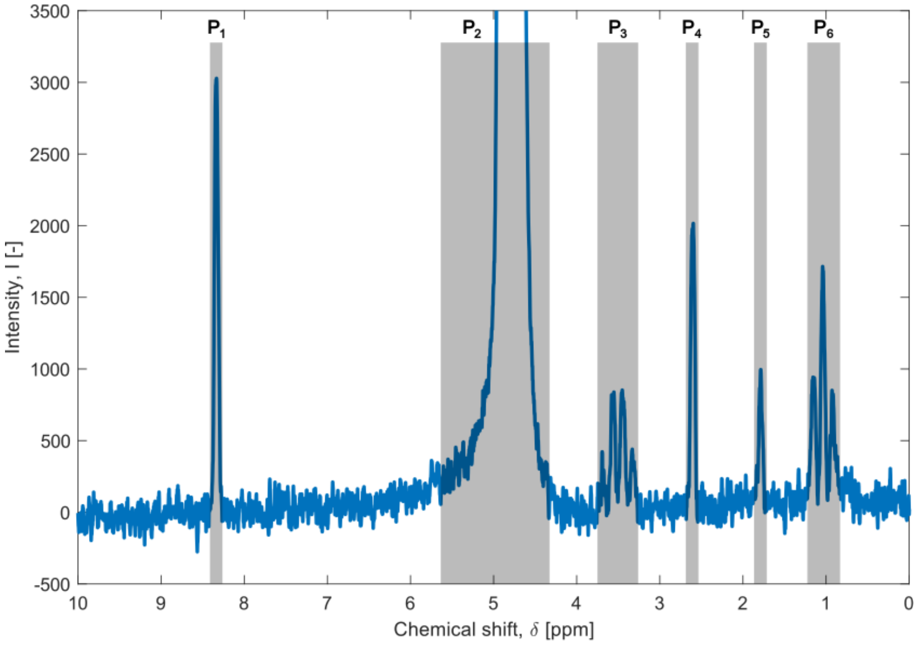


Figure 10: *^{1}H NMR spectrum on a sample of catholyte after $t = 20\,\mathrm{min}$ of electrocatalytical $\mathrm{CO_2}$ reduction with internal dimethyl sulfoxide (DMSO) standard adapted from Rihm et al. [58]. The six identified peaks and their respective ranges that were defined are marked in gray.*

and with $x < 5$ and $z < 10$ (Fig. 11). An additional filter on the boiling point (T_b) is added to remove all the species that are not expected to be found in the liquid phase at room temperature. The selected $T_b = 15\,°\mathrm{C}$ is lower than then the experimental temperature ($T = 25\,°\mathrm{C}$) as to ensure all species that might partially be in the liquid phase are included. The IRIs of 187 species are returned as result of the respective query shown in Fig. 11.

```
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 ?speciesIRI ?formula
WHERE {
   # select a species and its IUPAC name
   # ?formula - species chemical formula
   ?speciesIRI rdf:type os:Species ; rdfs:label ?formula ;
   FILTER( regex(?formula, "C[0-5]{0,1}H[0-9]{0,2}O[0-9]{0,2}$") )

   # select species with boiling point higher than 288 K
   ?speciesIRI os:hasBoilingPoint ?BP .
   ?BP os:value ?BPValue ; os:unit ?u .
   ?u rdfs:label "K" .
   FILTER(?BPValue > 288)
}
```

Figure 11: *SPARQL query selecting all the possible products of the electrochemical reaction expected to be found in the catholyte.*

The IRI of each species is then used in a second query to get the ^{1}H NMR spectral information (peaks shift and intensity) as shown in Fig. 11. A filter is used in the query to select only the first ^{1}H NMR spectra added to OntoSpecies from the PubChem record. However, different kinds of filters that select the spectra according to the solvent used or frequency can be constructed instead. 79 species out of the 186 species have recorded ^{1}H NMR spectral information.

```
PREFIX os: <http://www.theworldavatar.com/ontology/ontospecies/OntoSpecies.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?spectra ?chemicalshift ?intensity
WHERE {
   # select the chemical shift and intensity of each peach of the 1H NMR spectra
   # <#IRI#> in the query text is replaced with the species IRI
   <#IRI#> os:has1HNMRSpectra ?spectra .
   ?spectra os:hasSpectraGRaph ?graph .
   ?graph os:hasPeak ?peak .
   ?peak os:hasX1 ?chemicalshift ;
      os:hasY ?intensity .
   FILTER( regex(str(?spectra), "_1_") )
   # the filter selects only the 1H NMR spectra with index 1
}
```

Figure 12: *SPARQL query selecting ^{1}H NMR spectra peaks information (chemical shift and intensity) of a specific species. #IRI# in the shown query text is replaced with the species IRI.*

The recorded ^{1}H NMR spectrum in Fig. 10 presents six main peaks. The chemical shift (δ_i), the multiplicity, and the expected species contributing to each peak P_i are reported in Tab. 3. The DMSO peak was easily identified by its known chemical shift of 2.6 ppm used as reference point for all other species. The large peak at 4.9 ppm corresponds to the water solvent. A software agent is used to detect which among the 79 species have peaks with chemical shift in the range $[\delta_i - 0.2\,\text{ppm}, \delta_i + 0.2\,\text{ppm}]$, excluding the DMSO chemical shift. Prior to this, all peaks with intensity below 20% of the species' highest peak were discarded. 12 possible species are selected by the agent. We can narrow down the selection even further by considering the peak multiplicity (singlet, triplet, quartet) and counting reported peaks within the given range of chemical shift. The final result obtained from the agent is a list of four possible main species produced in the electrochemical reaction: formic acid, acetic acid, ethanol, ethoxyethane.

Table 3: *Chemical shifts and multiplicity of the peaks seen in the ^{1}H NMR spectrum in Fig. 10 and the expected species contributing to each peaks returned as result from the software agent.*

Peak ID	Chemical Shift, δ [ppm]	Multiplicity	Possible Species
1	$\delta_1 = 8.35$	singlet	Formic Acid ($HCOOH$)
2	$\delta_2 = 4.9$	singlet	Water (H_2O)
3	$\delta_3 = 3.51$	quartet	Ethanol (CH_3CH_2OH), Ethoxyethane ($(C_2H_5)_2O$)
4	$\delta_4 = 2.6$	singlet	DMSO ($(CH_3)_2SO$)
5	$\delta_5 = 1.83$	singlet	Acetic Acid (CH_3COOH)
6	$\delta_6 = 1.04$	triplet	Ethanol (CH_3CH_2OH), Ethoxyethane ($(C_2H_5)_2O$)

As shown in Tab. 3, formic acid and acetic acid can be unmistakably identified, while the triplet and quartet could be caused by ethanol or ethoxyethane. At this point, additional reasoning has to be done: ethanol is a much more likely candidate as it is a known product of electrochemical CO_2 reduction [43] and despite its much higher boiling temperature was identified in the gas phase [58]. This reasoning could also be implemented via a software agent but this is not feasible in this example. Finally, some minor products might exhibit small signals in Fig. 10 that overlap with peaks P_3 and P_6 as their signal pattern does not appear completely symmetrical. Another software agent was searching for such species and returned a list of possible subproducts as shown in appendix A.2.

5.2 Data enrichment

The enrichment of databases is fundamental to maintain them, as well as the consistency and accuracy of the data [33]. In this section, we demonstrate the data enrichment capabilities of our approach with few examples.

CASE 1: Assign properties from species similarities.

As mentioned previously, the ability of easily investigating properties of classes can also enable prediction of unknown values. In the simplest case, unknown properties of a molecule that is part of a group with a well-known trend along a single variable can be extrapolated.

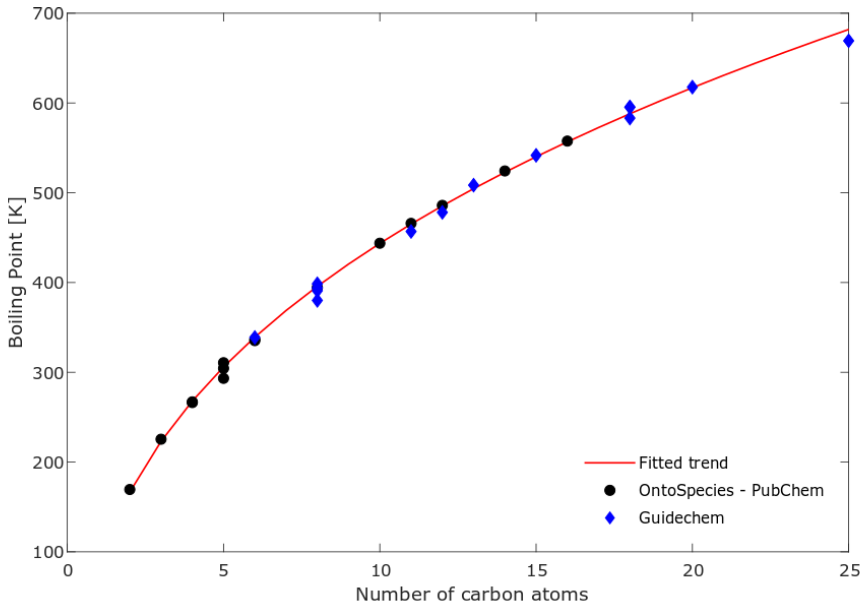


Figure 13: *Boiling points for alkanes as a function of number of carbon atoms. Black circles and blue diamonds show the experimental values taken from OntoSpecies/PubChem or Guidechem databases, respectively. Red line represents the fitted trend using a cube root function.*

For example, boiling points of alkenes as a function of number of carbon atoms follow a defined trend as shown in Fig. 7. However, in OntoSpecies there are species classified as alkenes where the boiling point is not reported. The remaining values can be extrapolated fitting the trend of alkene boiling points taken from OntoSpecies/PubChem with a cube root function using an agent.

