Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

OntoPESScan: An Ontology for the Exploration of Potential Energy Surfaces

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

released: March 17, 2022

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

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

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

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

[5] The Alan Turing Institute
London
United Kingdom

Preprint No. 294

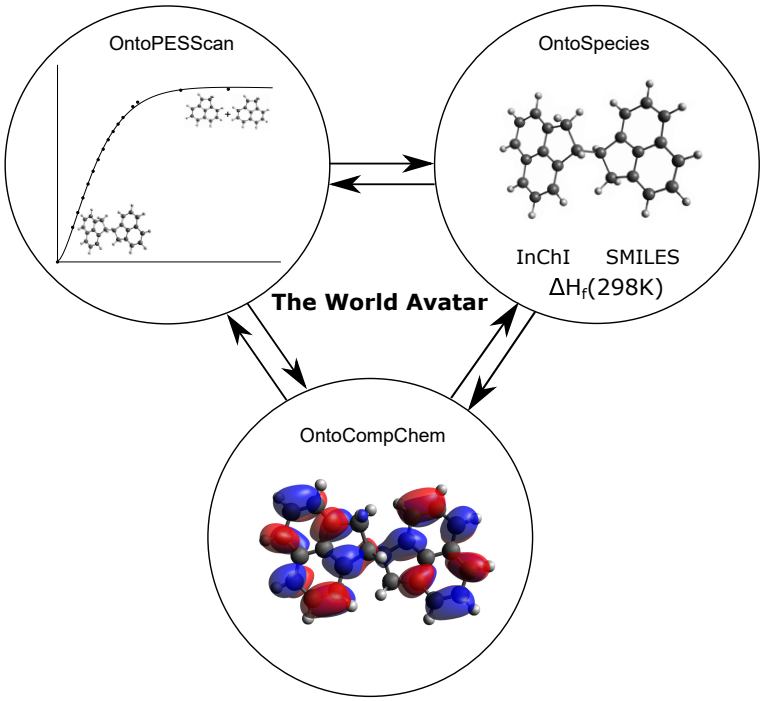
Keywords: Potential Energy Surface, Ontology, The World Avatar, OntoPESScan

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

In this work, a new OntoPESScan ontology is developed for the semantic representation of potential energy surfaces (PES), a central concept in computational chemistry. This ontology is developed in line with knowledge graph principles and The World Avatar (TWA) project. OntoPESScan is linked to other ontologies for chemistry in TWA, including OntoSpecies, which helps uniquely identify species along the PES and access their properties, and OntoCompChem, which allows association of potential energy surfaces with the quantum chemical calculations and concepts used to derive them. A forcefield fitting agent is also developed that makes use of the information in the OntoPESScan ontology to fit force fields to reactive surfaces of interest on the fly by making use of the empirical valence bond methodology. This agent is demonstrated to successfully parametrise two cases, a PES on ethanol, and a PES on a localized π-radical PAH hypothesized to play a role in soot formation during combustion. OntoPESSCan is an extension to the capabilities of TWA, and in conjunction with potential further ontological support for molecular dynamics and reactions, will further progress towards an open, continuous, and self-growing knowledge graph for chemistry.



Highlights

- The OntoPESScan ontology is developed for representing potential energy surfaces.
- OntoPESScan is linked to the existing OntoSpecies and OntoCompChem ontologies.
- A forcefield fitting agent is developed to fit data in OntoPESScan on-the-fly.
- The agent successfully fits forcefields to surfaces of ethanol and π-radical PAHs.

Contents

1 Introduction

The potential energy surface (PES) is one of the key concepts of computational chemistry, representing the relationship between the energy of a molecule or system of molecules, and the geometry or coordinates of said system [48]. The potential energy surface thus allows representations and changes in a molecule or system of molecule's shape and geometry to be related to the system electronic energy, enabling the application of Schrödinger's equation to molecules. As a consequence, determination and descriptions of potential energy surfaces find a wide variety of use in computational chemistry. This includes accurate computation of rate coefficients of chemical reactions, as chemical reactions are naturally represented by potential energy surface and a good description of the PES is necessary to compute the rate of the reaction described by the surface [53]. Potential energy surfaces are also central to the development of force fields for molecular dynamics simulations, which must capture how interactions between different chemical species result in changes in system energies, thus requiring representation of the underlying PES [4]. Achieving this all requires chemical data on potential energy surfaces, reactions, and the chemical species they describe.

An increasing amount of chemical information is stored in databases online to help facilitate the sharing and manipulation of such data across all corners of computational chemistry [52]. Examples include the Computational Chemistry Comparison and Benchmark DataBase (CCCBDB) from the National Institute of Standards and Technology (NIST) which houses energetic, vibrational, and thermochemical information for a variety of species determined by various initio quantum chemistry methods [37] and the Alexandria library of calculations for force field development [28]. Well known general chemistry databases such as Pubchem [42, 43] also provides information on cheminformatics identifiers to uniquely define and link different data on chemical species as well as geometry and crystal structure information relevant to computational chemistry efforts. Information science and mathematical methods such as graph theory and machine learning to such chemical data is increasingly used to further progress, with examples seen in organic reaction network analysis [20, 36], predictive combustion chemical kinetics [21], using machine learning to suggest retrosynthetic pathways in Reaxys [29, 47], and applying machine learning to fit potential energy surfaces [59].

A key group of methods that facilitate the access and manipulation of such chemical data are Semantic Web technologies. Semantic Web approaches such as knowledge graphs (KGs) and ontologies provide frameworks for the storage, representation, and annotation of information in an consistent and well-defined manner. They also allow for clear logical approaches in the querying and manipulation of data, and have been seeing increasing use in chemistry. Some key examples include the chemical information ontology [30], the chemical ontology built on methontology [49], the ChEBI ontology for information on chemical species of biological interest and applications, [11, 31, 32], and PubChem's RDF representation of its data and annotations [25, 43]. Additionally, there are several chemistry related ontologies as part of the dynamic, cross-domain knowledge graphs comprising the J-Park simulator (JPS) and The World Avatar (TWA) projects [15, 44]. The World Avatar is discussed in detail in section 2, but briefly, the chemistry ontologies include OntoCompChem for representing quantum chemistry calculations [45], OntoSpecies for

representing chemical species [19], OntoKin for representing chemical reaction mechanisms [18], and OntoChemExp for representing chemical experiments [2, 3].

