Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

Linking Reaction Mechanisms and Quantum Chemistry: An Ontological Approach

Feroz Farazi[1], Nenad B. Krdzavac[1,4], Jethro Akroyd[1,4],
Sebastian Mosbach[1,4], Angiras Menon[1,4], Daniel Nurkowski[2],
Markus Kraft[1,3,4]

released: 08 August 2019

[1] Department of Chemical Engineering
and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom
Email: mk306@cam.ac.uk

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

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

[4] Cambridge Centre for Advanced Research and
Education in Singapore (CARES)
CREATE Tower
1 Create Way
Singapore, 138602

Preprint No. 236

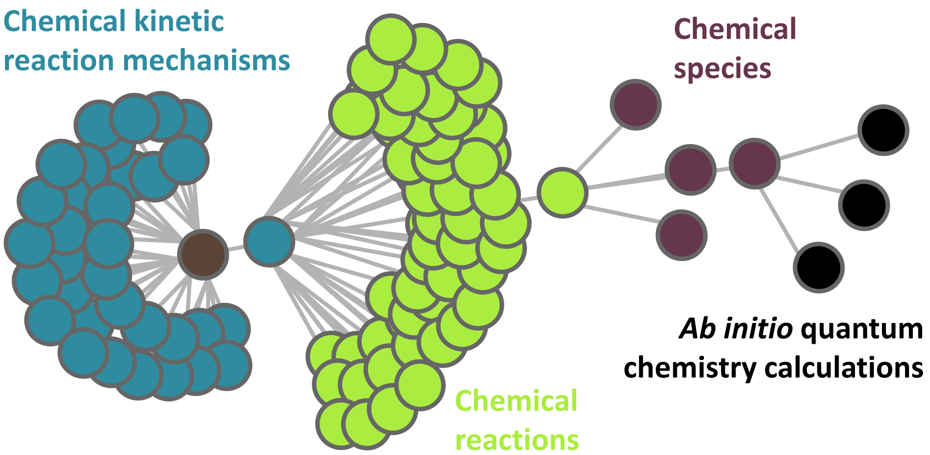
Keywords: Chemical kinetic reaction mechanism, quantum chemistry, linked data, chemical ontology, thermodynamic data, knowledge-graph, J-Park Simulator

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

E-Mail: c4e@cam.ac.uk
World Wide Web: `https://como.ceb.cam.ac.uk/`

Abstract

In this paper, a linked-data framework for connecting species in chemical kinetic reaction mechanisms with quantum calculations is presented. A mechanism can be constructed from thermodynamic, reaction rate, and transport data that has been obtained either experimentally, computationally, or by a combination of both. This process in practice requires multiple sources of data, which raises several issues. For example, the same species may have been given different names by different authors, whereas other species may have been given the same name even though they are distinct entities. Secondly, thermodynamic, reaction rate, and transport data may be inconsistent, with large variations outside stated error bounds between different sources. A linked data-centric knowledge-graph approach is taken in this work to address these challenges. In order to implement this approach, two existing ontologies, namely OntoKin, for representing chemical kinetic reaction mechanisms, and OntoCompChem, for representing quantum chemistry calculations, are extended. In addition, a new ontology, which we call OntoSpecies, is developed for uniquely representing chemical species. The framework also includes agents to populate and link knowledge-bases created through the instantiation of these ontologies. In addition, the developed knowledge-graph and agents naturally form a part of the J-Park Simulator (JPS) – an Industry 4.0 platform which combines linked data and an eco-system of autonomous agents for cross-domain applications. The functionality of the framework is demonstrated via a use-case based on a hydrogen combustion mechanism.


Chemical kinetic reaction mechanisms
Chemical species
Ab initio quantum chemistry calculations
Chemical reactions

Highlights

- A framework is built to link chemical species in mechanisms and quantum calculations.
- Existing ontologies are extended to represent species interconnections and provenance.
- OntoSpecies, an ontology for uniquely representing chemical species, is created.
- Knowledge-representational agents are implemented to populate knowledge-bases.
- A computational agent to calculate thermodynamic data for each species is developed.

Contents

1 Introduction

The current trend of boosting data exchange and automation in industry is frequently referred to as Industry 4.0. Major aspects of this include the Internet of Things (IoT) [4] – devices and sensors that can interact and exchange data, and cyber-physical systems (CPS) [36] – physical systems that are monitored and controlled by deeply integrated software. Every item in an Industry 4.0 environment must be accompanied by a corresponding digital representation, often termed Digital Twin, that provides information about the entity and that, crucially, can interact with others. This connectedness offers tremendous potential, which has also been noted in the context of the Semantic Web, where Linked Data [7] provides connections between previously unrelated information, thus enhancing its accessibility and enabling it to be identified and processed by automated software agents.

Within Industry 4.0, this connectedness can increase productivity and resource efficiency, leading to lower energy consumption and reduced emissions [34]. One aspect of this is the reduction of the carbon footprint and emissions arising from the use of fuels for transport, for instance in the shipping industry. In this case, a digital twin may include computational models to describe certain aspects of a ship's behaviour quantitatively, such as the production of pollutant emissions as a function of speed or choice of fuel. This requires the digital twin to have access to chemical models, known as chemical mechanisms.

A major barrier to using chemical models in an Industry 4.0 environment is that there is significant inconsistency between chemical mechanisms, both in terms of naming of chemical species and in terms of thermodynamic, kinetic and transport data [16, 33]. In common file-formats, the species contained in a mechanism are identified by arbitrary string labels. This can cause two types of uniqueness problems. Firstly, the same species may have different names assigned to them in chemical mechanisms developed by different authors. Secondly, distinct species may be given the same name in different mechanisms. For example, many mechanisms for the combustion of hydrocarbon fuels include benzene. The USC mechanism II [55] refers to benzene as C_6H_6, its molecular formula. However, other mechanisms such as the ABF mechanism [2], or the USC mechanism [35] refer to benzene as A1, in reference to its number of aromatic rings. While systematic species naming conventions exist, such as SMILES [58] and InChI [25], they are only partially successful in the sense that some uniqueness problems remain and that for instance stereoisomers and electronic states (spin and excitations) cannot be distinguished. In order to circumvent these challenges, a common approach, taken for example by the CAS registry [1] and PrIMe [16], is to introduce arbitrary but unique identifiers. In addition to naming problems, reaction mechanisms often suffer from inconsistencies in thermodynamic, kinetic and transport data, with different mechanisms sometimes showing significant differences in what should purportedly be the same quantity [33]. This is a problem because it means that alternative mechanisms for the same fuel may give inconsistent results, and because it makes it very difficult to combine chemistries from different models.

The complexity of chemical mechanisms and the inconsistency problems are such that it is infeasible to solve these problems by hand. Instead, a systematic and automated approach is required. A number of attempts have been made in the literature to automate the generation of mechanisms [19, 56] and the calculation of thermodynamic [28, 38] and kinetic data [6, 52]. For instance, Keçeli et al. [28] have automated the prediction of thermodynamic data for sets of species created by the Reaction Mechanism Generator (RMG) [19], purely from first principles, without involving existing experimental or computational data, and have applied this methodology to *n*-butane combustion. By generating lists of species from scratch and using only a single source of data, this approach avoids issues with provenance, curation of the 'best' currently available data, or naming and data inconsistencies.

The purpose of this paper is to create a single knowledge-graph that connects species in chemical mechanisms to computational chemistry data. We achieve this by taking a linked-data approach to the species naming inconsistency problems found in chemical mechanisms. This involves introducing an ontology and knowledge-base for unique species, as well as suitably extending existing ontologies for chemical mechanisms and quantum calculations. In this work, we consider only species which are intended to represent real physical molecules. Links to computational chemistry data enable the disambiguation of species beyond what is possible using methods such as InChI or SMILES. The created knowledge-graph provides a means not only to identify inconsistencies between different data sets but also to derive many quantities, such as species thermodynamic data, from first principles. The developed knowledge-graph and agents furthermore form an integral part of the J-Park Simulator (JPS)[1] – an intelligent simulation platform that uses ontology-based linked data and reasoning in cross-domain applications. In this paper, we consider an example based on hydrogen combustion.

The paper is structured as follows. Section 2 describes the J-Park Simulator (JPS), which is the context of this work. Section 3 provides an overview of current chemical databases and ontologies. Section 4 gives a detailed description of the proposed approach and framework. Section 5 demonstrates an instantiation of the framework using a hydrogen mechanism. Strengths of the approach and shortcomings of the present implementation are discussed in section 6. Conclusions are drawn in section 7.