The extrapolated boiling point trends (red line in Fig. 13) have been compared with experimental values taken from Guidechem database [27] (blue diamonds in Fig. 13) and they give a good prediction of alkene boiling points as shown in Fig. 13. Obtained values are listed in Tab. A4 in appendix A.3.

The extrapolated boiling points can be instantiated and assigned to the respective species in OntoSpecies by the agent using a SPARQL update query.

CASE 2: Find molecules that miss classification tags.

Classification in OntoSpecies is taken from ChEBI. ChEBI is manually maintained and as such cannot easily scale to the full scope of public chemical data. Many compounds in PubChem databases are not reported in ChEBI so they miss the classification tag. As an example, if we query for species that are classified as “alkene” in OntoSpecies only 35 species are returned. However, if we consider that alkenes are molecules with sum formula C_xH_{2x} that also have one double bond and we apply this constraint in our query we find that 78 species miss the “alkene” classification tag. We manually checked that all the returned species (e.g., pent-2-ene, dec-4-ene, etc.) were actually belonging to the alkene class of molecules. The chemical class IRI representing “alkene” molecules can then be linked to these species using the SPARQL update query in Fig. 14.

```
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#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

INSERT DATA {?speciesIRI os:hasChemicalClass ?classIRI}
WHERE {
   # select a species and with chemical formula CxH2x
   ?speciesIRI rdf:type os:Species ; rdfs:label ?formula .
   FILTER(regex(?formula, "C[0-9]{0,10}H[0-9]{0,10}$"))
   FILTER(xsd:float(strbefore(strafter(str(?formula),'C'),'H')) =
   ↪  1/2*xsd:float(strafter(str(?formula),'H')))

   # select species that have a double bond
   ?speciesIRI os:hasAtomicBond ?bond . ?bond os:hasBondOrder 2 .

   # select species that miss the classification tag "alkene"
   FILTER NOT EXISTS{
      ?speciesIRI (rdf:|!rdf:)* ?x .
      ?x ?y ?z .
      ?z (rdf:|!rdf:)* ?classIRI .
      ?classIRI rdfs:label "alkene" . }
}
```

Figure 14: *SPARQL query that find alkene species without classification and link to them the chemical class IRI that represents “alkene” molecules.*

This improves the reusability and findability of the data, and it also allows inference through the navigation of the ChEBI hierarchy (i.e. a species is an alkene therefore a hydrocarbon).

We can also check for inconsistency of data. We indeed noted that there is a compound that does not have chemical formula of the type C_xH_{2x} but it is tagged as “alkene” in ChEBI: 7,11,15-trimethyl-3-methylene-hexadec-1-ene with a sum formula of $C_{20}H_{38}$ which is indeed an alkadiene as it exhibits two double bonds.

CASE 3: Introduce new concepts of chemical classes.

Some chemical classes are not included in ChEBI, such as the “alkanol” class. Alkanol species are alcohols that have a chemical formula of the type $C_xH_{2x+2}O$. The use of ChEBI classification in combination with a filter on the chemical formula as previously done in section 5.1 permits to select such species using a SPARQL query. The alkanol species returned by the query (e.g., ethanol, propan-1-ol, propan-2-ol, etc.) are manually verified to ensure the accuracy of the classification. A new chemical class labeled "alkanol" can now be linked to these species and assigned as subclass of the class labeled "alcohol" using a SPARQL update query. This demonstrates the ability of our approach to add knowledge that is not included in the data sources.

CASE 4: Check coherence of data reported in PubChem.

PubChem gathers data from different sources. Identification of discrepant data in aggregated databases is a key step in data curation and remediation. PubChem data is stored in the form of strings. To get data from PubChem and put it in the ontological format, an agent is used to process the strings and get the numerical value and its related unit. Different strings for the same properties might be present in the PubChem record, so the agent also selects a recommended value as described in section 4. This exercise also serves to identify properties of a species with discrepancies between reported values. Using OntoSpecies this can be done using a query that identifies species with a specific property values that is 20% higher or lower than the recommended value. Fig. 15 reports the query example for boiling point.

For example, methyl acetate is selected as a species with discrepancies in the boiling point values. The list of boiling points of methyl acetate taken from PubChem is reported in Tab. 4. If we look at the boiling point values we see that one is much higher than the other ones (red entry in Tab. 4), due to most likely a typo in the original string. All the other values are similar to the one picked by the agent as recommended value (blue entry in Tab. 4). As PubChem collects its data from several external sources, properties are linked to their respective sources through the predicate `os:hasProvenance`. We can then find out that the discrepant entry is taken from Human Metabolome Database (HMDB) (http://www.hmdb.ca/metabolites/HMDB0031523). Preserving the provenance of information is crucial for ensuring the integrity, reliability, and usability of information over time. It helps to promote transparency and to facilitate collaboration and sharing [33].

```
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 ?speciesIRI ?iupacstring ?BPvalue1 ?BPvalue2
WHERE{
   ?speciesIRI rdf:type os:Species ;os:hasIUPACName ?iupac .
   ?iupac os:value ?iupacstring .
   FILTER(regex(str(?iupac) ,'_1_'))

   ?speciesIRI os:hasBoilingPoint ?BP1 .
   ?BP1 os:value ?BPvalue1 ;
      os:hasProvenance ?provenance .
   ?speciesIRI os:hasBoilingPoint ?BP2 .
   ?BP2 os:value ?BPvalue2 .

   ?provenance rdfs:label "PubChem agent" .

   FILTER( ?BPvalue2/?BPvalue1 > 1.2 || ?BPvalue2/?BPvalue1 < 0.8 )
}
```

Figure 15: *SPARQL query selecting boiling point values of species that are 20% higher or lower than the recommended value.*

Table 4: *Boiling point values listed in PubChem for methyl acetate. The blue entry is selected as recommended value by the software agent. The discrepant entry is coloured in red.*

os:BoilingPoint			
index	os:originalDataString	os:value	os:unit
1	134.4 °F at 760 mmHg (NTP, 1992)	330.03888	K
2	56.7 °C	329.85	K
3	535.70 °C. @ 760.00 mm Hg (est)	808.85	K
4	57 °C	330.15	K
5	135 °F	330.37222	K

CASE 5: Navigate TWA chemistry domain with federated queries.

The inherent interoperability of TWA and the according accessibility of other domains and subdomains that are linked to OntoSpecies enables data enrichment. For example, in the OntoCompChem ontology, information on the input and output of density functional theory (DFT) calculation are reported. Every conducted computational chemistry calculation (`occ:G09`) is linked to a species in OntoSpecies by the predicate (`occ:hasUniqueSpecies`). Information like optimized geometry, HOMO-LUMO gap and rotational constants can then be accessed through a federated query. Fig. 16 shows a federated query that returns the HOMO-LUMO gaps and the level of theory and basis set of the DFT calculations performed for the carbon dioxide species.

Four calculations on HOMO-LUMO gaps are stored in OntoCompChem for carbon dioxide and they are listed in Tab. 5.