The purpose of this paper is to extend the chemistry section of The World Avatar knowledge graph by developing and introducing an ontology for the representation of potential energy surfaces, OntoPESScan. This enables another key set of computational chemical data to be integrated in the dynamic knowledge graph of TWA, and also enables leveraging of the linked data principles on which the knowledge graph is built. This includes, for example linking chemical species seen in the potential energy surfaces with the data and properties available in OntoSpecies, and providing additional information on the methodology and results of potential energy surface scans through OntoCompChem. The ontological structure of OntoPESScan is outlined, and it's ability to describe and represent data on different types of potential energy surface scans is demonstrated. Additionally, an agent is constructed that can query OntoPESScan and linked chemistry ontologies to fit a force field based on the data found on the scan by making use of the empirical valence bond (EVB) method [39]. The development of OntoPESScan and the construction of the force field agent serve to show how the knowledge graph framework can further progress data automation in chemistry, as well as the representation and analysis of potential energy surfaces and force fields that find wide applications in chemistry.

2 The World Avatar

2.1 Overview

The World Avatar is a dynamic knowledge graph (dKG) which makes use of linked data principles and the semantic web to represent concepts as nodes of a graph, and relationships between said concepts as vertices of a graph. The linking of information is a key strength of a dKG, as it enables concepts and information across different domains to be linked together. The World Avatar makes use of ontologies to define the relationships and key concepts for a given domain or use case in what is known as the terminological component, or TBox. Instances, data, and facts about the concepts and relations in the TBox form the assertion components, or ABox. These are often instantiated in a triple store, which contains the formal subject-predicate-object relations for the data.

The aim of The World Avatar is to eventually have a digital representation of every real world concept or relation. This can be thought of as a Universal Digital Twin, with the original concept of Digital Twins arising in smart city planning [17] and infrastructure development [7] being extended to all domains. The World Avatar itself has its origins in the J-Park Simulator (JPS), which strove to create virtual representations of the processes and industrial entities on the Jurong Island Eco-Industrial Park in Singapore [44]. The JPS is an example of the well-known industry 4.0 concept, where each industrial device or process has a digital representation to help facilitate what-if case analysis [46]. The JPS was developed with the dynamic knowledge graph principles in mind, containing several ontologies whose data could be semantically linked, with the linked data concept being implemented through the use of generalized web addresses known as Internationalised Resource Identifiers (IRIs).

A key aspect of dynamic knowledge graphs are agents, autonomous software that can continuously and independently act on the knowledge graph, essentially leveraging the data and structure of the KG to perform 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 (ABoxes) in the KG, updating existing instances in the KG with improved information where appropriate, and updating the concepts, definitions of concepts, and relationships between concepts in the Ontological TBoxes. Agents perform these tasks with the aim of producing a self-growing, self-updating, and self-improving knowledge graph. Additionally, agents can also compose composite agents to answer more complex queries or perform more complex tasks. Agents themselves are integrated as part of The World Avatar knowledge graph by means of an ontology that semantically describes agents, OntoAgent [74], and a market for using and identifying new agents [75].

Ontologies, instances, and agents form the main components of the knowledge graph, illustrated schematically in Figure 1.

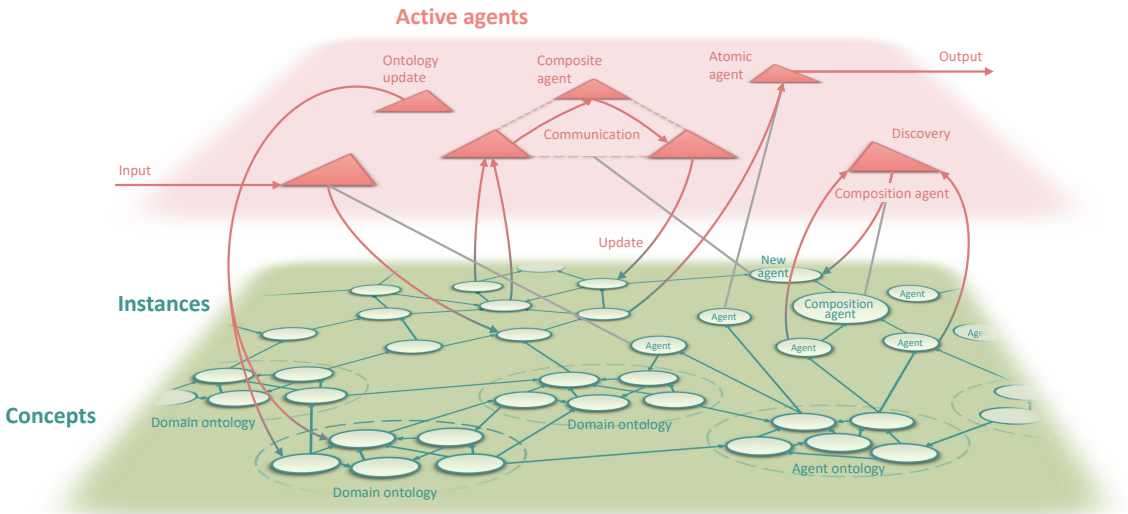


Figure 1: *Structure of The World Avatar knowledge graph. Image reproduced from Akroyd et al. [1] under a CC BY 4.0 licence.*

The World Avatar 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 [50] as well as OntoEIP for describing the functions and interactions underlying Eco-Industrial parks [71–73]. 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 [15] and the Weather Ontology [64]. Finally, several ontologies have been developed in the chemistry domain including OntoKin [18], OntoSpecies [19], OntoCompChem [45], and OntoChemExp [2], which are discussed in detail in the next section.

Several examples of agents and cross-domain applications also exist in TWA. This in-

cludes the thermochemistry agent that determines thermochemical parameters for chemical species using information in OntoSpecies, which has then been combined with agents for modelling engines using kinetics in OntoKin and an Atmospheric Dispersion Agent to demonstrate how underling chemical mechanisms in ship engines can impact predicted pollutant distributions in nearby cities [55]. Additional agents include those that can predict the power conversion efficiency of organic solar cells based on SMILES strings and organic donor properties in OntoSpecies in conjunction with machine learning approaches [16] as well as several city planning agents [9]. Other cross domain examples include ELChemo, which made use of an OntoTwin ontology to combine the chemical and electrical components of industrial power plants by linking relations and information in OntoEIP with OntoPowerSys [13], and efforts to use ontologies and the knowledge graph as the foundation for a universal digital twin of the United Kingdom [1] by considering virtual representations of everything from land use and infrastructure to gas and power systems. There are also several links between the chemistry domain ontologies, discussed further below.