2 J-Park Simulator (JPS)

The J-Park Simulator (JPS) [13, 45] is an Industry 4.0 application maintaining Digital Twins of smart buildings (*e.g.* Cambridge CARES Lab in Singapore – equipped with many sensors), energy networks (*e.g.* Jurong Island power network [11, 12]), and chemical industry (*e.g.* a biodiesel plant [59]) amongst others within a knowledge-graph that contains in particular also chemical models (*i.e.* chemical mechanisms) and software agents [61] (*e.g.* computational agents). Connectivity among these physical and virtual entities is established and maintained by autonomously operating software agents.

Figure 1 shows the conceptual diagram of JPS depicting real-world entities, their Digital Twins as well as connectivity and interactions between them. An agent reads the status of a real-world entity via a sensor and updates the Digital Twin of the entity by operating on the knowledge-graph. Based on the new status, the agent uses the reasoning tool and

[1] http://www.theworldavatar.com

deduces that an action has to be performed on the entity by involving other relevant agents. Finally, the action is carried out via an actuator.

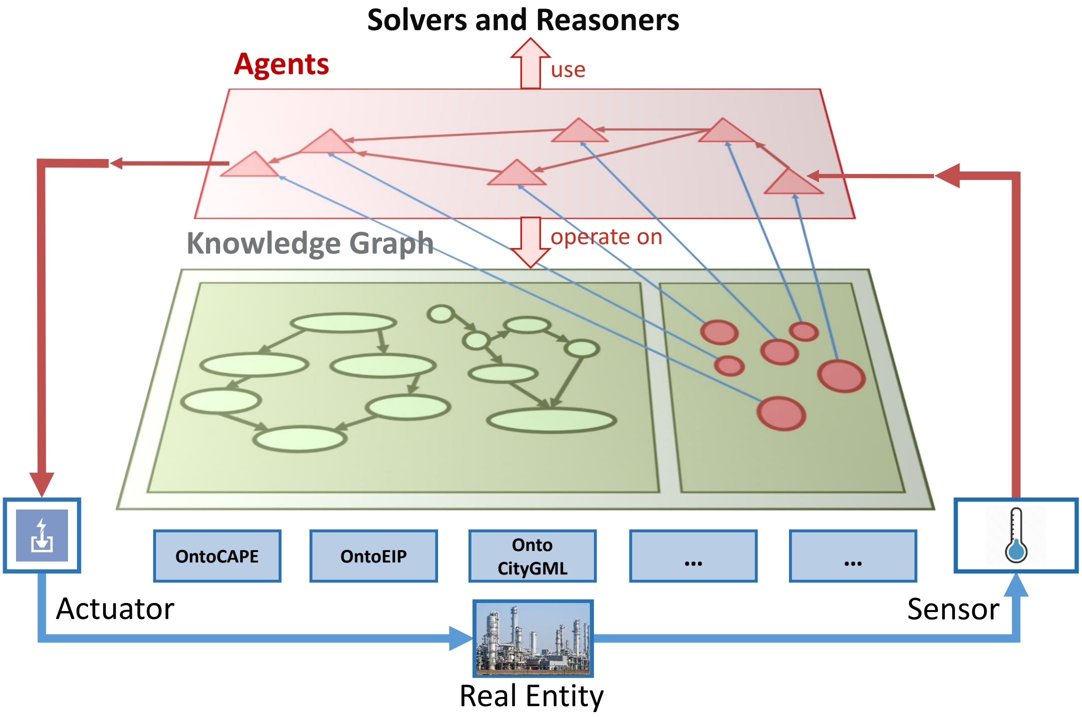


Figure 1: *The JPS knowledge-graph and agents [61]. The former connects multiple domains and the latter interact with the knowledge-graph and different software.*

The knowledge-graph encompasses all Digital Twins that are part of the JPS, and is maintained by agents. The knowledge-graph accommodates the involved cross-domain knowledge by means of several ontologies including OntoCAPE [39], OntoEIP [60], OntoCityGML [13], OntoPowSys [60], OntoKin [14] and OntoAgent [61]. In addition, several knowledge-bases from the Linked Open Data (LOD) Cloud[2] are used, such as DBpedia[3] [37]. OntoCAPE was created for representing chemical process knowledge, while OntoEIP was designed for codifying resources, transportation networks and chemical plants for managing an Eco-Industrial Park (EIP). The OntoPowSys ontology was developed for representing electrical power systems, whereas the OntoKin ontology was designed for describing chemical mechanisms. OntoCityGML can describe cities and landscapes, and OntoAgent can represent software agents developed for various purposes.

Figure 2 demonstrates an existing cross-domain use-case of the JPS to explain how multiple components are connected together. In this example, the components with the red background represent agents and the components with the light-blue background are data sources. The figure shows that JPS uses the SRM Ship, created by customising SRM Engine Suite [30, 32, 43, 53], a toolset designed and developed to model the performance of and emissions from internal combustion engines, to estimate the exhaust emissions from

[2] https://lod-cloud.net/

[3] http://dbpedia.org/

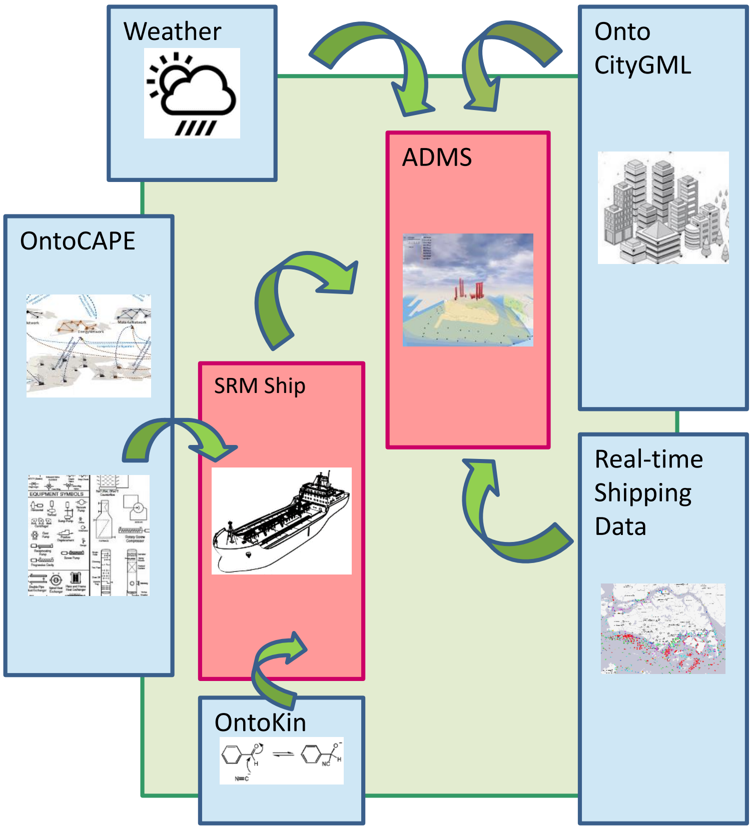


Figure 2: *A cross-domain use-case demonstrating the atmospheric dispersion of pollutant emission from ships, integrating real-time data of ship locations, combustion simulation, chemical models, 3D-representations of buildings, live weather data, and atmospheric dispersion models – all made interoperable through the use of ontologies.*

ships. JPS then uses ADMS [8], the Atmospheric Dispersion Modelling System, to simulate the dispersion of these emissions in the vicinity of each ship. The SRM simulations use chemical mechanisms obtained via OntoKin and the corresponding knowledge-base. The ADMS simulations use live weather data acquired from a public website through an API to calculate the distribution of the emissions, which is then visualised in JPS as an overlay on a 3D map. This example illustrates the potential of chemical models being part of a knowledge graph, *i.e.* in the form of Linked Data, with the associated benefits of semantics and interoperability, as well as the value of higher-quality chemical models, to solve wider cross-domain problems such as environmental pollution.

Further examples exist where detailed chemical reaction mechanisms play a central role as part of cross-domain applications that integrate multiple ontologies. These include industrial air pollution scenarios, similar to the ADMS ship emissions example above, involving power plants, namely "*Heizkraftwerk Mitte*" in Berlin and "*Energiecentrale*" in The Hague [13].

3 Chemical databases and ontologies

3.1 Chemical databases

The world's largest freely accessible database of chemical information is PubChem [29], which stores information in three primary categories: compounds, substances, and bioactivities. Currently, PubChem has information on 97 million compounds, 242 million substances, and 280 million bioactivities. Information in PubChem can be queried by standard means, such as text search, molecular formula, chemical structure. For a common molecule, such as benzene for example, PubChem contains a variety of properties. This includes 2D and 3D structures as well as any crystal structures which can be downloaded in standard formats such as JSON, XML, or CIF for crystal structures. PubChem also computes the standard identifier for the species in question, such as the IUPAC name, the canonical SMILES identifier, or the InChI format, as well as other vendor/chemical agency identifiers. Key computed and experimental chemical and physical properties for the structure, such as molecular weight, formal charge, melting point, boiling point, vapor pressure and others are also provided, as is any available spectral (UV-Vis, NMR, IR, Mass spectrometry) data that has been linked to the structure. PubChem also provides a large amount of information on the biological aspects of such structures, including drug information, solubility, toxicity, and biological activity.