Other information, like the outputs from a vibrational frequency calculation can be obtained in the same way and processed by an agent to calculate other relevant properties such as heat capacities, entropy, and enthalpy [42].

```
PREFIX occ: <http://www.theworldavatar.com/ontology/ontocompchem/ontocompchem.owl#>
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#>
PREFIX gc: <http://purl.org/gc/>

SELECT DISTINCT ?formula ?leveloftheory ?basisset ?homolumo
WHERE{
  # select CO2 species
  ?speciesIRI rdf:type os:Species ; rdfs:label ?formula .
  FILTER(str(?formula) = 'CO2')
  # retrive HOMO-LUMO info from OntoCompChem
  SERVICE <http://www.theworldavatar.com/blazegraph/namespace/ontocompchem/sparql> {
     ?calculation occ:hasUniqueSpecies ?speciesIRI ; gc:isCalculationOn ?lumo ;
     ↪  gc:isCalculationOn ?homo ; occ:hasInitialization ?i .
     ?i gc:hasParameter ?p1 ; gc:hasParameter ?p2 .
     ?p1 gc:hasBasisSet ?basisset .
     ?p2 occ:hasLevelOfTheory ?leveloftheory .
     ?lumo rdf:type occ:LumoEnergy ; occ:hasLumoEnergy ?lumoen .
     ?lumoen gc:hasValue ?lv .
     ?homo rdf:type occ:HomoEnergy ; occ:hasHomoEnergy ?homoen .
     ?homoen gc:hasValue ?hv .
     BIND( (xsd:float(?hv) - xsd:float(?lv))*27.2114 AS ?homolumo )
  }
}
```

Figure 16: *Federated SPARQL query selecting the HOMO-LUMO gaps, and the level of theory and basis set of the DFT calculations performed to calculate those on the carbon dioxide species.*

Table 5: *Level of theory and basis set of the DFT calculations performed on the carbon dioxide species and the calculated HOMO-LUMO gaps obtained as results of the federated query.*

index	level of theory	basis set	HOMO-LUMO gap [eV]
1	RB3LYP	6-31G(d,p)	-6.0270534
2	RB3LYP	6-31G(d,p)	-10.344957
3	RB3LYP	6-311G(d,p)	-11.055719
4	RB3LYP	6-311+G(d,p)	-9.965902

5.3 Automation of research laboratories

The automation of research activities in chemical laboratories plays an important role in accelerating scientific discovery and saving resources. The term “lab automation” is commonly used to describe the development of platforms comprising of robotic handlers that can carry out simple synthesis experiments and analyse results, driven by data-driven algorithms [59]. These activities aim to realise a “self-driving laboratory” but fundamental challenges around data availability, human-in-the-loop and decision-making remain [74]. We argue, that current approaches are not sufficient to meet these challenges and might even limit further development [2]: In order to represent and automate all aspects of experimental management, planning, execution, analysis, and reporting, the embedding of deep chemical knowledge (as present in OntoSpecies) is necessary. To achieve this, siloed nature of existing systems has to be overcome [55], integrating different data sources and aspects of digitalisation such as Lab Inventory Management Systems (LIMS) and

Electronic Lab Notebooks (ELN) [33, 73].

Organisational and management tasks are critical for smooth operations in a research lab and most importantly, ensure human safety. For example, the linking of GHS hazard statements to respective chemicals (see Sec. 3.1) is essential for automating chemical hazard identification and risk assessment. Moreover, the availability of big data on chemical structures and associated properties can help in predicting potential products' characteristics, such as toxicology [45]. We demonstrated in section 5.1, how OntoSpecies can be used in multi-objective experimental planning tasks such as solvent selection.

Accurate representation of experimental steps also requires access to tacit chemical knowledge. This does not only increase experimental repeatability but most importantly enables advanced experimental design algorithms to fully leverage all available data to optimise reaction conditions [2, 28]. In the analytical stage after execution of an experiment, i.e. the actual performance of the laboratory test, OntoSpecies can help in tasks like identification of products by NMR spectra (see Sec. 5.1).

In the post-analytical stage, OntoSpecies in combination with other ontology domains of TWA will help sample management, and reporting. As an example, the use of sensors has been designed to archive tested samples, reducing errors from mislabeling or incorrect storage. Tracking results, samples or assets is inherently straightforward as the uniqueness of identifiers (IRIs) is built in by design. Standardised reporting of procedures, observations, and results, which is considered a key challenge [28, 73], will enabled by the usage of interconnected dynamic knowledge graphs. By connecting more and more relevant domain ontologies, we are able to represent and subsequently automate research-related tasks of increasing complexity - working towards the vision of an "AI scientist" that can make Nobel-worthy discoveries, an idea that has been recently introduced in this context [40].

5.4 Implementation limits

Although OntoSpecies makes the gathering and processing of chemical data easier compared to non-semantic databases, we identified some weaknesses:

1. To submit a query, the user needs to know the SPARQL language and how the knowledge has been structured in the KG. In some cases of user error, the queries return no data but are considered formally correct, and no warnings are reported. For this reason, a more user-friendly interface with the KG is desired. In TWA context, "Marie", a question-answering interface that allows the users to type their questions in their natural language, has been developed [78, 79]. However, "Marie" is not acting on the full scope of TWA yet and is currently under further development.

2. The results of the queries reported in the use cases (see section 5.1 and 5.2) depend on the species that have been already instantiated in OntoSpecies. This means that if a species has not been instantiated yet, but satisfies the requirements of the query, it won't be shown in the result, leaving out species that can be important in the query context. However, OntoSpecies is automatically growing instantiating many new species every day and those can be selected based on use cases.

6 Conclusions

This paper introduces OntoSpecies, a web ontology designed to semantically represent chemical species and their properties. It serves as core ontology in the TWA chemistry domain. Compared to its previous implementation, it has been extended to include a wide range of identifiers, chemical and physical properties, chemical classifications and applications, plus spectral information associated with each species, and the provenance and attribution metadata, making it the most comprehensive semantic databases on chemical species. The information on a chemical species are collected from the respective PubChem data on the compound using a software agent. In this way, the ontology is enriched with a vast amount of chemical information, resulting in a comprehensive and reliable source of chemical data that can be accessed through a SPARQL endpoint.

We believe that our approach represents a significant advancement in the field of chemical data management. It offers a standardised way to represent chemical data and provides a powerful means for navigating and analysing chemical information in a way that is not possible using traditional databases technologies, making it a valuable tool for scientific research. The ontological format permits advanced queries and easy data analysis and visualisation. To demonstrate the usefulness of our approach, we presented several use cases that showcase how OntoSpecies can be used to compare chemical properties of similar compounds, find compounds with required characteristics, and automate laborious data gathering by researchers. We show how complex tasks such as the identification of species in an unknown mixture based on NMR measurements, the selection of suitable solvents based on multiple criteria, or the investigation of trends in chemical properties can be addressed using SPARQL queries in combination with the use of software agents to process the information obtained. We also show how the ontological format is beneficial to maintain and enrich the data as well as to check its consistency and accuracy. Finally, the link between OntoSpecies and other ontologies in TWA is discussed in the context of laboratory automation and cross-domain applications in TWA ecosystem.

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. Part of this work was supported by Towards Turing 2.0 under the EPSRC Grant EP/W037211/1 and The Alan Turing Institute. S. D. Rihm acknowledges financial support from Fitzwilliam College, Cambridge, and the Cambridge Trust. M. Kraft gratefully acknowledges the support of the Alexander von Humboldt Foundation. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Nomenclature

ABox	Assertion Component (of an ontology)
AI	Artificial Intelligence
CBU	Chemical Building Unit
ChEBI	Chemical Entities of Biological Interest
CHEMINF	Chemical Information Ontology
CHMO	Chemical Methods Ontology
DFT	Density Functional Theory
DMSO	Dimethyl Sulfoxide
DOI	Digital Object Identifier
ELN	Electronic Laboratory Notebook
FAIR	Findability, Accessibility, Interoperability, and Reusability
IDSM	Integrated Database of Small Molecules
InChI	International Chemical Identifier
IRI	Internationalised Resource Identifier
IUPAC	International Union of Pure and Applied Chemistry
KG	Knowledge Graph
LIMS	Laboratory Information Management System
ML	Machine Learning
MOP	Metal-Organic Polyhedra
MS	Mass-Spectrometry
NMR	Nuclear Magnetic Resonance
OWL	Web Ontology Language
PES	Potential Energy Surface
RDF	Resource Description Framework
SCF	Self-Consistent Field
SMILES	Simplified Molecular-Input Line-Entry System

SPARQL SPARQL Protocol and RDF Query Language

TBox Terminology Component (of an ontology)

TWA The World Avatar (project)

URL Uniform Resource Locator

UUID Universal Unique Identifier

WSDL Web Services Description Language

A Appendix

A.1 Classes in OntoSpecies

A list of all classes in OntoSpecies and their description is reported in Tab. A1 in alphabetical order. The table also includes the parent class and the corresponding CHEMINF or CHMO equivalent class when applicable. A class is linked to the parent class and equivalent class through the predicates `rdfs:subClassOf` and `owl:equivalentClass`, respectively.