2.2 Chemistry in The World Avatar

This section summarizes the current main ontologies and semantic web technologies in TWA that support the chemistry domain.

2.2.1 OntoSpecies

The OntoSpecies ontology is meant to serve a central chemistry ontology, with entries in the ontology consisting of unique chemical species and some of their key properties. This is in line with other major chemistry resources such as Pubchem [43] for example. The full ontology is found at `http://www.theworldavatar.com/ontology/ontospecies/OntoSpecies.owl`. Each entry in OntoSpecies is assigned a unique IRI. This is done so that the IRI serves as a way to uniquely identify a chemical species and enable linking a chemical species and its associated information to instances and concepts in other related ontologies. The basic properties of species stored in OntoSpecies include semantic descriptions of molecular formula, charge, molecular weight, and spin multiplicity. Isotopes, different charges, and different spin states are treated as different chemical species to the most common standard state. A key thermodynamic property of a chemical species in the standard enthalpy of formation is also represented in OntoSpecies, as this has use in several reactor simulations. The reference temperature, reference state and provenance for the standard enthalpy of formation are also included in the ontology to add the necessary context of where the value in the ontology comes from.

OntoSpecies also includes concepts for key cheminformatics identifiers widely used in the field, namely InChI and its associated hash representation (InChIKey) from IUPAC [33] and SMILES [67, 68]. This enables users to search for entries in OntoSpecies by their InChI or SMILES. Other identifiers supported in OntoSpecies include those from major chemistry databases in PubChemCID for PubChem and CASID for the chemical abstracts service where available. This facilitates searching for additional information on a chemical species in these external resources if it is not found in OntoSpecies. The final

category of concept in OntoSpecies consists of geometric properties of a species. This includes a list of bonds in the species, as well as a semantic representation for the full three-dimensional geometry of a species that serves as a curated reference geometry. This reference geometry serves two purposes. First, it provides unique IRIS for each atom in a chemical species, which means these atoms can also be uniquely identified in addition to just the overall chemical species. This can be important when comparing different quantum chemical calculations on the species where the order of atoms is often flexible. Additionally, this geometry can also be used as a starting point for the aforementioned quantum chemical calculations for an agent. Whilst it is possible for geometries to be derived from InChI and SMILES strings through format translators and molecular force fields as implemented in OpenBabel [56], this approach can encounter issues when dealing with metals for example, and so having a ready out-of-the-box geometry can prove advantageous in such situations.

2.2.2 OntoKin

The OntoKin ontology is used to semantically represent chemical reaction mechanisms. The full ontology is published at `http://www.theworldavatar.com/ontology/ontokin/OntoKin.owl` and covered in detail in Farazi et al. [18]. The key concepts in OntoKin essentially cover those that are required to describe the most common chemical mechanism format, with CHEMKIN [41] in particular being a main reference. Concepts in OntoKin include definitions of a chemical mechanism, which consists of a set of chemical reactions. Chemical reactions occur among different chemical species that consist of chemical elements and react in ratios defined by stoichiometric coefficients and can be reversible and irreversible. Additionally, the phase concept is also defined in OntoKin, enabling representation of both gas phase and surface (solid) phase reactions. The rates of these reactions are represented using rate models (i.e. Arrhenius-type) which are used to compute rate-coefficients. Thermodynamic and transport model concepts are also associated with species. Of note, multiple rate models can be associated with the same reaction, as can multiple thermodynamic or transport models with a given species. This allows OntoKin to support facile comparison of different rate thermodynamic, or transport models in the literature for the user [19]. Additionally, species in OntoKin can be linked to a species instance in OntoSpecies, enabling a unique identification. This helps resolve naming inconsistencies of species across different chemical mechanisms where benzene could be called "A1" in one mechanism but "C_6H_6" in another [19].

2.2.3 OntoCompChem

The OntoCompChem ontology is used to semantically represent the results and details of computational chemical calculations. The full ontology is published at `http://www.theworldavatar.com/ontology/ontocompchem/ontocompchem.owl` and covered in detail in Krdzavac et al. [45]. OntoCompChem includes concepts that cover the inputs and outputs of the most common computational chemistry calculations and builds on concepts defined in the Gainesville Core (GNVC) ontology [57]. On the input side, this includes the level of theory used to perform the calculation in terms of the functional

(i.e. B3LYP) and basis set (i.e. 6-31G(d)), as well as charge and multiplicity. These are inputs that virtually all computational chemistry packages require. OntoCompChem also includes classes that help define what program is used to perform the quantum chemical calculation. Currently, such classes are available for the widely used Gaussian programs such as Gaussian09 and Gaussian16 [22–24] but is easily expandable to include other programs.

On the output side, the results of single point energy, geometry optimization, and frequency calculations are represented. For energies, this includes the final converged self-consistent-field (SCF) energy, and if an accompanying frequency calculation is present, the zero-point energy correction is also included. A recent addition also means that frontier orbital energies and near orbital energies are also represented, namely the HOMO-2, HOMO-1, HOMO, LUMO, LUMO+1, and LUMO+2 energies. For geometry optimization calculations, a full 3D representation of the optimized geometry is present in OntoCompChem with full coordinate values, atom types, and rotational constants all included. For frequency calculations, a full list of the computed vibrational frequencies are represented enabling easy confirmation of the type of stationary point a geometry or calculation corresponds to. Finally, much like OntoKin, OntoCompChem calculations also contain a link to OntoSpecies in terms of the "hasUniqueSpecies" concept, which points to an IRI in OntoSpecies and defines what species the calculation is run on. This supports a clear way to group together different quantum chemistry calculations performed on the same species, making it straightforward to compare how different methodologies impact the results.

2.2.4 OntoChemExp

The OntoChemExp ontology is developed to enable semantic representation of chemical experiments. The full ontology is published at `http://theworldavatar.com/ontology/ontochemexp/OntoChemExp.owl` and is covered in detail in Bai et al. [2]. Concepts were initially developed with representation of combustion experiments in mind, but the four-modular structure is broadly applicable. This structure includes the experiment module, which includes the instance of an experiment and its associated metadata in terms of source or publication from which the experiment and its data is taken from. OntoChemExp also includes setup modules where details on the apparatus and key experimental conditions are represented, a results module where the data collected from the experiment is abstracted in terms of a data group for each independent variable and data points that are measured in each data group, and finally a specification module where concepts such as values and uncertainties are defined that can then be assigned to conditions or data instances in the setup or results section. As with OntoKin and OntoCompChem, the "hasUniqueSpecies" concept is present to enable unambiguous identification of chemical species involved in experiments by connection to a unique OntoSpecies IRI.