Another major database for chemical data is Reaxys, run by Elsevier. Reaxys [22] contains much of the same information as PubChem and other chemical databases, such as structure, key identifiers, physical and chemical properties, spectral data, and biological activity for various compounds. Reaxys' query builder allows the user to search for information using a variety of methods, be it text search, or search by molecular formula, chemical properties, structure, synthesis route, or chemical reactions the species of interest is involved in. Reaxys has three key sets of information for a substance, namely preparations, reactions, and documents. Preparations displays key synthesis routes that can be used to prepare the substance in question. Similarly, the reaction set contains the list of reactions in the Reaxys database which includes the substance the user has queried. Finally, the documents set lists the journal publications, patents, conference papers, and books that Reaxys has access to that are linked to the queried substance.

Databases that focus more on storing computational chemical information include the Computational Chemistry Comparison and Benchmark DataBase (CCCBDB) [27] for thermochemical properties of species from the National Institute of Standards and Technology (NIST). Information is queried by chemical name or molecular formula. The CCCBDB stores computed information in the following main categories: energy, geometry, vibrations, electrostatics, entropy and heat capacity, and reaction. Energy contains a variety of energies related to the queried structure. This includes optimised ground state energies, internal rotation barriers, HOMO (highest occupied molecular orbital) and LUMO (lowest unoccupied molecular orbital) energies, nuclear repulsion energies, correlation data, vertical ionization energies, electron and proton affinities, as well as excited state energies for both singlet and triplet states. Complete basis set extrapolations for energies are also provided where available. The geometry category contains the optimised ground state geometry of the structure, which can be extracted in bond order, cartesian, and Zmatrix format. Also included in this category is rotational information such as rotational constant, moments of inertia, inertial defects, and any computed second moments, as well as symmetry and point group information available for the queried structure. Frequencies contains the computed vibrations and zero-point energies for the structure, as well as recommended scaling factors for the computed data when used in further kinetic or thermochemical calculations. Electrostatics contains the atomic partial charges in different formats (Mulliken, ESP, AIM, CHELP), multipole values (dipoles and quadrupole values), polarizability tensor values, and spin densities for the queried structure. Entropy and heat capacities are displayed as calculated at 298 K. Finally, the reactions category allows the user to search for reaction energies at 0 K and 298 K, reaction entropies, transition states, and isodesmic reactions for a given set of reactant structures. All of the computed properties are displayed for different levels of theory that they have been calculated at. This is split into four main categories: methods with pre-defined basis sets, methods with standard basis sets, methods with effective core potentials, and single-point energy calculations. The first category includes semi-empirical methods such as PM6 as well as the Gaussian series composite methods (G1, G2, G3, G4, *etc.*). The second category includes standard methods with their corresponding basis sets, namely Hartree-Fock (HF), Density Functional Theory (DFT), Møller-Plesset Perturbation (MP2 and MP4), Quadradic Configuration Interaction (QCISD), and Coupled Cluster (CC) methods along with the standard basis set used (either Pople-type or Dunning-type). The third category is reserved for basis sets where the core electrons are not described explicitly but instead replaced by a pseudo-potential (or core potential). This is typically done for larger atoms where treating all electrons would be prohibitively expensive computationally and includes basis sets such as LANL2DZ for example. Finally, single point energy calculations report any values computed using mixed methods, *i.e.* where the geometry was optimised at one level of theory and the property was computed at another, typically higher level of theory. The CCCBDB also crucially has a comparison feature, where the user can compare the results of theoretical calculations to any available experimental data in NIST's databases, as well as look at the effect of calculation details like the integration grid in DFT calculations, or even properties for very similar molecules.

Other more specialised databases also exist. For example, the Alexandria library developed by Ghahremanpour et al. [20] consists of molecular properties for force field development. Alexandria contains molecular structures and properties for 2704 compounds, many of which contain functional groups common to biomolecules and drugs. Alexandria contains similar information to CCCBDB such as enthalpies of formation, heat capacities, entropies, zero-point energies, and frequencies. Alexandria contains more extensive multipole and polarizability calculations, with calculations up to hexadecapole moments provided, as well as electrostatic potentials and partial atomic charges in the various formats (Mulliken, Hirshfeld, ESP, CM5). Key to Alexandria is also the fact that geometries, vibrational, and electrostatic properties are all provided at the same level of theory, namely B3LYP/aug-cc-PVTZ. High level calculations are also provided for thermochemistry properties (G2, G3, G4, CBS-QB3, W1U, W1BD). Most importantly, Alexandria provides the Gaussian input and output files from the calculations, making reproducing the stored information significantly easier.

In terms of more specialised databases, Hait and Head-Gordon [23] provide a benchmark database specifically for DFT calculations on dipole moments, spanning a variety of functionals in the process. The database by Simmie [50] is specifically for high-level enthalpies of formation for nitrogen based compounds. The GDB-17 database [48] specifically enumerates small organic molecules, using graph-theoretic methods to span 166 billion of such molecules with the aim of guiding new drug design. Ramakrishnan et al. [47] provide the QM9 dataset, which is the main benchmark for training new machine learning potentials. It contains DFT calculations on around 134,000 molecules, mainly at the B3LYP/6-31G(2df,p). The information in QM9 is standard, containing geometries, thermochemistry, electrostatic properties, and vibrations. Geometries are cross-checked by generating InChI and SMILES identifiers using Open Babel [44] and comparing to GDB-17. B3LYP energetics are validated against higher level calculations for a small subset of 100 molecules, enabling uncertainty estimates for the data provided in the databases. The ANI-1 data set [51] uniquely contains non-equilibrium DFT calculations, that is for molecules in conformers that are not their minimum energy ground state configuration (hence non-equilibrium). Calculations are provided at the wB97x/6-31G(d) level of theory and contains around 20 million molecular conformations for 57,462 molecules taken from the GDB database. Goldsmith et al. [21] have presented first-principles thermochemistry for a collection of 219 small combustion-relevant molecules at the RQCISD(T)/cc-PV∞QZ//B3LYP/6-311++G(d,p) level of theory. This highly accurate dataset has allowed identification of discrepancies and errors in established databases.

In short, a variety of databases exist that contain both experimental and computational chemical information for a variety of purposes, be it reactions, kinetics and thermodynamics, method development in quantum chemistry, machine learning and big data analytics of chemical networks, or synthetic planning. The above databases do not appear to support querying and reasoning capabilities based on a semantic language or formal logic, although Reaxys' smart query does give significant flexibility and combination possibilities in what one wants to search. These capabilities can be supported and enhanced by chemical ontologies.

3.2 Chemical ontologies

This subsection provides a brief description of chemical ontologies dealing with chemical species and/or molecular entities. The ontologies are described from the following perspectives: application, purpose, and focus.

Chemical Entities of Biological Interest (ChEBI) [10] is an ontology derived from a database developed for standardising terminology and describing chemical structures of molecular entities used in the biochemistry discipline. The applications of ChEBI include its exploitation as background knowledge for data mining applications that take into account the semantic similarity between chemical entities [15]. In order to assess similarity, these applications use the concept hierarchy and disjointness axioms of the ontology. The ontology models cross-domain knowledge and is divided into four sub-ontologies called molecular structure, biological role, application, and subatomic particle. ChEBI has been employed as a chemical ontology in the widely-used Gene Ontology (GO) [3] and PubChem along with several other ontologies. The ontology focuses on 'small' chemical compounds with low molecular weight, selected based on their suitability to be applied in researching biological functioning. Similarly to the parent database, the purpose of this ontology is to serve as a repository for such chemical compounds.

The BioChEBI ontology [26] is a biochemical semantic resource generated through the integration of Gene Ontology and the chemical part of ChEBI. Gene Ontology has revolutionised the way of performing the biological data analysis by introducing logical connectivity between genes and proteins. The ontology has the potential to be applied as background knowledge to data integration services that deal with biological and chemical data to assist in drug discovery. The purpose of BioChEBI is to create a coherent representation of chemical entities across GO and ChEBI. BioChEBI focuses on chemical molecules which are observed in biological processes in genes.

Gainesville Core [54] is an ontology consisting of logic-based formal definitions and textual descriptions of computational chemistry vocabulary and logical axioms. This ontology has been applied to integrate data generated by different computational chemistry software and packages. The purpose of the ontology is to offer a means to represent and publish both computational chemistry data and its semantics. Gainesville Core focuses on properties related to molecular systems (*e.g.* multiplicity and charge) and quantum calculations (*e.g.* basis sets and spin types).