Table A1: *List of classes in OntoSpecies and their description. Parent class and equivalent class are also reported when applicable.*

Class	Description
`os:11BNMRSpectra`	Boron-11 NMR spectroscopy (also known as 11B NMR) is a version of NMR spectroscopy used to elucidate the structure of boron-containing compounds. • Subclass of `1DNMRSpectra` • Equivalent to `CHMO_0000843`
`os:13CNMRSpectra`	Carbon-13 NMR spectroscopy (also known as 13C NMR) is a version of NMR spectroscopy used to elucidate the structure of carbon-containing compounds. • Subclass of `1DNMRSpectra` • Equivalent to `CHMO_0000837`
`os:15NNMRSpectra`	Nitrogen-15 NMR spectroscopy (also known as 15N NMR) is a version of NMR spectroscopy used to elucidate the structure of nitrogen-containing compounds. • Subclass of `1DNMRSpectra` • Equivalent to `CHMO_0000844`
`os:17ONMRSpectra`	Oxygen-17 NMR spectroscopy (also known as 17O NMR) is a version of NMR spectroscopy used to elucidate the structure of oxygen-containing compounds. • Subclass of `1DNMRSpectra` • Equivalent to `CHMO_0001189`
`os:19FNMRSpectra`	Fluorine-19 NMR spectroscopy (also known as 19F NMR) is a version of NMR spectroscopy used to elucidate the structure of fluorine-containing compounds. • Subclass of `1DNMRSpectra` • Equivalent to `CHMO_0000845`
`os:29SiNMRSpectra`	Silicon-29 NMR spectroscopy (also known as 29Si NMR) is a version of NMR spectroscopy used to elucidate the structure of silicon-containing compounds. • Subclass of `1DNMRSpectra` • Equivalent to `CHMO_0001955`
`os:31PNMRSpectra`	Phosphorus-31 NMR spectroscopy (also known as 31P NMR) is a version of NMR spectroscopy used to elucidate the structure of phosphorus-containing compounds. • Subclass of `1DNMRSpectra` • Equivalent to `CHMO_0000839`

os:1DNMRSpectra	One-dimensional NMR spectra. • Subclass of NMRSpectra • Equivalent to CHMO_0001928
os:1H13CNMRSpectra	Two-dimensional 1H-13C NMR spectra. • Subclass of 2DNMRSpectra • Equivalent to CHMO_0000890
os:1H1HNMRSpectra	Two-dimensional 1H-1H NMR spectra. • Subclass of 2DNMRSpectra • Equivalent to CHMO_0002420
os:1HNMRSpectra	Hydrogen-1 NMR spectroscopy (also known as H1 NMR or proton NMR) is a version of NMR spectroscopy used to elucidate the structure of hydrogen-containing compounds. • Subclass of 1DNMRSpectra • Equivalent to CHMO_0002419
os:2DNMRSpectra	Two-dimensional NMR spectroscopy is a set of nuclear magnetic resonance spectroscopy (NMR) methods, which give data plotted in a space defined by two frequency axes. • Subclass of NMRSpectra • Equivalent to CHMO_0000932
gc:Atom	A chemical entity constituting the smallest component of an element having the chemical properties of the element.
os:AtomChiralCount	Atom stereocenter (atom that is related to four distinct atoms) count. • Subclass of os:Property • Equivalent to CHEMINF_000205
os:AtomChiralDefCount	Defined atom stereocenter (atom that is related to four distinct atoms) count. • Subclass of os:Property • Equivalent to CHEMINF_000206
os:AtomChiralUndefCount	Undefined atom stereocenter (atom that is related to four distinct atoms) count. • Subclass of os:Property • Equivalent to CHEMINF_000212
os:AtomicBond	Bond between two atoms. • Equivalent to CHEMINF_000063
os:AtomicRadius	Radius of an atom. • Subclass of os:Property • Equivalent to CHEMINF_000125
os:AtomicWeight	Mass of an atom. • Subclass of os:Property • Equivalent to CHEMINF_000084
os:AutoignitionTemperature	The lowest temperature at which the substance will spontaneously ignite in a normal atmosphere without an external source of ignition (e.g., spark or flame). • Subclass of Property • Equivalent to CHEMINF_000444
os:BoilingPoint	The temperature at which this compound changes state from liquid to gas at a given atmospheric pressure. • Subclass of ThermoProperty • Equivalent to CHEMINF_000257

os:BondChiralCount	Bond stereocenter count. • Subclass of os:Property • Equivalent to CHEMINF_000213
os:BondChiralDefCount	Defined bond stereocenter count. • Subclass of os:Property • Equivalent to CHEMINF_000214
os:BondChiralUndefCount	Undefined bond stereocenter count. • Subclass of os:Property • Equivalent to CHEMINF_000215
os:Caco2Permeability	Caco-2 (Cancer coli-2) is a human colon epithelial cancer cell line (established from human colorectal adenocarcinoma cells). It is primarily used as a model of the intestinal epithelial barrier. The Caco-2 permeability of a chemical is used as a measure of its intestinal absorption in human. • Subclass of Property
os:CanonicalizedCompound	Indicate if the compound is canonicalized. • Subclass of os:Property
os:Charge	Total charge of a chemical entity. • Subclass of os:Property • Equivalent to CHEMINF_000131
os:ChebiID	Database identifier used by ChEBI. • Subclass of Identifier • Equivalent to CHEMINF_000407
os:ChemicalClass	Chemical classes are groupings that relate chemicals by similar features. Chemicals can be classified by their structure (e.g., hydrocarbons), uses (e.g., pesticides), physical properties (e.g., volatile organic compounds [VOCs]), radiological properties (e.g., radioactive materials), or other factors. ChEBI Ontology tree. ChEBI is an acronym for Chemical Entities of Biological Interest, which is a freely available dictionary of molecular entities focused on 'small' chemical compounds. ChEBI incorporates an ontological classification, whereby the relationships between molecular entities or classes of entities and their parents and/or children are specified. • Subclass of Classification
os:CID	Database identifier used by PubChem. • Subclass of Identifier • Equivalent to CHEMINF_000140
os:Classification	A set of concepts and categories in a subject area or domain that shows their properties and the relations between them.
os:CollisionCrossSection	Collision cross section represents the effective area for the interaction between an individual ion and the neutral gas through which it is traveling (e.g., in ion mobility spectrometry experiments). It quantifies the probability of a collision taking place between two or more particles. • Subclass of Property
os:CompoundComplexity	Indicator that denotes how complicated a structure is. • Subclass of os:Property • Equivalent to CHEMINF_000390

`os:CovalentUnitCount`	The number of covalent units in a chemical structure. • Subclass of `os:Property` • Equivalent to `CHEMINF_000280`
`os:Density`	Density (with unit) and specific gravity (without unit) of a compound. Density is mass of a unit volume of a compound and commonly expressed in units of kg/m3 or g/cm3. Specific gravity, also known as relative density, is a unit-less quantity, defined as the ratio of the density of a compound to that of a standard reference material (typically, water at 4 °C for liquids and air at room temperature [20 °C or 68 °F] for gases). • Subclass of `ThermoProperty` • Equivalent to `CHEMINF_000416`
`os:DissociationConstants`	A specific type of equilibrium constant that measures the propensity of a larger object to separate (dissociate) reversibly into smaller components, as when a complex falls apart into its component molecules, or when a salt splits up into its component ions. This includes pKa (the negative logarithm of the acid dissociation constant) and pKb (the negative logarithm of the base dissociation constant). • Subclass of `ThermoProperty`
`os:ElectronAffinity`	Amount of energy released when an electron attaches to a neutral atom or molecule in the gaseous state to form an anion. • Subclass of `os:Property`
`os:ElectronConfiguration`	Arrangement of electrons in orbitals around an atomic nucleus. • Subclass of `os:Property`
`os:Electronegativity`	Electronegativity is an atomic quality that describes its power to attract electrons to itself . • Subclass of `os:Property` • Equivalent to `CHEMINF_000121`
`pt:Element`	An element in the periodic table.
`os:ElementClassification`	Classification of elements in the periodic table. • Subclass of `os:Classification`
`os:ElementGroupNumber`	Group number of an element in the periodic table. • Subclass of `os:Property`
`os:ElementName`	Name of an element in the periodic table. • Subclass of `os:Identifier`
`os:ElementPeriodNumber`	Period number of an element in the periodic table. • Subclass of `os:Property`
`os:ElementSymbol`	Symbol of an element in the periodic table. • Subclass of `os:Identifier`
`os:EnthalpyOfSublimation`	The enthalpy (or heat) of sublimation is the amount of energy that must be added to a mole of solid at constant pressure to turn it directly into a gas (without passing through the liquid phase). • Subclass of `ThermoProperty`

os:ExactMass	Mass of the most intense molecule peak in an MS spec, and when calculated denotes the mass of a molecule containing most likely isotopic composition for a single random molecule • Subclass of os:Property • Equivalent to CHEMINF_000217
os:FlashPoint	The lowest temperature at which a liquid can gives off vapor to form an ignitable mixture in air near the surface of the liquid. • Subclass of ThermoProperty • Equivalent to CHEMINF_000417
os:Frequency	Frequency of the spectrometer.
os:FunctionalGroup	Specific groups of atoms within molecules that are responsible for the characteristic chemical reactions of those molecules. • Equivalent to CHEMINF_000068
os:GCMS	Data from gas chromatography-mass spectrometry (GC-MS) experiments. • Subclass of os:MassSpectrometry • Equivalent to CHMO_0000497
os:Geometry	Geometry of a molecule.
os:GHSHazardStatement	GHS (Globally Harmonized System of Classification and Labelling of Chemicals) is a United Nations system to identify hazardous chemicals and to inform users about these hazards. GHS has been adopted by many countries around the world and is now also used as the basis for international and national transport regulations for dangerous goods. • Subclass of Classification
os:GroundLevel	Ground level of an element in the periodic table. • Subclass of os:Property
os:HeatOfCombustion	The heat of combustion is the energy released as heat when a compound undergoes complete combustion with oxygen under standard conditions. • Subclass of ThermoProperty
os:HeatOfVaporization	The heat (or enthalpy) of vaporization is the quantity of heat that must be absorbed if a certain quantity of liquid is vaporized at a constant temperature. • Subclass of ThermoProperty • Equivalent to CHEMINF_000418
os:HeavyAtomCount	The number of non-hydrogen atoms. • Subclass of os:Property • Equivalent to CHEMINF_000300
os:HenrysLawConstant	Henry's law states that the amount of dissolved gas (in liquid, such as water) is proportional to its partial pressure in the gas phase. The proportionality factor is called the Henry's law constant and defined as the ratio of a compound's partial pressure in air to the concentration of the compound in water at a given temperature. • Subclass of ThermoProperty • Equivalent to CHEMINF_000433