2.2.5 Marie

The World Avatar also includes the Marie website, which serves as a question-and-answer system for the chemical information in the KG [75]. Marie was developed with the inten-

tion of lowering the barrier to interacting with the knowledge graph and various ontologies. One typically accesses information in the KG through query construction languages such as SPARQL [58]. Whilst SPARQL allows for logical and complex queries to be carried out, it can be a barrier for unfamiliar users, and so Marie provides a natural question-and-answer approach to access by which users can request or search for information much like they would in a typical search engine like Google or Wolfram Alpha. Marie makes use of natural language processing (NLP) techniques to map the question asked by the user to SPARQL queries that traverse the knowledge graph to find the information that answers the question. Marie can currently answer several types of questions on chemistry including finding kinetic and thermodynamic properties for species or finding out what reactions a species is involved in.

A recent extension also leverages the semantic framework for agents, where if a user's question is not answerable purely based on the static data in the KG, Marie will automatically identify and invoke an appropriate agent to try and derive an answer based on the information in the KG. The NLP framework also has an automated approach to help identify new question types that can be answered when a new agent is added in the KG. An example of this is calling the thermochemistry agent to calculate the thermodynamic properties of a chemical species requested by a user. Even if the thermodynamic properties are not explicitly stored in the KG, if the thermochemistry agent can identify an appropriate quantum chemical calculation for the requested species, it can calculate the required thermodynamic properties and return them to the user. As more instances, ontologies, and agents are added to TWA, Marie will automatically imrpove to answer a wider variety of questions from the user.

The World Avatar does currently include ontologies that can represent a variety of chemical concepts for species and their properties, reactions and kinetics, quantum chemistry calculations, and chemical experiments. However, representation of potential energy surfaces is not included in the above ontologies, and is a key concept in computational chemistry, derivation of reaction rates, and the large field of molecular dynamics. Thus, OntoPESScan is developed to fill this gap whilst also naturally linking to existing ontologies.

3 The OntoPESScan ontology

3.1 Main Structure

The OntoPESScan ontology is developed to be a compact representation of the key concepts necessary to semantically represent scans and explorations of potential energy surfaces. The linked data principles of the knowledge graph are used as guidelines so that information is not unnecessarily duplicated in this new ontology. This is achieved by linking the OntoPESScan ontology to existing chemistry ontologies in The World Avatar, namely OntoSpecies and OntoCompChem. The structure of the OntoPESScan ontology is shown in Figure 2. The full terminological component (TBox) of the ontology containing the full class and relational definitions is available at `http://theworldavatar.com/ontology/ontopesscan/OntoPESScan.owl`.

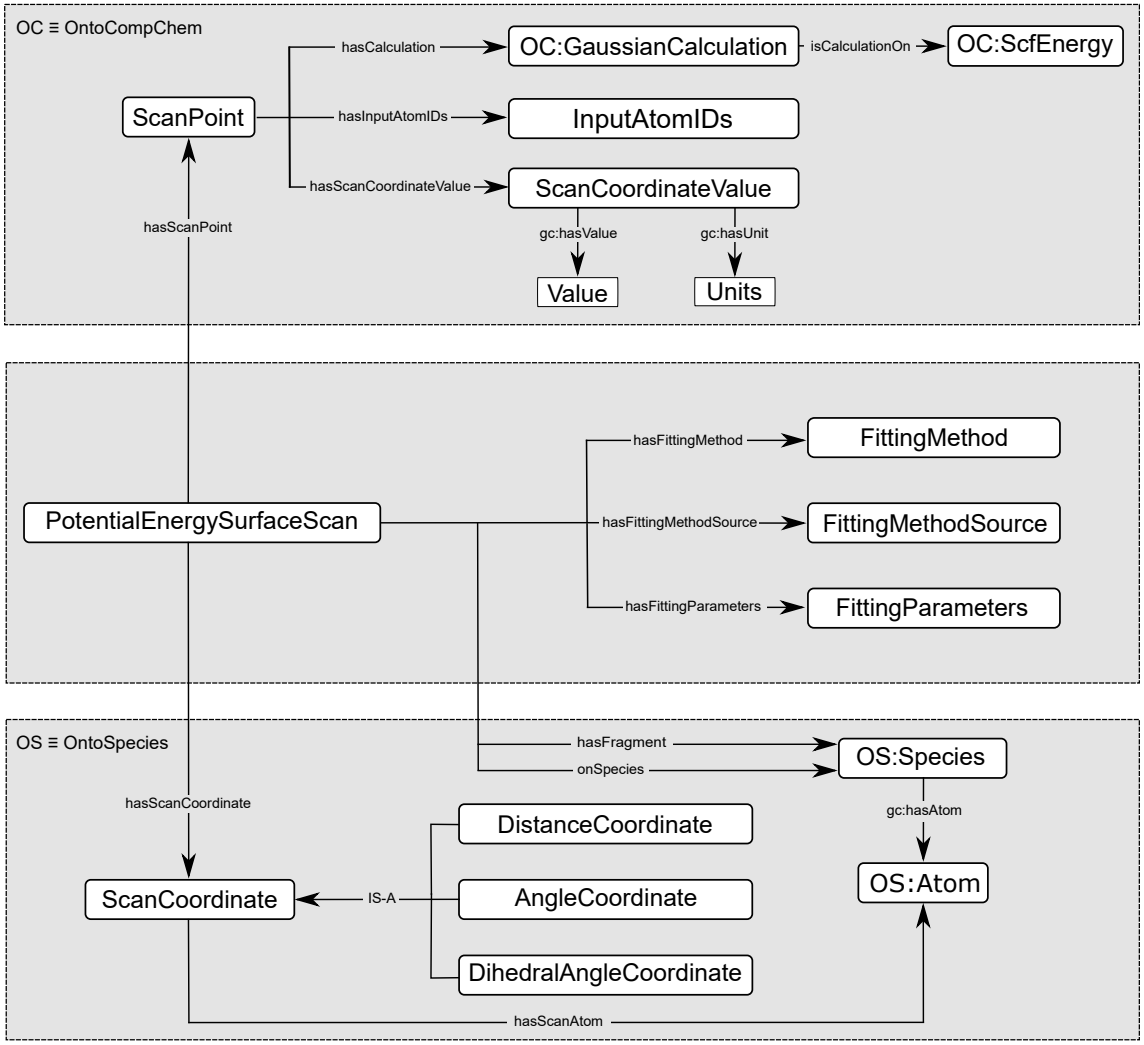


Figure 2: *The structure of the OntoPESScan ontology developed in this work. The main classes and concepts are illustrated by boxes, with the relations between these classes represented by the arrows. Links to OntoSpecies (OS) and OntoCompchem (OC) are also shown.*