PubChemRDF [18] is an RDF representation of PubChem, which is a database containing different types of entries including chemical entities and biological activities. One application of PubChemRDF is to allow researchers to use Semantic Web technologies, *e.g.* RDF triple stores for storage and SPARQL for querying. Another application is to integrate PubChem with other biochemical semantic resources by using external ontologies such as CHEMical INFormation ontology (CHEMINF) [24] for enhancing the data analysis experience of researchers. The purpose of PubChemRDF is to express the meaning of PubChem data in a format that is interoperable with the Semantic Web. It focuses on the representation of chemical substances, compounds and structures thereof.

OntoCAPE [42] is an ontology designed for capturing different aspects of chemical process engineering. The application of this ontology is manifold, including the annotation of documentation and specifications produced for process engineering tasks, automatic composition of software components for process modelling, and ontology-supported computer-aided process modelling. The purpose of OntoCAPE is to produce machine-readable and interpretable models of chemical processes for enhancing automation in relevant industries. The ontology focuses on the modelling of chemical processes and plants from the perspectives of their design and development. OntoCAPE includes the ontological modelling of chemical species and molecular entities.

OntoKin [14] is an ontology providing the Description Logic (DL)-dependent expressivity for representing gas-phase and surface-phase chemical mechanisms. The application of this ontology includes the development of a queryable reaction mechanism knowledge-base, a simulation tool to estimate the atmospheric dispersion of emissions from diesel generators, and a tool to visualise mechanisms represented on the Semantic Web. The purpose of the ontology is to move a stage closer to solving the data inconsistency issues across reaction mechanisms. The focus of the ontology is to model a reaction mechanism based on the phases it contains, reactions that occur within these phases, and species that participate in these reactions.

The OntoCompChem ontology [31] extends the Gainesville Core ontology by incorporating concepts from CompChem [46]. CompChem is a data format which is built upon Chemical Markup Language (CML) and is intended for the codification of generic chemical data semantics. OntoCompChem is thus able to represent quantum chemistry calculations as performed by the Gaussian [17] software in a fully CML-compliant manner. The application of CompChem is effective storage, management and retrieval of fast-growing quantum chemistry data originating from the extensive use of related software. Its purpose is to represent semantics of computational chemistry calculations. This enables the reuse of already performed calculations, hence reducing the computational resource consumption time. The focus of CompChem is the representation of computational chemistry calculations needed for analysing thermochemistry, whereas OntoCompChem has been applied to support interoperability between quantum chemistry and thermochemistry calculation tools. Similarly to CompChem, the purpose of the OntoCompChem ontology is to manage the storage of computational chemistry calculations along with their semantics to enable easy retrieval and greater reuse. The ontology focuses on the representation of quantum chemistry and thermochemistry calculations of chemical species.

4 A knowledge-graph for chemical species

4.1 Ontological resources

This subsection describes a formal approach based on the use of ontologies to link data about the species in chemical mechanisms with data about the species derived from computational chemistry calculations. We propose a new ontology called OntoSpecies as a core component of this approach.

Our approach includes the application of the OntoSpecies, OntoKin and OntoCompChem ontologies to meet both the data and semantics representational requirements for the successful execution of the knowledge-graph generation task. OntoSpecies has been developed from scratch while OntoKin and OntoCompChem have been extended to reach the end goal. The reason for the use of these ontologies in this work is their logical and infrastructural capabilities to manage different types of properties of chemical species. OntoSpecies has been designed to capture generic information about species, such as empirical formula and molecular weight. OntoKin, on the other hand, supports the encoding of properties that are relevant to chemical mechanisms, such as thermodynamic data and transport data. OntoCompChem covers features related to computational chemistry calculations of various properties of species, such as functional and basis set.

Figure 3 depicts these ontologies with their concepts, data properties and relations which are crucial to describe the aforementioned data linking approach. OntoSpecies helps resolve the naming issues described in section 1. The way it does so is that there is a set of unique entries, each of which corresponds to a unique real-world species. It also provides an infrastructure to link data about species derived from different sources with different levels of granularity; in this case, chemical mechanisms and computational chemistry calculations.

The figure includes the following concepts of OntoSpecies: Species, Empirical Formula,

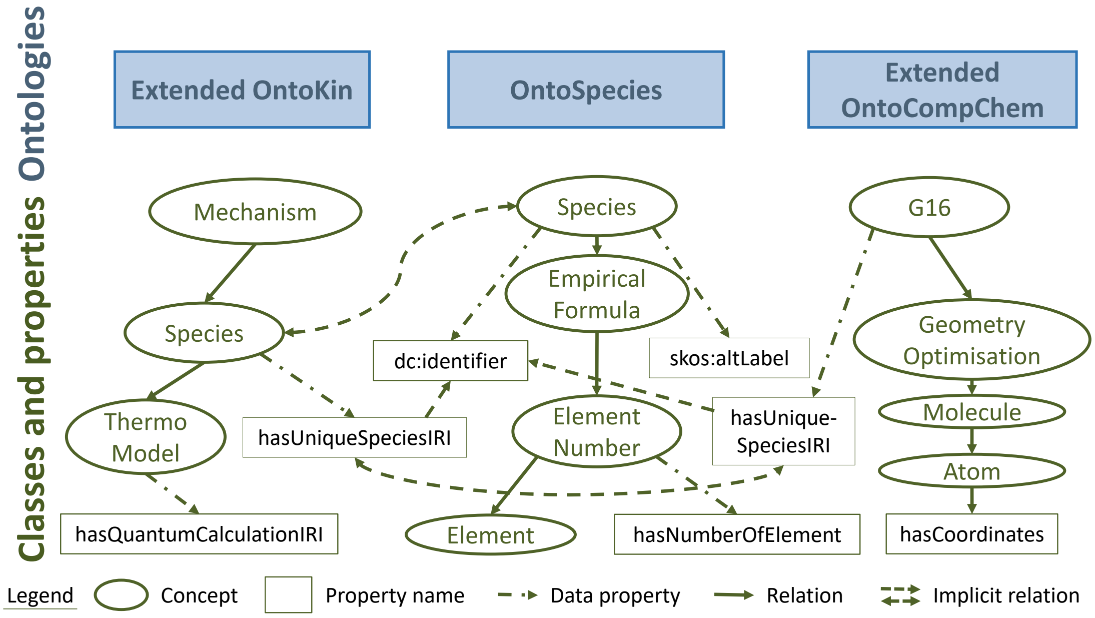


Figure 3: *Core concepts and properties of the OntoSpecies and extended OntoKin and OntoCompChem ontologies with implicit and explicit links between elements of these ontologies. OntoKin and OntoCompChem contain many more concepts and relations, which are omitted here for simplicity.*

Element Number and Element. The Species concept allows for the creation of a real-world species via instantiation. The Empirical Formula concept is provided to model the type and number of elements available in a species. Element is defined to codify the instance of an element or an atom, while Element Number connects an element with its multiplicity within a species. The data properties that belong to OntoSpecies are dc:identifier, skos:altLabel and hasNumberOfElement. By adopting best practices in ontology development, the identifier was reused from Dublin Core (dc) [57] as was altLabel from Simple Knowledge Organisation System (skos) [41]. The unique identifier of species is codified using dc:identifier, and alternative names are codified using skos:altLabel. The hasNumberOfElement data property has been defined to codify the amount of an element or an atom in a species. The speciality of this modelling choice is that it separates the names of a species from its identity. As a result, a species which has multiple names can still be recognised uniquely via its identifier. Not shown in the figure are concepts and properties for the standard enthalpy of formation as well as the corresponding reference temperature. The OntoSpecies ontology is available on the web[4].

Figure 3 shows the following concepts of OntoCompChem: G16, Geometry Optimisation, Molecule and Atom. From the aforementioned four concepts, G16 deals with the modelling of electronic structure calculations, while Geometry Optimisation, on the other hand, allows for the modelling of the molecular geometry of both stable minima and transition state species. Molecule enables the modelling of the constituent parts and data properties of a molecular entity. By contrast, Atom can model the data properties of a chemical element. The hasCoordinates object property is used for the codification of the

[4] http://www.theworldavatar.com/ontology/ontospecies/OntoSpecies.owl

3D geometry of a molecule. OntoCompChem is extended in this work by a data property called hasUniqueSpeciesIRI. The hasUniqueSpeciesIRI data property links computational chemistry calculations of a species to its corresponding representation in OntoSpecies by means of an IRI (Internationalised Resource Identifier).