`os:HydrogenBondAcceptorCount`	Number of hydrogen bond acceptors in a given molecular entity. This is usually the count of all negatively or partially negatively charged heteroatoms (e.g. alcohol oxygen) capable of accepting a hydrogen bond. • Subclass of `os:Property` • Equivalent to `CHEMINF_000245`
`os:HydrogenBondDonorCount`	Number of hydrogen bond donors in a given molecular entity. This is usually the count of all negatively or partially negatively charged heteroatoms (e.g. alcohol oxygen) that have covalently attached to them partially positively charged hydrogen atoms that are capable of participating in a hydrogen bond. • Subclass of `os:Property` • Equivalent to `CHEMINF_000244`
`os:Hydrophobicity`	Hydrophobicity is the physical property of a molecule that is seemingly repelled from a mass of water. • Subclass of `Property`
`os:Identifier`	Chemical names, synonyms, identifiers, and descriptors. • Equivalent to `CHEMINF_000061`
`os:InChI`	The IUPAC International Chemical Identifier (InChI) is a textual identifier for chemical substances, designed to provide a standard and human-readable way to encode molecular information and to facilitate the search for such information in databases and on the web. • Subclass of `Identifier` • Equivalent to `CHEMINF_000113`
`os:InChIKey`	The InChIKey is a fixed length (27 character) condensed digital representation of the InChI that is not human-understandable. • Subclass of `Identifier` • Equivalent to `CHEMINF_000059`
`os:InstrumentType`	Type of instrument used for the spectrometry.
`os:IonizationMode`	Ionization mode used in the mass spectrometry analysis.
`os:IonizationPotential`	Ionization potential, also called ionization energy, is the amount of energy required to remove an electron from an isolated atom or molecule. • Subclass of `Property` • Equivalent to `CHEMINF_000191`
`os:IsoelectricPoint`	The isoelectric point, sometimes abbreviated to IEP, is the pH at which a particular molecule or surface carries no net electrical charge. • Subclass of `Property`
`os:IsotopeAtomCount`	The sum of all atoms enriched with respect to a particular atom isotope. • Subclass of `os:Property` • Equivalent to `CHEMINF_000301`
`os:IUPACName`	An IUPAC name is a systematic name which is formulated according to the rules and recommendations for chemical nomenclature set out by the International Union of Pure and Applied Chemistry (IUPAC). • Subclass of `Identifier` • Equivalent to `CHEMINF_000107`

os:LCMS	Data from liquid chromatography-mass spectrometry (LC-MS) experiments. • Subclass of os:MassSpectrometry • Equivalent to CHMO_0000524
os:LogP	Log P is the partition coefficient expressed in logarithmic form. The partition coefficient is the ratio of concentrations of a compound in a mixture of two immiscible solvents at equilibrium. This ratio is therefore used to compare the solubilities of the solute in these two solvents. Because octanol and water are the most commonly used pair of solvents for measuring partition coefficients, the Log P values listed in this section refer to octanol/water partition coefficients, unless indicated otherwise. • Subclass of Property • Equivalent to CHEMINF_000251
os:LogS	The base-10 logarithm of the aqueous solubility of this compound. • Subclass of Property
os:MALDI	MALDI (matrix-assisted laser desorption/ionization) is an ionization technique that uses a laser energy absorbing matrix to create ions from large molecules with minimal fragmentation. • Subclass of os:MassSpectrometry • Equivalent to CHMO_0002203
os:MSMS	Data from tandem mass spectrometry (MS-MS) experiments. • Subclass of os:MassSpectrometry • Equivalent to CHMO_0000701
os:MassSpectrometry	Mass spectrometry (MS or mass spec) is a technique to determine molecular structure through ionization and fragmentation of the parent compound into smaller components. • Subclass of SpectralInformation • Equivalent to CHMO_0000470
os:MeltingPoint	The melting point is the temperature at which a substance changes state from solid to liquid at atmospheric pressure. When considered as the temperature of the reverse change (from liquid to solid), it is referred to as the freezing point. • Subclass of ThermoProperty • Equivalent to CHEMINF_000256
os:MolecularFormula	A molecular formula is a structure descriptor which identifies each constituent element by its chemical symbol and indicates the number of atoms of each element found in each discrete molecule of that compound. • Subclass of Identifier • Equivalent to CHEMINF_000042
os:MolecularWeight	Mass of a molecule. • Subclass of os:Property • Equivalent to CHEMINF_000216

os:MonoIsotopicWeight	The mass of a molecule calculated using the mass of the most abundant isotope of each element. E.g., Carbon has a monoisotopic mass of 12.000 g/mol. • Subclass of os:Property • Equivalent to CHEMINF_000218
os:NMRSpectra	Nuclear magnetic resonance spectrum. • Subclass of SpectralInformation • Equivalent to CHMO_0000835
os:OpticalRotation	Optical rotation is a property of chiral substances that is expressed as the angle to which the material causes polarized light to rotate at a particular temperature, wavelength, and concentration. • Subclass of ThermoProperty • Equivalent to CHMO_0002818
os:OtherMS	Other compound's mass spectrometry (MS) information. • Subclass of os:MassSpectrometry
os:OxidationStates	Oxidation states of an element in the periodic table. • Subclass of os:Property
os:Peak	A peak of the spectrum.
os:PolarSurfaceArea	The polar surface area is defined as the combined surface area belonging to oxygen and nitrogen atoms and hydrogen atoms bound to these electronegative atoms. • Subclass of os:Property • Equivalent to CHEMINF_000307
os:Property	Chemical or physical property of a chemical species or element in the periodic table.
okin:Reference	Provenance of data.
os:ReferenceState	Reference state of a thermodynamic property.
os:SMILES	A SMILES is a structure descriptor that denotes a molecular structure as a graph. • Subclass of Identifier • Equivalent to CHEMINF_000018
os:Solubility	The solubility of a substance is the amount of that substance that will dissolve in a given amount of solvent. The default solvent is water, if not indicated. • Subclass of ThermoProperty • Equivalent to CHEMINF_000258
os:Solvent	Solvent used in the NMR analysis.
os:Species	Various chemical and physical properties that are experimentally determined for this compound. See also the Safety and Hazard Properties section (if available), which has additional properties pertinent to chemical safety and hazards.
os:SpectraGraph	Spectral data collected in a graph.
os:SpectralInformation	Spectral data for this compound, including 1-D and 2-D NMR, Infrared (IR), Raman, and Ultraviolet (UV) spectroscopy, mass spectrometry (MS), and chromatography. • Equivalent to CHMO_0000800
os:SubStructureKeysFingerprint	• Subclass of os:Property

os:SurfaceTension	Surface tension is a contractive tendency of the surface of a liquid that allows it to resist an external force. It is measured as the energy required to increase the surface area of a liquid by a unit of area. • Subclass of ThermoProperty • Equivalent to CHEMINF_000420
os:TautomerCount	Count of tautomers. • Subclass of os:Property • Equivalent to CHEMINF_000202
os:ThermoProperty	Thermodynamic property of a chemical species. • Subclass of Property
om:Unit	Major uses of this chemical, including both consumer uses and industrial uses.
os:Use	Application and role of a chemical species.
os:VaporDensity	The density of a gas or vapor relative to that of the reference gas. While some resources use the hydrogen gas as the reference gas for the vapor density calculation, many resources (particularly in relation to safety considerations at commercial and industrial facilities in the U.S.) defines the vapor density with respect to the density of air, which has an arbitrary value of one. If a gas has a vapor density of less than one it will generally rise in air. If the vapor density is greater than one the gas will generally sink in air. • Subclass of ThermoProperty • Equivalent to CHEMINF_000440
os:VaporPressure	Vapor pressure (or equilibrium vapor pressure) is the pressure of a vapor in thermodynamic equilibrium with its condensed phases in a closed system. • Subclass of os:ThermoProperty • Equivalent to CHEMINF_000419
os:Viscosity	Viscosity is a measure of a fluid's resistance to flow. It describes the internal friction of a moving fluid. • Subclass of os:ThermoProperty
os:XCoordinate	Atom X coordinate in space.
os:YCoordinate	Atom Y coordinate in space.
os:ZCoordinate	Atom Z coordinate in space.

A.2 Queries results

USE CASE 2: list of suitable co-solvents

Tab. A2 reports a list of co-solvents for propan-2-ol to enable the easiest separation by distillation. First column reports a list of species selected by the SPARQL query in Fig. 9 as suitable co-solvents (49 species). The second column reports a check-mark if the species follow criterion 1; the third column reports a check-mark if the species follow criteria 1 and 2 together; the firth column reports a check-mark if the species follow criteria 1, 2 and 3 together. Species that follow criterion 1 but are discarded by criterion 2 are highlighted in red (4 species). Species that follow all the criteria are highlighted in green (13 species).