Figure 2 shows that the OntoPESScan TBox broadly contains three main conceptual sections. The first, in the centre contains the main class of the ontology namely "PotentialEnergySurfaceScan". This class is used to define instances of scans on potential energy surfaces in the Ontology. Connected to this class are three data properties, "hasFittingMethod", "hasFittingMethodSource", and "hasFittingParameters" which each connect the PotentialEnergySurfaceScan instance to string instances that provide contextual descriptions of the method or methods that are used to fit the potential energy surface scan instance. An example of this would be fitting a Morse potential to an intermolecular interaction, as is often done when using potential energy surface scans to compute the rate of reaction between two radicals in models such as the Gorin Model [62]. In this case, the "hasFittingMethod" data property would link the PotentialEnergySurfaceScan instance for the reaction of interest to "MorsePotential", with the "hasFittingMethodSource" being

assigned to a specific publication or reference that describes the Morse Potential method in detail. The "hasFittingParameters" would then be the Morse Potential parameters, typically D_e for the well depth, r_e for the equilibrium bond distance, and a for the width of the potential well, and their values.

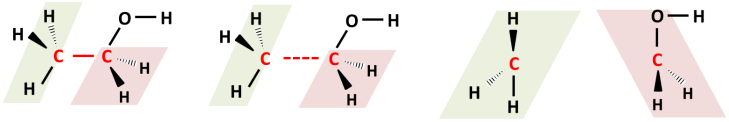
The next key relations are object properties that connect instances of the PotentialEnergySurfaceScan to instances of Species in the OntoSpecies ontology. These are the "onSpecies" and "hasFragment" object properties. The onSpecies object property enables a user to query the scans in OntoPESscan by what species the scan is performed on, and makes use of the OntoSpecies species IRIs being unique to resolve any ambiguities. This is an identical approach to what is adopted in OntoCompChem, OntoKin, and OntoChemExp as discussed previously. The hasFragment property serves a similar to onSpecies, but accounts for the fact that a potential energy surface can involve multiple species. Common examples of this would include bond forming processes between two radicals or the reverse process of bond breaking in a chemical species resulting in two species as products. As a result, the PotentialEnergySurfaceScan class needs to be able to be semantically linked to multiple OntoSpecies Species, and the hasFragment object property enables this.

3.2 Scan Coordinates

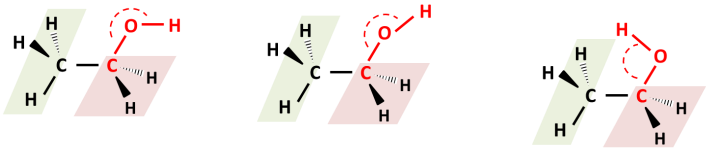
The second main conceptual section in Figure 2 concerns the definition of the scan coordinate, which is essentially the geometric variable with which the energy of the chemical system varies along the the potential energy surface. The central class is thus "ScanCoordinate", and instances of this class are linked to instances of PotentialEnergySurfaceScan through the object property "hasScanCoordinate". Three subclasses of Scan Coordinate are defined in the OntoPESScan TBox, inheriting the properties of ScanCoordinate through the IS-A relation. These are "DistanceCoordinate", "AngleCoordinate", and "DihedralAngleCoordinate" with each of these coordinates defining a different geometric type of scan.

Figure 3 illustrates three different scans on an ethanol molecule that correspond to the three scan coordinate subclasses defined in OntoPESScan. The first is a potential energy surface derived from a scan along a bond, with the scan coordinate in this case being the carbon-carbon bond length in an ethanol molecule. This would be defined by a "DistanceCoordinate" instance. The DistanceCoordinate instance would then need to be defined in terms of the atoms that comprise this coordinate, namely the two carbon atoms in ethanol. To uniquely define the scan coordinate, OntoSpecies is again used, with instances of the ScanCoordinate class being linked to atom IRIs in OntoSpecies through the "hasScanAtom" object property. By using IRIs, the problem of different users uploading variations of the same scan on ethanol but with different ordering of atoms in their computational calculation is circumvented. This is because there is a unique instance for ethanol in OntoSpecies, with each atom having their own unique IRI and fully defined coordinates, enabling unique identification of the atoms. Such bond scans have a variety of applications, being necessary to compute rate constants for bond-forming or bond-breaking reactions as in Smith and Golden [62] for example. They are also crucial when developing reactive force fields such as ReaxFF [60], where they are used as references for fitting the necessary force field parameters that can then be applied to larger systems

a) Scan along a bond:



b) Scan on an angle:



c) Scan of a dihedral or torsional angle:

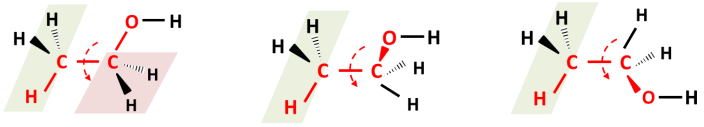


Figure 3: *Examples of the three main different types of scans represented semantically in OntoPEScan, shown for ethanol. This includes a) bond scans b) plane angle scans, and c) dihedral angle scans. The atoms that define the scan coordinate are displayed in red.*

of molecules.