The concepts of OntoKin shown in Fig. 3 are Mechanism, Species and Thermo Model. OntoKin contains many more concepts and relations than shown in the figure, in particular concerning kinetic and transport data, but since the aim of this work is to establish links to quantum chemistry calculations, we focus here mainly on the concepts affected by this. The Mechanism concept was defined to model the data and metadata including the reference of a chemical mechanism. The Species concept can be employed to model data properties and relations of a chemical species. Thermo Model allows for the codification of thermodynamic models that can be defined for a chemical species. OntoKin is extended in this work by two data properties: hasQuantumCalculationIRI and hasUniqueSpeciesIRI. The hasQuantumCalculationIRI data property is an IRI specified to establish a link between the thermodynamic model and computational chemistry calculations of a chemical species, whereas the hasUniqueSpeciesIRI data property is an IRI defined to link a species in a mechanism with its corresponding representation in OntoSpecies.

The relation drawn between the Mechanism and Species concepts is called hasSpecies, which connects an instance of the former concept to an instance of the latter. Similarly, the relation that exists between Species and Thermo Model is termed hasThermoModel and can connect the instances of these concepts. The relations in OntoSpecies and the extended OntoCompChem have similar interpretations.

4.2 Linking between ontologies

This subsection describes how data resulting from computational chemistry calculations is linked to chemical mechanisms, and how this extends to linking multiple pieces of data. The description is based around a conceptual example that uses the ontological concepts and properties presented in the previous subsection.

Figure 4 shows how a species belonging to a mechanism can be linked to multiple computational chemistry calculations, each generated at a different level of theory and the same species belonging to multiple mechanisms can be linked to the same computational chemistry calculation. The knowledge-bases ‘OntoSpecies KB’, ‘Extended OntoKin KB’ and ‘Extended OntoCompChem KB’ were created for populating with instances of the concepts belonging to the OntoSpecies, extended OntoKin and extended OntoCompChem ontologies, as described in the previous subsection, respectively.

OntoSpecies KB was populated with a group of species including H_2O and H_2, each with properties such as the unique ID, alternative labels (names) and the empirical formula. Ideally, each real-world species appears exactly once in this knowledge-base. On the other hand, the extended OntoCompChem KB was populated with computational chemistry calculations for the same group of species. The extended OntoKin KB was populated with a collection of mechanisms including Mechanism 1 and Mechanism 2, each containing a set of species. Both Mechanism 1 and 2 contain the species H_2O and H_2, each codified with a set of properties including hasUniqueSpeciesIRI and hasQuantumCalculationIRI.

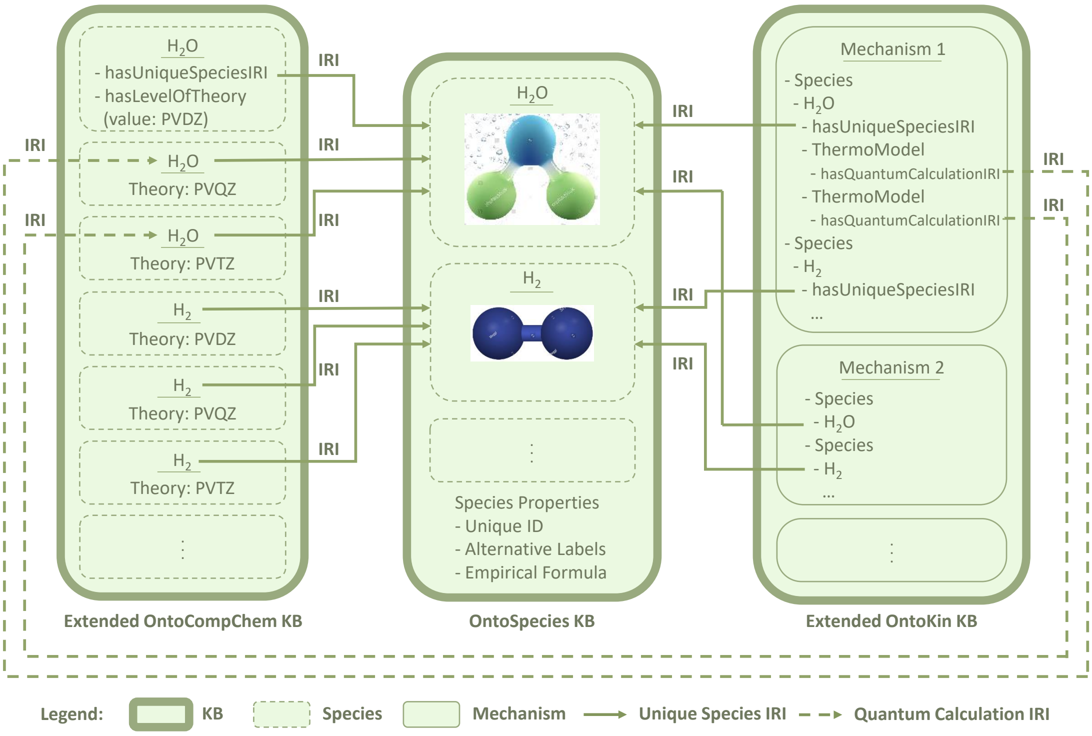


Figure 4: *Links between individuals in the OntoSpecies, extended OntoKin, and extended OntoCompChem knowledge-bases. Instances of quantum chemistry calculations at various levels of theory as well as instances of species within mechanisms are connected to unique species by means of IRIs. Similarly, thermodynamic data associated with species in mechanisms can be connected to quantum calculations through IRIs.*

The hasUniqueSpeciesIRI property links H_2O in Mechanism 1 and 2 to the unique representation of H_2O in OntoSpecies KB. Similarly, the representations of H_2 in these mechanisms and OntoSpecies KB are linked. Each link uses the IRI of the corresponding unique species. The hasQuantumCalculationIRI property, on the other hand, links H_2O in Mechanism 1 to H_2O in the extended OntoCompChem KB to refer to multiple calculations performed at the levels of theory PVQZ and PVTZ. In these links, the IRIs of the calculations were used. The hasUniqueSpeciesIRI property of H_2O in the extended OntoCompChem KB contains the IRI of H_2O in OntoSpecies KB.

This approach is consistent with the data linking approach proposed by Berners-Lee [5]. IRIs are used to identify all concepts, instances, data properties and relations. In order to make the IRIs resolvable, they are appended to the base address of a HTTP server and converted to HTTP IRIs. HTTP IRIs of species, computational chemistry calculations and mechanisms show meaningful information when they are looked up. Finally, to enable agents to discover more knowledge about the objects of interest, species in mechanisms are linked to the generic information about species and their computational chemistry calculations.

4.3 Computational and representational agents

We want to populate the knowledge-graph and use the computational chemistry data to calculate thermodynamic properties for each species, and link the thermodynamic property data such that it can be used in the context of the chemical mechanisms. This is performed using a mixture of manual and automated tasks conducted by computational agents. The long-term intention is that all processes will be automated.

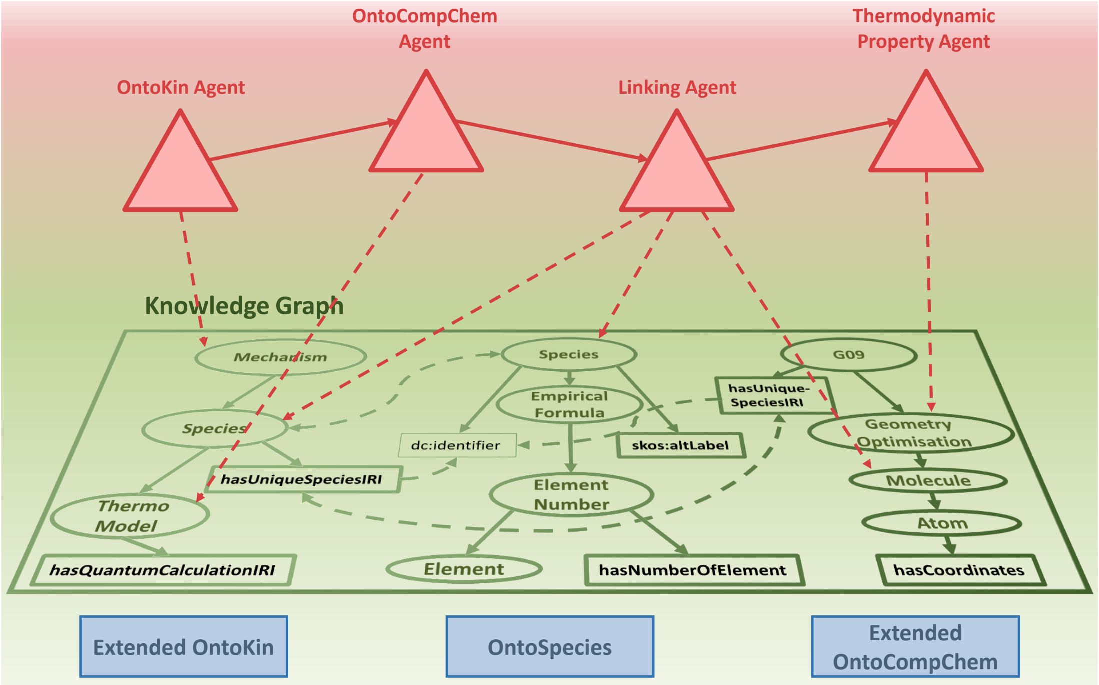


Figure 5: *The computational and representational agents interacting via the knowledge-graph.*