Table A2: *List of co-solvents for propan-2-ol to enable the easiest separation by distillation*

Species	Criterion 1	Criteria 1-2	Criteria 1-2-3
2,2-bis(hydroxymethyl)propane-1,3-diol	✓	✓	×
2,6-dimethylheptan-4-ol	✓	✓	×
2-(2-butoxyethoxy)ethanol	✓ [10]	✓	× [10]
2-(2-ethoxyethoxy)ethanol	✓	✓	×
2-(2-hydroxyethoxy)ethanol	✓	✓	×
2-(2-methoxyethoxy)ethanol	✓	×	-
2-(4-methylcyclohex-3-en-1-yl)propan-2-ol	✓	✓	×
2-[2-(2-hydroxyethoxy)ethoxy]ethanol	✓	✓	×
2-[2-(2-methoxyethoxy)ethoxy]ethanol	✓	✓	×
2-[2-[2-(2-hydroxyethoxy)ethoxy]ethoxy]ethanol	✓	✓	×
2-butoxyethanol	✓	✓	×
2-butyloctan-1-ol	✓	✓	✓
2-ethoxyethanol	✓	×	-
2-ethyl-2-(hydroxymethyl)propane-1,3-diol	✓	✓	×
2-ethylhexan-1-ol	✓	✓	×
2-furylmethanol	✓	×	-
2-hydroxyacetic acid	✓	✓	✓
2-methoxyethanol	✓ [10]	× [7]	-
2-methylbutan-1-ol	✓	✓	✓
2-methylbutan-2-ol	✓	✓	✓
2-methylpentane-2,4-diol	✓	✓	×
2-methylpropan-1-ol	✓	✓	✓
2-phenylpropan-2-ol	✓	✓	×
3-methylbutan-2-ol	✓	✓	✓
butan-1-ol	✓ [10]	✓	✓ [10]
butan-2-ol	✓	✓	✓
butane-1,2-diol	✓	✓	×
butane-1,3-diol	✓	✓	×
butane-1,4-diol	✓	✓	×
cyclohexanol	✓	✓	×
decan-1-ol	✓	✓	✓
dodecan-1-ol	✓	✓	×
ethyl 2-hydroxypropanoate	✓	✓	×
ethylene glycol	✓ [10]	✓	× [10]
glycerol	✓	✓	×
hexadecan-1-ol	✓	✓	×
hexan-1-ol	✓	✓	×
hexane-1,6-diol	✓	✓	×
icosan-1-ol	✓	✓	×
methanol	✓ [10]	✓	✓ [10]
octadecan-1-ol	✓	✓	×
octan-2-ol	✓	✓	×
pentan-1-ol	✓	✓	✓
pentan-2-ol	✓	✓	✓
pentane-1,5-diol	✓	✓	×
phenylmethanol	✓	✓	×
propan-1-ol	✓	✓	✓
propane-1,2-diol	✓	✓	×
propane-1,3-diol	✓	✓	×

USE CASE 3: subproducts identification

In the NMR spectra in Fig. 10, the triplet signals at 1.04 ppm do not exactly follow the expected 1:2:1 pattern (i.e. 25%:50%:25%). Calculated peak ares give 30%:42%:27%, which indicates that another signal might overlap on the left side of the triplet in Fig. 10. Also the signals of the quartet at 3.51 ppm do not exactly follow the expected 1:3:3:1 pattern (i.e. 13%:37%:37%:13%). Calculated peak ares give 12%:33%:38%:17%, which indicates that another signal might overlap on the right side of the quartet in Fig. 10. A list of possible subproducts can be then found querying for species whose highest peak can interfere with these two peaks (P_3 and P_6). The SPARQL query result is reported in Tab. A3. Check-marks in columns P_3 and P_6 indicate that the species interferes with P_3 and P_6 respectively. Species that interfere with both peaks are highlighted in blue.

Table A3: *List of possible subproducts for the NMR spectra in Fig. 10.*

Species	P_3	P_6
(2S,3R)-butane-1,2,3,4-tetrol	✓	×
(2S,4R)-pentane-1,2,3,4,5-pentol	✓	×
2-ethoxyethanol	✓	✓
2-methoxy-2-methyl-propane	×	✓
2-methylbutan-2-ol	×	✓
2-methylbutanal	×	✓
2-methylbutanoic acid	×	✓
2-methylpropan-1-ol	✓	✓
2-methylpropanal	×	✓
2-methylpropanoic acid	×	✓
3-methyl-2-oxo-butanoic acid	×	✓
3-methylbutan-2-ol	×	✓
3-methylbutanoic acid	×	✓
butan-1-ol	✓	✓
butan-2-ol	✓	✓
butanal	×	✓
butane-1,3-diol	✓	✓
butane-2,3-diol	✓	✓
butyric acid	×	✓
ethyl propanoate	×	✓
ethylene glycol	✓	×
glycerol	✓	×
methanol	✓	×
methyl 2-methylpropanoate	✓	✓
methyl 3-oxobutanoate	✓	×
methyl acetate	✓	×
pent-1-en-3-ol	×	✓
pent-1-en-3-one	×	✓
pentan-1-ol	✓	✓
pentan-2-ol	×	✓
pentanal	×	✓
propan-1-ol	✓	✓
propan-2-ol	×	✓
propionic acid	×	✓
tetrahydrofuran	✓	×

A.3 Data enrichment

CASE 1: boiling points of alkenes

Table A4: *List of species classified as alkenes in OntoSpecies ordered by number of carbon atoms, their experimental boiling points ($T_{b-experimental}$) taken from PubChem or GuideChem (black or blue color respectively) and their extrapolated boiling points obtained fitting PubChem data with a cube root function ($T_{b-predicted}$).*

Species		$T_{b-experimental}$ [K]	$T_{b-predicted}$ **[K]**
ethylene	C_2H_4	169.43	166.57
prop-1-ene	C_3H_6	225.43	223.04
2-methylprop-1-ene	C_4H_8	266.26	267.99
but-1-ene	C_4H_8	267.04	267.99
2-methylbut-1-ene	C_5H_{10}	304.35	305.94
2-methylbut-2-ene	C_5H_{10}	310.65	305.94
3-methylbut-1-ene	C_5H_{10}	293.25	305.94
(E)-hex-3-ene	C_6H_{12}	338.75	339.12
2-methylpent-1-ene	C_6H_{12}	335.26	339.12
hex-1-ene	C_6H_{12}	336.65	339.12
(E)-oct-2-ene	C_8H_{16}	398.55	395.76
(E)-oct-3-ene	C_8H_{16}	394.05	395.76
3-methyleneheptane	C_8H_{16}	390.85	395.76
4,5-dimethylhex-1-ene	C_8H_{16}	380.05	395.76
oct-1-ene	C_8H_{16}	394.43	395.76
(E)-2,6-dimethyloct-3-ene	$C_{10}H_{20}$	-	443.58
dec-1-ene	$C_{10}H_{20}$	443.76	443.58
(E)-7-methyldec-4-ene	$C_{11}H_{22}$	-	465.12
5-methyldec-1-ene	$C_{11}H_{22}$	456.85	465.12
undec-1-ene	$C_{11}H_{22}$	465.87	465.12
(E)-dodec-2-ene	$C_{12}H_{24}$	-	485.38
(E)-dodec-3-ene	$C_{12}H_{24}$	-	485.38
8-methylundec-1-ene	$C_{12}H_{24}$	478.15	485.38
dodec-1-ene	$C_{12}H_{24}$	485.93	485.38
(E)-tridec-2-ene	$C_{13}H_{26}$	508.45	504.56
tetradec-1-ene	$C_{14}H_{28}$	524.26	522.77
pentadec-1-ene	$C_{15}H_{30}$	541.65	540.13
hexadec-1-ene	$C_{16}H_{32}$	557.55	556.74
heptadec-1-ene	$C_{17}H_{34}$	-	572.67
(E)-octadec-7-ene	$C_{18}H_{36}$	595.35	587.98
(E)-octadec-9-ene	$C_{18}H_{36}$	595.35	587.98
octadec-1-ene	$C_{18}H_{36}$	-	587.98
octadec-9-ene	$C_{18}H_{36}$	583.15	587.98
7,11,15-trimethyl-3-methylene-hexadec-1-ene	$C_{20}H_{38}$	617.65	616.99
pentacos-1-ene	$C_{25}H_{50}$	669.25	681.90

References

[1] 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. *Journal of Chemical Information and Modeling*, 61:1701–1717, 4 2021. doi:10.1021/acs.jcim.0c01322.

[2] J. Bai, L. Cao, S. Mosbach, J. Akroyd, A. A. Lapkin, and M. Kraft. From platform to knowledge graph: Evolution of laboratory automation. *JACS Au*, 2:292–309, 2 2022. doi:10.1021/jacsau.1c00438.