The second type of scan is a potential energy surface derived from a scan along a plane angle, in this case the angle formed by the carbon-oxygen bond and the oxygen-hydrogen bond in ethanol, illustrated in b) in Figure 3. This would be represented by an Angle-Coordinate instance, which would be connected to the appropriate carbon, oxygen, and hydrogen atom in OntoSpecies through the hasScanAtom object property as with a bond scan. This is a less commonly used type of scan, but are necessary for parametrizing bond-bending and flexural interactions in reactive force fields [60], and angle changes do occur in organic dehydration and dehydrogenation mechanisms. The third type of scan corresponds to a potential energy surface derived from a scan along a dihedral or torsional angle, in this case illustrated for the hydrogen-carbon-carbon-oxygen dihedral in ethanol. This would be represented by a DihedralAngleCoordinate instance, with four atoms from the OntoSpecies entry for ethanol being connected to the instance in OntoPESScan in this case, akin to the previous bond and plane angle cases. Dihedral and torsional angles are also widely used, again being crucial for parametrizing reactive force fields, but also for deriving potential energy surfaces and rate constants for cis-trans isomerisation reactions as shown in Figure 3 c) for ethanol. Fitting of dihedral angle scans are also used to apply hindered rotor corrections to partition functions when computing rate constants which is seen in several rate coefficient determining computer programs such as MultiWell [6], Reaction Mechanism Generator [26], and VaReCoF [27].

3.3 Scan Points along the PES

The third main conceptual section in Figure 2 concerns the definition of the scan point. Here, an instance of the ScanPoint class is connected to an instance of PotentialEnergySurfaceScan through the "hasScanPoint" object property, and allows for each point on a potential energy surface to be its own instance, whilst having a common association with a single scan. Each scan point is connected to an instance of the "ScanCoordinateValue" class, through the "hasScanCoordinateValue" object property. The ScanCoordinateValue instance is then linked to associated value and unit through data and object properties borrowed from the GainesVilleCore ontology [57] and unit definitions from NASA's QUDT ontology [34]. For example, the first scan point along the carbon-carbon bond scan in ethanol would have its own instance, with this instance being linked to a ScanCoordinateValue instance, which would then have a value of 1.51 and a unit of the Ångström class from QUDT.

The ScanPoint section also contains the "hasCalculation" Object property, which connects a ScanPoint instance to an instance of the GaussianCalculation class defined in OntoCompChem. This enables OntoPESScan to make use of the classes and relationships defined in and the data stored in OntoCompChem, and essentially means that each ScanPoint in a PotentialEnergySurfaceScan is associated with its own entry in OntoCompChem. Through OntoCompChem, properties such as the SCF energy shown in Figure 2 and full three-dimensional geometry are available for each scan point, meaning these concepts and this data does not need to explicitly stored and duplicated in OntoPESScan. This link also helps distinguish between two commonly supported types of scans that can be performed using computational chemistry packages, namely relaxed scans where a full geometry optimization is carried out at each point along the scan, and rigid scans, where only the defined scan-coordinate is modified while leaving other geometric coordinates unchanged where possible. These two types of scans are illustrated in Figure 4:

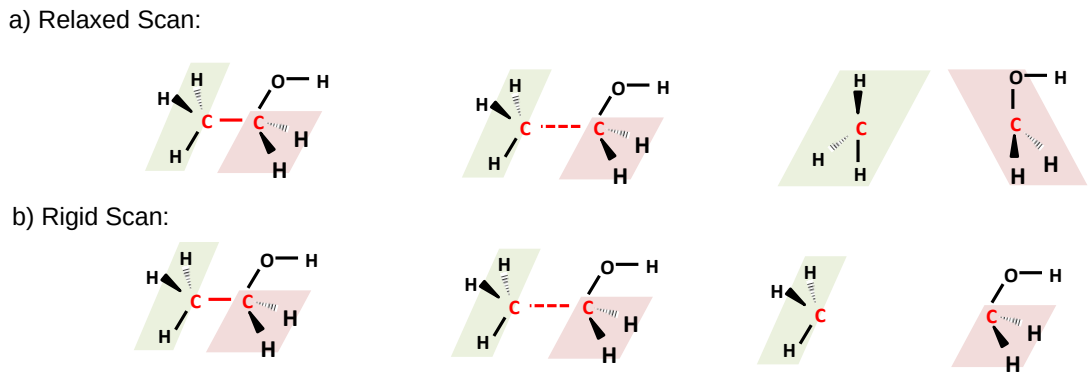


Figure 4: *Illustration of a relaxed carbon-carbon bond scan on ethanol (a) and of a rigid carbon-carbon bond scan on ethanol (b).*

Figure 4 illustrates the difference between a relaxed and rigid scan for the carbon-carbon bond scan in ethanol. The fragments at a long distance have their geometries changed from when they are bonded together in ethanol, as a geometry optimization is performed

for each point along the surface. In contrast, for a rigid scan, the bond length is elongated but the fragments do not change in geometry. Notably, in terms of scan-coordinate the atoms that define the scan coordinate are identical, and the values of the scan coordinate can also be the same. However, they can be distinguished by differences in the electronic energy and of course the geometries along the surfaces will be substantially different, meaning such scans can be distinguished by leveraging the linked data in OntoCompChem. To facilitate this, a simple data property, "hasInputAtomIDs" links a "ScanPoint" to a string that defines what atom indices in the input quantum chemistry calculation file were used to define the scan to make querying and finding information in the associated OntoCompChem entry simpler. Each ScanPoint having their own OntoCompChem calculation and InputAtomIDs also supports definitions of a potential energy surface scan from multiple different quantum chemistry jobs in addition to the cases where the scan is performed in a single job, giving flexibility.

3.4 Population and Querying

Given the ontological structure, an ABox of OntoPESScan in the knowledge graph must also be accompanied by one ABox in OntoCompChem for every ScanPoint linked to the PotentialEnergySurfaceScan instance. To help with creation of and uploading entries to the KG, a set of software agents have been developed to process output log files from a Gaussian program calculation and create the necessary OWL files for both the OntoPESScan and OntoCompChem instances, streamlining the population process. These are freely accessible at https://github.com/cambridge-cares/TheWorldAvatar The OntoSpecies speciesIRI that the OntoPESScan entry is linked to provided as input as well so that the OntoPESScan instance in the KG is linked to both OntoCompChem and OntoSpecies at upload. This enables utilizing information in these ontologies when querying OntoPESScan. For example, a federated query can be used to first search OntoSpecies for Species instances with an InChI that matches that of CO_2, and then searching OntoPESScan for PotentialEnergySurfaceScan instances that have this species instance as the target of onSpecies. This essentially finds scans in the KG that are performed on CO_2:

```
PREFIX OntoPESScan: <http://www.theworldavatar.com/ontology/
   ontopesscan/OntoPESScan.owl#>
    PREFIX OntoSpecies: <http://www.theworldavatar.com/
       ontology/ontospecies/OntoSpecies.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX gc: <http://purl.org/gc/>
    SELECT DISTINCT ?pesIRI
    WHERE {
  SERVICE <http://www.theworldavatar.com/blazegraph/namespace
     /ontospecies/sparql>  {
  ?species OntoSpecies:inChI "InChI=1S/CO2/c2-1-3"^^xsd:
     string . }
  ?pesIRI OntoPESScan:onSpecies ?species .
}
```