In Fig. 5, it is illustrated how the computational chemistry and knowledge-representational agents interact via the knowledge-graph to carry out cross-domain tasks and at the same time contribute to the enrichment and evolvement of the knowledge-graph with information that will enhance the performance of computational agents. Three knowledge-representational agents have been designed and deployed to populate the knowledge-bases built using the aforementioned ontologies. A computational agent has been developed to calculate thermodynamic properties and data for each chemical species. We have implemented these four agents such that they seamlessly integrate with the OntoAgent [61] ontology and thus extend the existing agent eco-system of the JPS. More specifically, the agents can be briefly described as follows, where we use the same agent classification, into Types 0, 1, . . . , 4, as in previous work [13]:

- OntoKinInp Agent: Adds chemical mechanisms (represented in the extended OntoKin KB) to the knowledge-graph. This agent is an input agent and hence of Type 0.
- OntoCompChemInp Agent: Adds the results of computational chemistry calculations (represented in the extended OntoCompChem KB) to the knowledge-graph.

At this stage, we assume computational chemistry calculations are already completed. This agent is another input agent and hence also of Type 0.

- Linking Agent: Links computational chemistry data in the knowledge-graph by linking entries in the extended OntoCompChem KB to OntoSpecies KB. This agent modifies the structure of the knowledge-graph (by creating new links) and hence is a Type 2 agent.
- Thermodynamic Property Agent: Calculates thermodynamic properties for each species using computational chemistry data, and links the thermodynamic property data such that it can be used in the context of the chemical mechanisms. This requires a number of computations:
  - Partition function calculations are required to calculate the thermochemical properties (*e.g.* entropy, heat capacity, enthalpy) [40]. The standard enthalpy of formation and its associated reference temperature, both of which are needed as inputs to the agent, are taken from the corresponding species instance in the OntoSpecies KB.
  - The calculated thermochemical properties must be parameterised in a form that is suitable for chemical mechanisms (in this case 7-coefficient NASA polynomials).
  - The parameterised thermochemical properties must be added to the knowledge-graph where they are represented in the extended OntoKin KB.

  As a calculation agent, this agent is of Type 1.

We link mechanism data in the knowledge-graph by manually linking entries in the extended OntoKin KB to OntoSpecies KB.

This approach is how we perform data conversion, as per the principles proposed by Berners-Lee [5]. The goal is to produce standardised well-linked data. By publishing mechanisms, unique chemical species, and computational chemistry calculations using OWL and RDF, we achieve four stars, where the highest rating is five stars, which is awarded when four-star data is linked to external sources. We reach the five-star rating by having linked this four-star data to NIST, thus producing data that allows the agents to gather more knowledge about chemical species and further extend the knowledge-graph.

5 A hydrogen-mechanism use-case

We have selected the hydrogen combustion mechanism for the use-case mainly for three reasons. Firstly, it is the simplest gas-phase combustion system that can be considered. Secondly, it forms the basis of all hydrocarbon combustion mechanisms. Finally, for the above reason, it is also the most well-established mechanism in terms of species, reactions, thermodynamic data and kinetic data.

The hydrogen combustion mechanism used in this work consists of 10 species and 40 reactions [9]. To have a consistent thermodynamic and kinetic data for the mechanism, DFT calculations were performed on all species in the mechanism using Gaussian 16 [17]. The widely applicable DFT functional B3LYP was used for all of the calculations. Geometry optimizations were performed using the tight convergence criteria. Frequencies were calculated for molecules to ensure that a minimum on the potential energy surface was located. To test the effect of basis set, three different Dunning-type basis sets were used: cc-PVDZ, cc-PVTZ, and cc-PVQZ. All log files for the calculation were set to the 'verbose' print standard so that as much information was included in the log file as possible. The resulting Gaussian 16 log files were parsed by Thermodynamic Property Agent, enabling the calculation of thermodynamic properties and NASA polynomial fittings for thermochemical data of the mechanism.

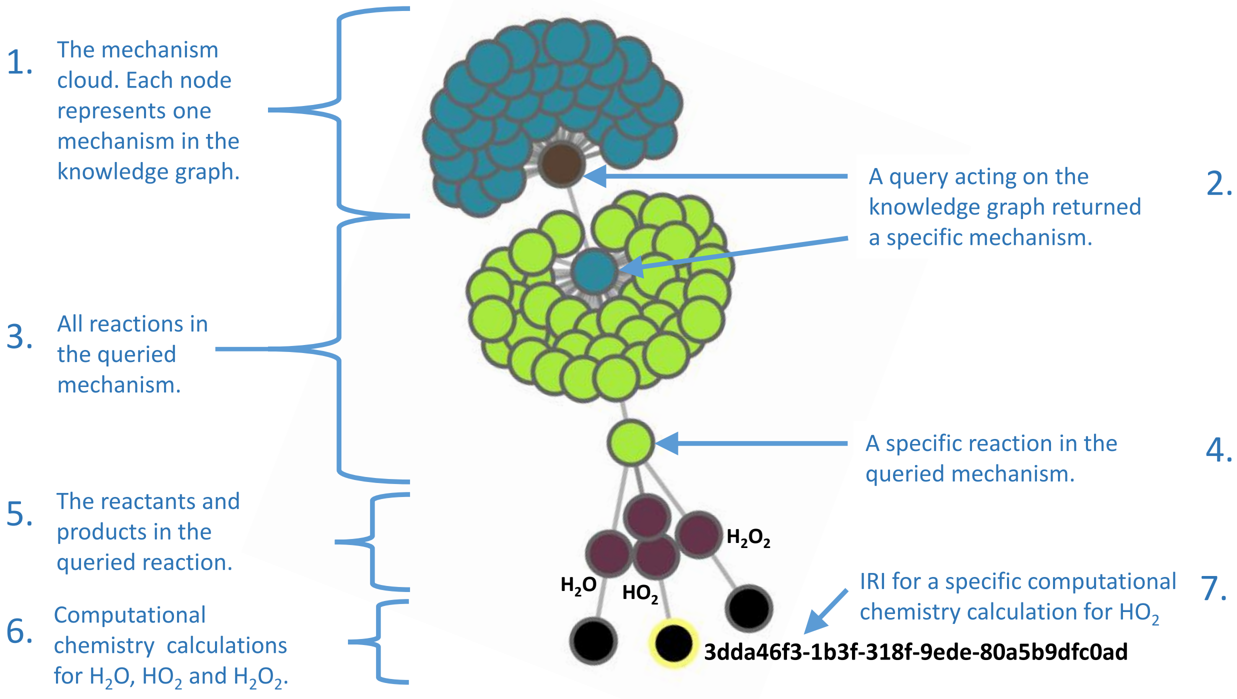

1. The mechanism cloud. Each node represents one mechanism in the knowledge graph.
2. A query acting on the knowledge graph returned a specific mechanism.
3. All reactions in the queried mechanism.
4. A specific reaction in the queried mechanism.
5. The reactants and products in the queried reaction.
6. Computational chemistry calculations for H_2O, HO_2 and H_2O_2.
7. IRI for a specific computational chemistry calculation for HO_2
H_2O_2
H_2O
HO_2
3dda46f3-1b3f-318f-9ede-80a5b9dfc0ad

Figure 6: *Annotated screenshot demonstrating how chemical species in reaction mechanisms are connected to computational chemistry calculations via the linked-data approach. Clicking on individual nodes expands the visible part of the knowledge-graph. Most of the node-labels have been suppressed to aid legibility.*

Chemical species thermodynamic data was obtained using the Thermodynamic Property agent. The agent calculates heat capacities, enthalpies and entropies from the species molecular partition functions using the rigid-rotor harmonic-oscillator treatment (RRHO), which includes translational, vibrational, rotational and electronic energy contributions. Standard enthalpies of formation, together with their reference temperature, are also provided as inputs to the agent for each species.