[3] I. K. Bakulin and M. A. Orekhov. Basic principles underlying the size dependence of the hydrocarbon ionization energy. *Journal of Experimental and Theoretical Physics*, 135(5):611–616, 2022. doi:10.1134/S1063776122110012.

[4] F. Belleau, M. A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41:706–716, 10 2008. doi:10.1016/j.jbi.2008.03.004.

[5] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284:34–43, 5 2001. doi:10.1038/SCIENTIFICAMERICAN0501-34.

[6] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry*, 4:217–241, 1 2008. doi:10.1016/S1574-1400(08)00012-1.

[7] F. P. Byrne, S. Jin, G. Paggiola, T. H. M. Petchey, J. H. Clark, T. J. Farmer, A. J. Hunt, C. Robert McElroy, and J. Sherwood. Tools and techniques for solvent selection: green solvent selection guides. *Sustainable Chemical Processes*, 4(1):1–24, 2016. doi:10.1186/s40508-016-0051-z.

[8] C. Cobas. NMR signal processing, prediction, and structure verification with machine learning techniques. *Magnetic Resonance in Chemistry*, 58(6):512–519, 2020. doi:10.1002/mrc.4989.

[9] S. J. Coles, N. E. Day, P. Murray-Rust, H. S. Rzepa, and Y. Zhang. Enhancement of the chemical semantic web through the use of inchi identifiers. *Organic & Biomolecular Chemistry*, 3:1832–1834, 5 2005. doi:10.1039/B502828K.

[10] A. Curzons, D. Constable, and V. Cunningham. Solvent selection guide: a guide to the integration of environmental, health and safety criteria into the selection of solvents. *Clean Technologies and Environmental Policy*, 1(2):82–90, 1999. doi:10.1007/s100980050014.

[11] K. Degtyarenko, P. D. 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 Research*, 36:D344, 1 2008. doi:10.1093/NAR/GKM791.

[12] A. Devanand, G. Karmakar, N. Krdzavac, R. Rigo-Mariani, Y. S. F. Eddy, I. A. Karimi, and M. Kraft. OntoPowSys: A power system ontology for cross domain interactions in an eco industrial park. *Energy and AI*, 1:100008, 2020. doi:10.1016/J.EGYAI.2020.100008.

[13] A. Devanand, G. Karmakar, N. Krdzavac, F. Farazi, M. Q. Lim, Y. S. F. Eddy, I. A. Karimi, and M. Kraft. ElChemo: A cross-domain interoperability between chemical and electrical systems in a plant. *Computers & Chemical Engineering*, 156:107556, 1 2022. doi:10.1016/J.COMPCHEMENG.2021.107556.

[14] A. Eibeck, M. Q. Lim, and M. Kraft. J-Park Simulator: An ontology-based platform for cross-domain scenarios in process industry. *Computers & Chemical Engineering*, 131:106586, 12 2019. doi:10.1016/J.COMPCHEMENG.2019.106586.

[15] European Bioinformatics Institute (EMBL-EBI). ChEMBL Database, 2023 (accessed April 25, 2023). URL `https://www.ebi.ac.uk/chembl/`.

[16] European Bioinformatics Institute (EMBL-EBI). Chemical Methods Ontology, 2023 (accessed April 25, 2023). URL `https://www.ebi.ac.uk/ols/ontologies/chmo`.

[17] European Bioinformatics Institute (EMBL-EBI). Chemical Entities of Biological Interest (ChEBI) - ChEBI statistics, 2023 (accessed April 25, 2023). URL `https://www.ebi.ac.uk/chebi/statisticsForward.do`.

[18] FAIR Principles. FAIR Principles - GO FAIR, 2023 (accessed April 25, 2023). URL `https://www.go-fair.org/fair-principles/`.

[19] P. Fantke, C. Cinquemani, P. Yaseneva, J. D. Mello, H. Schwabe, B. Ebeling, and A. A. Lapkin. Transition to sustainable chemistry through digitalization. *Chem*, 7: 2866–2882, 11 2021. doi:10.1016/J.CHEMPR.2021.09.012.

[20] F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca, and M. Kraft. Ontokin: An ontology for chemical kinetic reaction mechanisms. *Journal of Chemical Information and Modeling*, 60:108–120, 1 2020. doi:10.1021/ACS.JCIM.9B00960.

[21] G. Fu, C. Batchelor, M. Dumontier, J. Hastings, E. Willighagen, and E. Bolton. PubChemRDF: Towards the semantic annotation of PubChem compound and substance databases. *Journal of Cheminformatics*, 7:1–15, 7 2015. doi:10.1186/s13321-015-0084-4.

[22] J. Galgonek and J. Vondrášek. IDSM ChemWebRDF: SPARQLing small-molecule datasets. *Journal of Cheminformatics*, 13, 12 2021. doi:10.1186/S13321-021-00515-1.

[23] H. Gao, L. T. Zhu, Z. H. Luo, M. A. Fraga, and I. M. Hsing. Machine learning and data science in chemical engineering. *Industrial & Engineering Chemistry Research*, 61:8357–8358, 6 2022. doi:10.1021/ACS.IECR.2C01788.

[24] G. V. Gkoutos, P. Murray-Rust, H. S. Rzepa, and M. Wright. Chemical markup, XML, and the World-Wide Web. 3. Toward a signed semantic chemical web of trust. *Journal of Chemical Information and Computer Sciences*, 41:1124–1130, 2001. doi:10.1021/ci000406v.

[25] G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf. The Units Ontology: a tool for integrating units of measurement in science. *Database*, 2012, 10 2012. doi:10.1093/database/bas033.

[26] A. Gogleva, D. Polychronopoulos, M. Pfeifer, V. Poroshin, M. Ughetto, M. J. Martin, H. Thorpe, A. Bornot, P. D. Smith, B. Sidders, J. R. Dry, M. Ahdesmäki, U. McDermott, E. Papa, and K. C. Bulusu. Knowledge graph-based recommendation framework identifies drivers of resistance in egfr mutant non-small cell lung cancer. *Nature Communications*, 13:1–14, 3 2022. doi:10.1038/s41467-022-29292-7.

[27] Guidechem. Guidechem Chemical Network, 2023 (accessed April 25, 2023). URL `https://www.guidechem.com/`.

[28] A. J. Hammer, A. I. Leonov, N. L. Bell, and L. Cronin. Chemputation and the standardization of chemical informatics. *Journal of the American Chemical Society*, 1(10):1572–1587, 2021. ISSN 15205126. doi:10.1021/jacsau.1c00303.

[29] J. Hastings, L. Chepelev, E. Willighagen, N. Adams, C. Steinbeck, and M. Dumontier. The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLoS One*, 6:e25513, 10 2011. doi:10.1371/journal.pone.0025513.

[30] J. Hastings, P. D. Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck. The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Research*, 41, 1 2013. doi:10.1093/nar/gks1146.

[31] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44:D1214–D1219, 2016. doi:10.1093/NAR/GKV1031.

[32] A. Howarth, K. Ermanis, and J. M. Goodman. DP4-AI automated NMR data analysis: straight from spectrometer to structure. *Chemical Science*, 11(17):4351–4359, 2020. doi:10.1039/d0sc00442a.

[33] K. M. Jablonka, L. Patiny, and B. Smit. Making the collective knowledge of chemistry open and machine actionable. *Nature Chemistry*, 14(4):365–376, 2022. doi:10.1038/s41557-022-00910-7.

[34] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. L. Novère, H. Parkinson, E. Birney, and A. M. Jenkinson. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, 30:1338–1339, 5 2014. doi:10.1093/bioinformatics/btt765.

[35] S. Kim, P. A. Thiessen, E. E. Bolton, and S. H. Bryant. PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. *Nucleic Acids Research*, 43:W605–W611, 7 2015. doi:10.1093/NAR/GKV396.

[36] S. Kim, P. A. Thiessen, and E. E. Bolton. Programmatic retrieval of small molecule information from PubChem using PUG-REST. *Methods in Pharmacology and Toxicology*, pages 1–24, 2018. doi:10.1007/7653_2018_30.

[37] S. Kim, P. A. Thiessen, T. Cheng, B. Yu, and E. E. Bolton. An update on PUG-REST: RESTful interface for programmatic access to PubChem. *Nucleic Acids Research*, 46:W563–W570, 7 2018. doi:10.1093/NAR/GKY294.

[38] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton. PubChem 2019 update: improved access to chemical data. *Nucleic acids research*, 47:D1102–D1109, 1 2019. doi:10.1093/NAR/GKY1033.

[39] S. Kim, P. A. Thiessen, T. Cheng, J. Zhang, A. Gindulyte, and E. E. Bolton. PUG-View: Programmatic access to chemical annotations integrated in PubChem. *Journal of Cheminformatics*, 11:1–11, 8 2019. doi:10.1186/S13321-019-0375-2.

[40] H. Kitano. Nobel Turing Challenge: creating the engine for scientific discovery. *npj Systems Biology and Applications*, 7(1):1–12, 2021. ISSN 20567189. doi:10.1038/s41540-021-00189-3.