Listing 1: *Query to retrieve scan instances on CO_2.*

Similarly, a federated query can also be used to retrieve the electronic energies from OntoCompChem at each scan point for one of PotentialEnergySurfaceScan instances found in the previous query:

```
    PREFIX OntoPESScan: <http://www.theworldavatar.com/ontology/
       ontopesscan/OntoPESScan.owl#>
        PREFIX OntoCompChem: <http://www.theworldavatar.com/
           ontology/ontocompchem/ontocompchem.owl#>
    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
        PREFIX gc: <http://purl.org/gc/>
        SELECT ?scanval ?elenval
        WHERE {
<http://www.theworldavatar.com/kb/ontopesscan/
   PotentialEnergySurfaceScan_b6d609c6-fe80-4bb8-aaf7-
   d4675ffd2638> OntoPESScan:hasScanPoint ?scanpoint .
       ?scanpoint OntoPESScan:hasScanCoordinateValue ?
          scanvalueIRI .
       ?scanpoint OntoPESScan:hasCalculation ?ccIRI .
       ?scanvalueIRI <http://purl.org/gc/hasValue> ?scanval .
       SERVICE <http://www.theworldavatar.com/blazegraph/
          namespace/ontocompchem_test/sparql>  {
       ?ccIRI gc:isCalculationOn ?scfenIRI .
           ?scfenIRI gc:hasElectronicEnergy ?elenIRI .
           ?elenIRI gc:hasValue ?elenval. }
    }
        ORDER BY ASC(?scanval)
```

Listing 2: *Query to retrieve electronic energies for each scan point in a potential energy surface.*

Overall, the OntoCompChem ontology is designed to concisely represent the key concepts and relations for scans on potential energy surfaces. It makes use of links to OntoSpecies to uniquely identify species in the scan and atoms in the scan coordinate, and enables the user to query for any associated species properties such as InChI for example. OntoPESScan is also connected to OntoCompChem, so that properties defined in OntoCompChem such as geometry or SCFEnergy are available for each scan point, which allows this information to be accessed without duplication among ontologies.

4 Force Field Fitting Agent

As mentioned previously, determination and descriptions of potential energy surfaces find a wide variety of use in computational chemistry. This includes accurate computation of rate coefficients of chemical reactions [53] and the development of force fields for molecular dynamics (MD) simulations [4]. Achieving this requires chemical data on potential energy surfaces, reactions, and the chemical species they describe.

In this work, we show how to benefit from the data stored in the KG to parametrize reactive force fields as an initial demonstration. Designing reactive force fields for MD simulations is very challenging, since it necessarily involves dealing with a multi-dimensional problem, where the interactions need to be modelled by highly complicated functional forms with many strongly coupled parameters that are optimised via a difficult search. Unfortunately, a general parametrization for the commonly used reactive force fields is not available yet and, instead, parameters are tuned to specific chemical systems and environments. In this work, in order to avoid the complications of building reliable reactive force fields, we selected the Empirical Valence Bond (EVB) force field coupling method [14] to join two classical force fields from literature, forming a truly reactive force field as a result. The main advantage of the EVB method when compared to reactive force fields is that simulation of reactive processes is conducted via the coupling of non-reactive force fields that are already available in literature to describe the chemically different states. Moreover, compared to the fitting of reactive force fields that requires a large set of quantum mechanical structure and energy data, fitting the EVB force field only requires the potential energy surface for the reaction of interest to the study. In addition, despite the initial task of calibrating the coupling terms against reference data, research has demonstrated that these couplings are invariant to the surrounding electrostatics, making it possible to simulate the same reactive unit in different environments [35]. These features of the EVB method have increased its recognition as a practical and reliable tool within the computational chemistry community [14, 39]. However, as soon as the KG will be populated with more data, the fitting procedure developed in this work can be easily extended to calibrate other force fields and functional forms.

In the EVB method, a classical force field is assigned to any different chemical state for the system. The EVB method defines a Hamiltonian whose matrix representation has each of the computed energies of the involved chemical states as diagonal components, whereas the off-diagonal terms are given by the coupling terms between the force fields in the reaction. Matrix diagonalization at each time step allows computation of reactive energy landscapes that account for the change in chemistry when sampling conformations between the participating chemically different states. More details on EVB theory are reported in the Appendix A.1.

The EVB coupling term calibration is performed by an executable that is designed to accept the energies and geometries of each scan point as input to perform parameter estimation for the target force field. To achieve this, the EVB executable calculates the energies and adjusts the parameters within the target force field to replicate the potential energy surface. This framework is intended to act as an agent within TWA ecosystem, following the agent template designed by Mosbach et al. [55] with few changes. A Unified Modeling Language (UML) activity diagram of the Force Field Fitting agent is provided in Fig. 5. The process starts by querying the information required by the executable from the knowledge graph giving as input an OntoPESScan IRI. SCF energies and geometries of each scan point are retrieved from the OntoCompChem instances linked to the OntoPESScan entry. The agent creates the input files in the format required by the executable and transfers them to the HPC platform. A SLURM job is set up and submitted to the HPC system. The job is then monitored using a status file associated to job. Finally, in case of a successful run, the fitting parameters for the force field are then retrieved by the agent and added to the appropriate scan in the knowledge-graph.