Figure 6 illustrates the functionality of the proposed linked-data compliant knowledge-graph approach. The hydrogen mechanism was represented using OWL and published on the mechanism cloud, which forms part of the knowledge-graph. A SPARQL query was performed on the knowledge-graph to visualise the hydrogen mechanism. An interactive interface was designed to navigate through the reactions and species. As an example, when the reaction $OH + H_2O_2 \rightleftharpoons H_2O + HO_2$ is expanded, its reactants and products are visualised. These reactants and products are linked to the corresponding species in OntoSpecies KB and these species are linked to computational chemistry calculations performed at different levels of theory for each species. The figure shows three different quantum calculations associated to three of the species involved in the selected reaction. We remark that the reason reactions appear here, even though they do not feature directly in the links between species in mechanisms and quantum calculations, and hence in the discussion in section 4, is simply that this is one out of several possible ways how a user might 'explore' a chemical mechanism.

6 Discussion

The connectivity offered by a linked-data approach is a critical prerequisite for automated processing by autonomous pieces of software. It has long been recognised that in order to be feasible in practice, chemical model generation or problem resolution with existing ones inevitably has to be systematic, repeatable, and hence automated, because the tasks involved would be too time-consuming or error-prone to conduct manually. For this reason, several attempts at automating chemical model creation have been made in the literature or are currently in progress. All attempts known to the present authors have in common, though, that no reference to the outside world is made, and no integration with existing models or data, in particular experimental data, or more generally external sources is considered. Hence, the necessity of a linked-data approach also extends to the automated creation of chemical models.

Similarly, in our wider review of existing chemical databases and ontological approaches, we have found a wealth of data, however also a near-ubiquitous lack of interoperability and linked-data principles. In this work, we have developed a framework which integrates ontologies, knowledge-bases, and agents into a single knowledge-graph and thus we have taken a first step towards realising linked data for chemical models. The knowledge-graph is the core component of the developed framework. In assembling it, we followed the linked-data principles as proposed by Berners-Lee [5], which are broadly applicable to data conversion, linking, and sharing [49], and recommend to make the semantics of data explicit [7].

The developed framework allows links between chemical species and computational chemistry calculations to be established. In order to avoid problems with naming species uniquely, as encountered by standard naming conventions such as SMILES and InChI, we introduced arbitrary but unique identifiers – an approach also taken for instance by the CAS registry and PrIMe. This, of course, does not address the problem of which unique identifier should be associated with an existing species in a mechanism for example. This remains an open problem and is an area of ongoing research [33]. The linked nature of the framework enables it to be used to curate existing knowledge, and furthermore to address data inconsistencies between multiple sources.

While developing the ontological component for representing thermodynamic data of species, a choice needed to be made as to where to accommodate this. In principle, any of the three ontologies, OntoKin, OntoCompChem, or OntoSpecies, are potential candidates, and a case can be made for each of them as they all include in some way a concept of species. Being post-processed from quantum calculations might suggest to include thermodata into OntoCompChem, whereas it is common (*e.g.* [27]) to curate thermodata alongside unique species, thus suggesting OntoSpecies. Ultimately, we made the choice to represent thermodynamic data within OntoKin simply for practical reasons, mainly because the necessary concepts and relations were already present, but also for ease of importing, exporting, and reusing parts of chemical mechanisms. However, this choice may be revisited in the future.

Since the focus of this paper is to establish proof-of-principle of a linked-data approach, the partition function calculations conducted by the Thermodynamic Property Agent we have created for deriving thermo-data from quantum chemistry calculations are relatively basic, ignoring for example hindered rotors. This, however, can in some cases lead to substantial errors. In order to achieve a thermochemical knowledge-base of high quality, it would be necessary to employ more advanced treatments of anharmonicities [28].

7 Conclusions

In this paper, we have connected species occurring in chemical kinetic reaction mechanisms to quantum chemistry calculations in a single knowledge-graph. We have achieved this by creating a new ontology, called OntoSpecies, and integrating it with extended versions of OntoKin, an ontology for mechanisms, and OntoCompChem, an ontology for quantum chemistry calculations. We have implemented four software agents – three knowledge-representational ones which add mechanisms and quantum calculations to the knowledge-graph and create links between them, and a computational one which deduces thermochemical data from quantum calculations. Using these four agents, we have instantiated the three ontologies, *i.e.* created and populated knowledge-bases for them, for an example use-case of a hydrogen combustion mechanism. The created knowledge-graph and agents seamlessly integrate into the knowledge-graph and eco-system of autonomously operating software agents, respectively, of the J-Park Simulator (JPS) – an intelligent Industry 4.0 simulation platform built upon ontology-based linked data. The use of agents for evolving the knowledge-graph in an automated manner is also briefly discussed. Whilst we have chosen here an example from combustion, the methodology is equally applicable to atmospheric chemistry and other areas.

Acknowledgements

This work was partly funded by the National Research Foundation (NRF), Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme, and by the European Union Horizon 2020 Research and Innovation Programme under grant agreement 724145. Markus Kraft gratefully acknowledges the support of the Alexander von Humboldt foundation.

References

[1] American Chemical Society. CAS Registry, 2019. URL https://www.cas.org/support/documentation/chemical-substances. Accessed 17 May 2019.

[2] J. Appel, H. Bockhorn, and M. Frenklach. Kinetic modeling of soot formation with detailed chemistry and physics: laminar premixed flames of C_2 hydrocarbons. *Combustion and Flame*, 121(1):122 – 136, 2000. doi:10.1016/S0010-2180(99)00135-2.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000. doi:10.1038/75556.

[4] L. Atzori, A. Iera, and G. Morabito. The internet of things: A survey. *Computer Networks*, 54(15):2787–2805, 2010. doi:10.1016/j.comnet.2010.05.010.

[5] T. Berners-Lee. Linked data – design issues, 2006. URL https://www.w3.org/DesignIssues/LinkedData.html. Accessed 10 May 2019.

[6] P. L. Bhoorasingh, B. L. Slakman, F. Seyedzadeh Khanshan, J. Y. Cain, and R. H. West. Automated transition state theory calculations for high-throughput kinetics. *Journal of Physical Chemistry A*, 121(37):6896–6904, 2017. doi:10.1021/acs.jpca.7b07361.

[7] C. Bizer, T. Heath, and T. Berners-Lee. Linked data – the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009. doi:10.4018/jswis.2009081901.

[8] Cambridge Environmental Research Consultants (CERC). Atmospheric Dispersion Modelling System (ADMS), 2019. URL http://www.cerc.co.uk/environmental-software/ADMS-model.html. Accessed 16 July 2019.

[9] M. O. Connaire, H. J. Curran, J. M. Simmie, W. J. Pitz, and C. Westbrook. A comprehensive modeling study of hydrogen oxidation. *International Journal of Chemical Kinetics*, 36:603–622, 2004. doi:10.1002/kin.20036.

[10] K. Degtyarenko, P. 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(suppl_1): D344–D350, 2008. doi:10.1093/nar/gkm791.

[11] A. Devanand, I. A. Karimi, and M. Kraft. Optimal site selection for modular nuclear power plants. *Computers & Chemical Engineering*, 125:339–350, 2019. doi:10.1016/j.compchemeng.2019.03.024.

[12] A. Devanand, G. Karmakar, N. Krdzavac, L. K. Aditya, R. Rigo-Mariani, A. Krishnan, E. Y. S. Foo, I. A. Karimi, and M. Kraft. OntoPowerSys: A power systems ontology for cross domain interactions in an eco industrial park, 2019. Submitted for publication.

[13] A. Eibeck, M. Q. Lim, and M. Kraft. J-Park Simulator: An ontology-based platform for cross-domain scenarios in process industry, 2019. Submitted for publication.

[14] F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, and M. Kraft. OntoKin: An ontology for chemical kinetic reaction mechanisms, 2019. Submitted for publication.

[15] J. D. Ferreira, J. Hastings, and F. M. Couto. Exploiting disjointness axioms to improve semantic similarity measures. *Bioinformatics*, 29(21):2781–2787, 2013. doi:10.1093/bioinformatics/btt491.

[16] M. Frenklach. Transforming data into knowledge – Process Informatics for combustion chemistry. *Proceedings of the Combustion Institute*, 31(1):125–140, 2007. doi:10.1016/j.proci.2006.08.121.

[17] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian 16 Revision B.01, 2016. Gaussian Inc. Wallingford CT.

[18] 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):34, 2015. doi:10.1186/s13321-015-0084-4.

[19] C. W. Gao, J. W. Allen, W. H. Green, and R. H. West. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Computer Physics Communications*, 203:212–225, 2016. doi:10.1016/j.cpc.2016.02.013.

[20] M. M. Ghahremanpour, P. van Maaren, and D. van der Spoel. Alexandria library. Zenodo, 2017. doi:10.5281/zenodo.1004711.