[41] A. Kondinski, A. Menon, D. Nurkowski, F. Farazi, S. Mosbach, J. Akroyd, and M. Kraft. Automated rational design of metal-organic polyhedra. *Journal of the American Chemical Society*, 144:11713–11728, 7 2022. doi:10.1021/JACS.2C03402.

[42] 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. *Journal of chemical information and modeling*, 59:3154–3165, 5 2019. doi:10.1021/ACS.JCIM.9B00227.

[43] K. P. Kuhl, E. R. Cave, D. N. Abram, and T. F. Jaramillo. New insights into the electrochemical reduction of carbon dioxide on metallic copper surfaces. *Energy and Environmental Science*, 5(5):7050–7059, 2012. doi:10.1039/c2ee21234j.

[44] M. Lipfert, M. K. Rout, M. Berjanskii, and D. S. Wishart. Automated tools for the analysis of 1D-NMR and 2D-NMR spectra. In G. A. N. Gowda and D. Raftery, editors, *NMR-Based Metabolomics: Methods and Protocols*, volume 2037 of *Methods in Molecular Biology*, pages 429–449. Springer New York, New York, NY, 2019. ISBN 978-1-4939-9690-2. doi:10.1007/978-1-4939-9690-2_24.

[45] G. Markert, J. Born, M. Manica, G. Schneider, and M. R. Martinez. Chemical representation learning for toxicity prediction. *Digital Discovery*, 2023. doi:10.1039/d2dd00099g.

[46] W. Marquardt, J. Morbach, A. Wiesner, and A. Yang. *OntoCAPE: A Re-Usable Ontology for Chemical Process Engineering*. Springer, 2010.

[47] A. McNaught and A. Wilkinson. *IUPAC Compendium of Chemical Terminology*. Blackwell Scientific Publications: Oxford, UK, (the "gold book") edition, 1997.

[48] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. D. Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, and A. R. Leach. Chembl: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47:D930–D940, 1 2019. doi:10.1093/NAR/GKY1075.

[49] A. Menon, L. Pascazio, D. Nurkowski, F. Farazi, S. Mosbach, J. Akroyd, and M. Kraft. OntoPESScan: An ontology for potential energy surface scans. *ACS Omega*, 8(2):2462–2475, 2023. doi:10.1021/acsomega.2c06948.

[50] S. Mosbach, A. Menon, F. Farazi, N. Krdzavac, X. Zhou, J. Akroyd, and M. Kraft. Multiscale cross-domain thermochemical knowledge-graph. *Journal of Chemical Information and Modeling*, 60:6155–6166, 2020. doi:10.1021/acs.jcim.0c01145.

[51] P. Murray-Rust, H. S. Rzepa, S. M. Tyrrell, and Y. Zhang. Representation and use of chemistry in the global electronic age. *Organic & Biomolecular Chemistry*, 2: 3192–3203, 11 2004. doi:10.1039/B410732B.

[52] National Center for Biotechnology Information. PubChem, 2023. URL `https://pubchem.ncbi.nlm.nih.gov/`.

[53] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. Industry-scale knowledge graphs: Lessons and challenges. *Communications of the ACM*, 62(8): 36–43, 2019. doi:10.1145/3331166.

[54] Ostlund, N. S. and Sopek, M. GNVC: Gainesville Core Ontology - standard for publishing results of computational chemistry, 2023 (accessed April 25, 2023). URL `http://ontologies.makolab.com/gc/gc07.owl`.

[55] B. G. Pelkie and L. D. Pozzo. The laboratory of Babel : highlighting community needs for integrated materials data management. *Digital Discovery*, 2023. doi:10.1039/d3dd00022b.

[56] W. Phadungsukanan, M. Kraft, J. A. Townsend, and P. Murray-Rust. The semantics of chemical markup language (cml) for computational chemistry : Compchem. *Journal of Cheminformatics*, 4:15, 8 2012. doi:10.1186/1758-2946-4-15.

[57] D. Prat, J. Hayler, and A. Wells. A survey of solvent selection guides. *Green Chemistry*, 16(10):4546–4551, 2014. doi:10.1039/c4gc01149j.

[58] S. D. Rihm, M. Kovalev, A. A. Lapkin, J. W. Ager III, and M. Kraft. On the role of C4 and C5 products in electrochemical CO2 reduction via copper-based catalysts. *Energy & Environmental Science*, 16:1697–1710, 2023. doi:10.1039/D2EE03752A.

[59] M. Seifrid, R. Pollice, A. Aguilar-Granda, Z. Morgan Chan, K. Hotta, C. T. Ser, J. Vestfrid, T. C. Wu, and A. Aspuru-Guzik. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Accounts of Chemical Research*, 55(17):2454–2466, 2022. ISSN 15204898. doi:10.1021/acs.accounts.2c00220.

[60] P. Staroch. *A weather ontology for predictive control in smart homes*. PhD thesis, Vienna University of Technology, 2013.

[61] The Internet Society. RFC 3987: Internationalized Resource Identifiers (IRIs), 2023 (accessed April 25, 2023). URL `https://www.rfc-editor.org/rfc/rfc3987`.

[62] United Nations Economic Commission for Europe (UNECE). Globally harmonized system of classification and labelling of chemicals (ghs rev. 9, 2021), 2023 (accessed April 25, 2023). URL `https://unece.org/transport/standards/transport/dangerous-goods/ghs-rev9-2021`.

[63] K. P. C. Vollhardt and N. E. Schore. *Organic Chemistry: Structure and Function*. W. H. Freeman & Company, 8 edition, 2018. ISBN 9781319189044.

[64] W3C. OWL 2 web ontology language document overview (second edition), 2023 (accessed April 25, 2023). URL `https://www.w3.org/TR/2012/REC-owl2-overview-20121211/`.

[65] W3C. RDF schema 1.1, 2023 (accessed April 25, 2023). URL `https://www.w3.org/TR/2014/REC-rdf-schema-20140225/`.

[66] W3C. SKOS simple knowledge organization system reference, 2023 (accessed April 25, 2023). URL `https://www.w3.org/TR/skos-reference/`.

[67] W3C. SPARQL 1.1 query language, 2023 (accessed April 25, 2023). URL `https://www.w3.org/TR/2013/REC-sparql11-query-20130321/`.

[68] W3C. Semantic web, 2023 (accessed April 25, 2023). URL `https://www.w3.org/standards/semanticweb/`.

[69] D. Walsh. unit-parse · PyPI, 2023 (accessed April 25, 2023). URL `https://pypi.org/project/unit-parse/`.

[70] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. V. D. Lei, E. V. Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:1–9, 3 2016. doi:10.1038/sdata.2016.18.

[71] W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle, and E. V. Anslyn. The evolution of data-driven modeling in organic chemistry. *ACS Central Science*, 7(10):1622–1637, 2021. ISSN 23747951. doi:10.1021/acscentsci.1c00535.

[72] E. L. Willighagen, A. Waagmeester, O. Spjuth, P. Ansell, A. J. Williams, V. Tkachenko, J. Hastings, B. Chen, and D. J. Wild. The ChEMBL database as linked open data. *Journal of Cheminformatics*, 5:23, 5 2013. doi:10.1186/1758-2946-5-23.

[73] C. Willoughby and J. G. Frey. Data management matters. *Digital Discovery*, 1(3): 183–194, 2022. doi:10.1039/d1dd00046b.

[74] Y. Xie, K. Sattari, C. Zhang, and J. Lin. Toward autonomous laboratories: Convergence of artificial intelligence and experimental automation. *Progress in Materials Science*, 132(December 2021):101043, 2023. ISSN 00796425. doi:10.1016/j.pmatsci.2022.101043.

[75] C. Zhang, A. Romagnoli, L. Zhou, and M. Kraft. Knowledge management of eco-industrial park for efficient energy utilization through ontology-based approach. *Applied Energy*, 204:1412–1421, 10 2017. doi:10.1016/J.APENERGY.2017.03.130.

[76] L. Zhou, M. Pan, J. J. Sikorski, S. Garud, L. K. Aditya, M. J. Kleinelanghorst, I. A. Karimi, and M. Kraft. Towards an ontological infrastructure for chemical process simulation and optimization in the context of eco-industrial parks. *Applied Energy*, 204:1284–1298, 10 2017. doi:10.1016/J.APENERGY.2017.05.002.

[77] L. Zhou, C. Zhang, I. A. Karimi, and M. Kraft. An ontology framework towards decentralized information management for eco-industrial parks. *Computers & Chemical Engineering*, 118:49–63, 10 2018. doi:10.1016/J.COMPCHEMENG.2018.07.010.

[78] X. Zhou, D. Nurkowski, S. Mosbach, J. Akroyd, and M. Kraft. Question answering system for chemistry. *Journal of Chemical Information and Modeling*, 61:3868–3880, 8 2021. doi:10.1021/ACS.JCIM.1C00275.

[79] X. Zhou, D. Nurkowski, A. Menon, J. Akroyd, S. Mosbach, and M. Kraft. Question answering system for chemistry - a semantic agent extension. *Digital Chemical Engineering*, 3:100032, 6 2022. doi:10.1016/J.DCHE.2022.100032.