[21] C. F. Goldsmith, G. R. Magoon, and W. H. Green. Database of small molecule thermochemistry for combustion. *Journal of Physical Chemistry A*, 116(36):9033–9057, 2012. doi:10.1021/jp303819e.

[22] J. Goodman. Computer software review: Reaxys. *Journal of Chemical Information and Modeling*, 49(12):2897–2898, 2009. doi:10.1021/ci900437n.

[23] D. Hait and M. Head-Gordon. How accurate is density functional theory at predicting dipole moments? An assessment using a new database of 200 benchmark values. *Journal of Chemical Theory and Computation*, 14(4):1969–1981, 2018. doi:10.1021/acs.jctc.7b01252.

[24] 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(10):e25513, 2011. doi:10.1371/journal.pone.0025513.

[25] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7(1):23, 2015. doi:10.1186/s13321-015-0068-4.

[26] D. Hill, N. Adams, M. Bada, C. Batchelor, T. Berardini, H. Dietze, H. Drabkin, M. Ennis, R. Foulger, M. Harris, J. Hastings, N. Kale, P. Matos, C. Mungall, G. Owen, P. Roncaglia, C. Steinbeck, S. Turner, and J. Lomax. Dovetailing biology and chemistry: integrating the gene ontology with the ChEBI chemical ontology. *BMC Genomics*, 14(1):513, 2013. doi:10.1186/1471-2164-14-513.

[27] R. D. Johnson III. NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, Release 19, 2018. doi:10.18434/T47C7Z.

[28] M. Keçeli, S. N. Elliott, Y.-P. Li, M. S. Johnson, C. Cavallotti, Y. Georgievskii, W. H. Green, M. Pelucchi, J. M. Wozniak, A. W. Jasper, and S. J. Klippenstein. Automated computational thermochemistry for butane oxidation: A prelude to predictive automated combustion kinetics. *Proceedings of the Combustion Institute*, 37 (1):363–371, 2019. doi:10.1016/j.proci.2018.07.113.

[29] S. Kim, P. Thiessen, E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. Bryant. PubChem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, 2016. doi:10.1093/nar/gkv951.

[30] M. Kraft, P. Maigaard, F. Mauss, M. Christensen, and B. Johansson. Investigation of combustion emissions in an HCCI engine – measurements and a new computational model. *Proceedings of the Combustion Institute*, 28(1):1195–1201, 2002. doi:10.1016/S0082-0784(00)80330-6.

[31] 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(7):3154–3165, 2019. doi:10.1021/acs.jcim.9b00227.

[32] J. Lai, O. Parry, S. Mosbach, and A. Bhave. Evaluating emissions in a modern compression ignition engine using multi-dimensional PDF-based stochastic simulations and statistical surrogate generation. *SAE Technical Paper No.* 2018-01-1739, 2018. 10.4271/2018-01-1739.

[33] V. R. Lambert and R. H. West. Identification, correction, and comparison of detailed kinetic models. In *9th US National Combustion Meeting*, 2015. URL `https://pdfs.semanticscholar.org/84bc/0933b0c29bdb7960e9106fcc51b6f024451e.pdf`.

[34] H. Lasi, P. Fettke, H. G. Kemper, T. Feld, and M. Hoffmann. Industry 4.0. *Business & Information Systems Engineering*, 6(4):239–242, 2014. doi:10.1007/s12599-014-0334-4.

[35] A. Laskin and H. Wang. On initiation reactions of acetylene oxidation in shock tubes: A quantum mechanical and kinetic modeling study. *Chemical Physics Letters*, 303(1):43 – 49, 1999. doi:10.1016/S0009-2614(99)00242-0.

[36] J. Lee, B. Bagheri, and H. A. Kao. A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3:18 – 23, 2015. doi:10.1016/j.mfglet.2014.12.001.

[37] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015. doi:10.3233/SW-140134.

[38] Y. P. Li, K. Han, C. A. Grambow, and W. H. Green. Self-evolving machine: A continuously improving model for molecular thermochemistry. *The Journal of Physical Chemistry A*, 123(10):2142–2152, 2019. doi:10.1021/acs.jpca.8b10789.

[39] W. Marquardt, J. Morbach, A. Wiesner, and A. Yang. *OntoCAPE - A Re-Usable Ontology for Chemical Process Engineering*. Springer-Verlag Berlin Heidelberg, 1st edition, 2010.

[40] D. A. McQuarrie. *Statistical Mechanics*. Harper & Row, New York, 1976.

[41] A. Miles and S. Bechhofer. SKOS simple knowledge organization system reference. Recommendation, World Wide Web Consortium (W3C), 2009. `http://www.w3.org/TR/skos-reference/`. Accessed 17 May 2019.

[42] J. Morbach, A. Yang, and W. Marquardt. OntoCAPE – A large-scale ontology for chemical process engineering. *Engineering Application of Artificial Intelligence*, 20 (2):147–161, 2007. doi:10.1016/j.engappai.2006.06.010.

[43] S. Mosbach, M. S. Celnik, A. Raj, M. Kraft, H. R. Zhang, S. Kubo, and K.-O. Kim. Towards a detailed soot model for internal combustion engines. *Combustion and Flame*, 156(6):1156–1165, 2009. doi:10.1016/j.combustflame.2009.01.003.

[44] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33, 2011. doi:10.1186/1758-2946-3-33.

[45] M. Pan, J. Sikorski, C. A. Kastner, J. Akroyd, S. Mosbach, R. Lau, and M. Kraft. Applying Industry 4.0 to the Jurong Island eco-industrial park. *Energy Procedia*, 75:1536–1541, 1015. doi:10.1016/j.egypro.2015.07.313.

[46] 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):1–16, 2012. doi:10.1186/1758-2946-4-15.

[47] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014. doi:10.1038/sdata.2014.22.

[48] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J. L. Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012. doi:10.1021/ci300415d.

[49] P. Shvaiko, F. Farazi, V. Maltese, A. Ivanyukovich, V. Rizzi, D. Ferrari, and G. Ucelli. Trentino government linked open geo-data: A case study. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. X. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, editors, *The Semantic Web – 11th International Semantic Web Conference (ISWC), Boston, MA, USA, November 11-15, 2012, Proceedings, Part II*, volume 7650 of *Lecture Notes in Computer Science*, pages 196–211. Springer, Berlin, Heidelberg, 2012. doi:10.1007/978-3-642-35173-0_13.

[50] J. M. Simmie. A database of formation enthalpies of nitrogen species by compound methods (CBS-QB3, CBS-APNO, G3, G4). *The Journal of Physical Chemistry A*, 119(42):10511–10526, 2015. doi:10.1021/acs.jpca.5b06054.

[51] J. S. Smith, O. Isayev, and A. E. Roitberg. ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific Data*, 4:170193, 2017. doi:10.1038/sdata.2017.193.

[52] Y. V. Suleimanov and W. H. Green. Automated discovery of elementary chemical reaction steps using freezing string and Berny optimization methods. *Journal of Chemical Theory and Computation*, 11(9):4248–4259, 2015. doi:10.1021/acs.jctc.5b00407.

[53] B. Wang, S. Mosbach, S. Schmutzhard, S. Shuai, Y. Huang, and M. Kraft. Modelling soot formation from wall films in a gasoline direct injection engine using a detailed population balance model. *Applied Energy*, 163:154–166, 2016. doi:10.1016/j.apenergy.2015.11.011.

[54] B. Wang, P. Dobosh, S. Chalk, M. Sopek, and N. Ostlund. Computational chemistry data management platform based on the semantic web. *The Journal of Physical Chemistry A*, 121(1):298–307, 2017. doi:10.1021/acs.jpca.6b10489.

[55] H. Wang, X. You, A. V. Joshi, S. G. Davis, A. Laskin, F. Egolfopoulos, and C. K. Law. USC Mech Version II. High-Temperature Combustion Reaction Model of H_2/CO/C1-C4 Compounds, 2007. http://ignis.usc.edu/USC_Mech_II.htm.

[56] V. Warth, F. Battin-Leclerc, R. Fournet, P. A. Glaude, G. M. Côme, and G. Scacchi. Computer based generation of reaction mechanisms for gas-phase oxidation. *Computers & Chemistry*, 24:541–560, 2000.

[57] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin Core Metadata for Resource Discovery. RFC 2413, 1998. doi:10.17487/RFC2413.

[58] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi:10.1021/ci00057a005.

[59] 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, 2017. doi:10.1016/j.apenergy.2017.05.002.

[60] 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, 2018. doi:10.1016/j.compchemeng.2018.07.010.

[61] X. Zhou, A. Eibeck, M. Q. Lim, N. Krdzavac, and M. Kraft. An agent composition framework for the J-Park Simulator – a knowledge graph for the process industry, 2019. Submitted for publication.