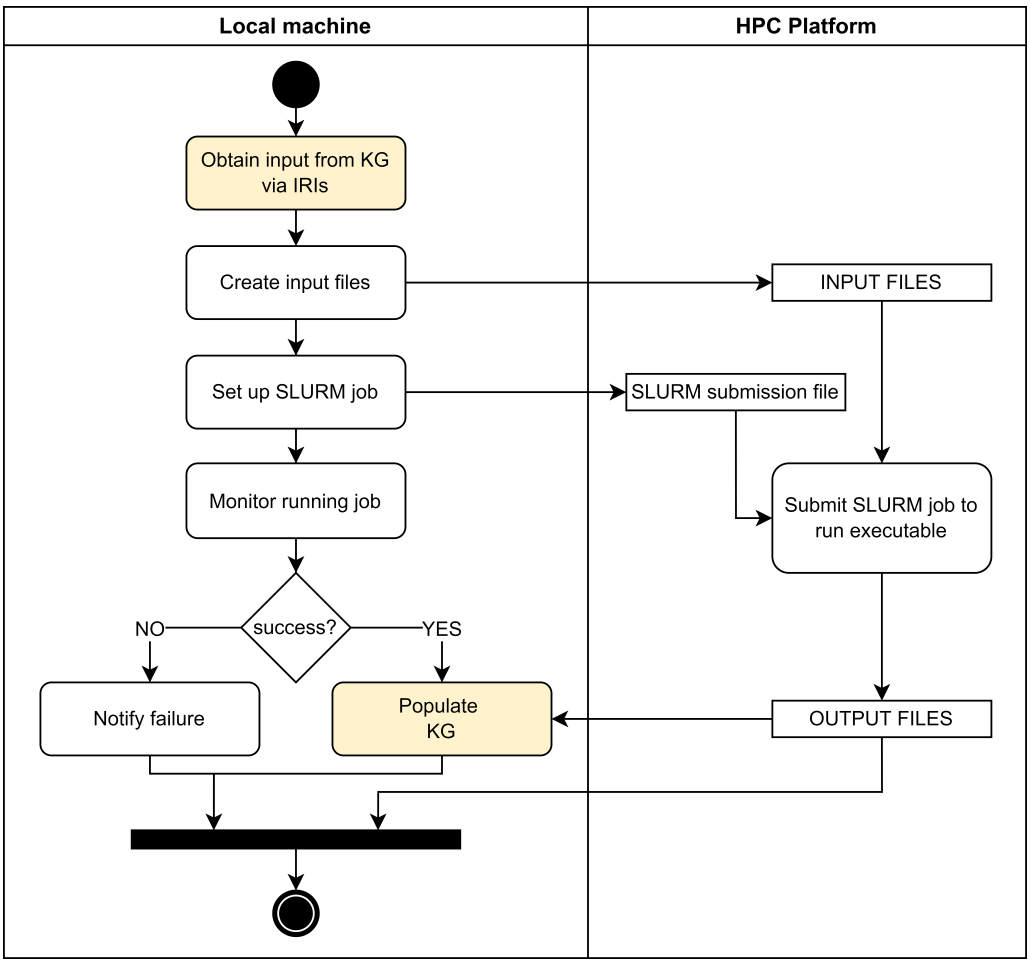


Figure 5: *UML activity diagram of Force Field Fitting agent. The agent enables the calibration job to be executed on a HPC platform. The yellow shaded actions indicates the data retrieving operation of agent over the knowledge-graph and the knowledge-graph populating operation.*

The EVB executable workflow covers different tasks that are performed with different software and it is detailed as follows:

1. The *XYZ* coordinates and the SCF energy of each scan point are given as input to the executable. A configuration file that contains information on the classic force field scheme to be used is also given as input.

2. Reactants and products state are identified from the PES curve. Local minima energies in the PES and energy at the scan coordinate maximum distance (in case of bond scans with one local minima in the PES) are selected by the executable as reactant and product chemical state.

3. Classic force field are assigned to each state using DL_FIELD [69]. DL_FIELD tool converts user's atomic configuration in simple *XYZ* coordinates, into identifiable atom types base on a particular user-selectable force field scheme looking at

the neighbouring atoms for each atom in the system. All the force field information for each chemical state are stored in a topology file (DL_POLY format). In the case DL_FIELD is unable to assign the topology to one of the state, the agent notify the job failure. In such circumstance, an user-defined topology for that state can be given as an input to the executable.

4. The two classic force fields energies for each geometry along the scan coordinate are calculated using DL_POLY [8, 65].

5. The EVB coupling term is calibrated using MoDS [10]. MoDS is an integration of multiple tools developed for various generic model development tasks, such as parameter estimation [2, 40], surrogate model creation [70], and experimental design [54]. The calibration procedure is described in the Appendix A.2.

We selected two use-cases as preliminary examples: the C-C bond scan in the ethanol molecule (Fig. 6a), and 1,2-dihydroacenaphthylen-1-yl dimer formation (Fig. 6b). The first case is selected to show that any bond breakage and formation can be accurately described using this methodology. The second one is a more interesting case that finds new applications in the combustion field. 1,2-dihydroacenaphthylen-1-yl is a localized π-radical. Recently, quantum mechanics/molecular mechanics (QM/MM) simulations showed that π-radicals bond strongly enough for stability at flame temperature and react rapidly through physically stabilized internal rotors towards soot nanoparticles [51]. However, current reactive force field parameterizations are unable to detect the bond formation between two localized π-radical sites so only ab initio or QM/MM approaches can be used to study such a system at the moment.

Fig. 6 shows the energies profiles for the two selected cases. Relaxed potential energy surface scans were performed using Gaussian 16 [23] along the C-C bond in the ethanol case and along the *pi*-two radical sites in the 1,2-dihydroacenaphthylen-1-yl dimer case. In both cases, geometries were optimized at B3LYP/cc-pVQZ level of theory and the energies were refined with single point energy calculations at M06-2X/cc-pVQZ level of theory. All DFT calculations were carried out using the spin-unrestricted formalism. The obtained DFT energy profiles (black lines in Fig. 6) were used as quantum-chemical reference energies for the force filed calibration. Scan points corresponding to local minima energy and the maximum distance are selected as state 1 and state 2 respectively by the executable. Force fields were generated with the DL_FIELD program using the OPLS2005 [5, 38] force field library for the two different chemical states. In the ethanol case, DL_FIELD fails to create the topology for state 1, because it is unable to assign any atom type for carbon atoms that contain three coplanar bonds with non-carbon atoms. So in this case, a user-defined topology is given as input for the $\dot{C}H_3$ and $\dot{C}H_2OH$ radicals. In the 1,2-dihydroacenaphthylen-1-yl case, where the OPLS2005 Lennard-Jones term is known to overestimate the dispersion energies, the isoPAHAP force field is used to describe the intermolecular interactions [66]. At each scan point, energies are calculated with the two selected force fields using DL_POLY. Finally, the EVB coupling term is calibrated to fit the DFT energy profiles using MoDS. The EVB potential energies exhibit good agreement with the DFT reference data in both cases, with a maximum residual of 11 kJ/mol for the ethanol case and 14 kJ/mol for the 1,2-dihydroacenaphthylen-1-yl dimer case.

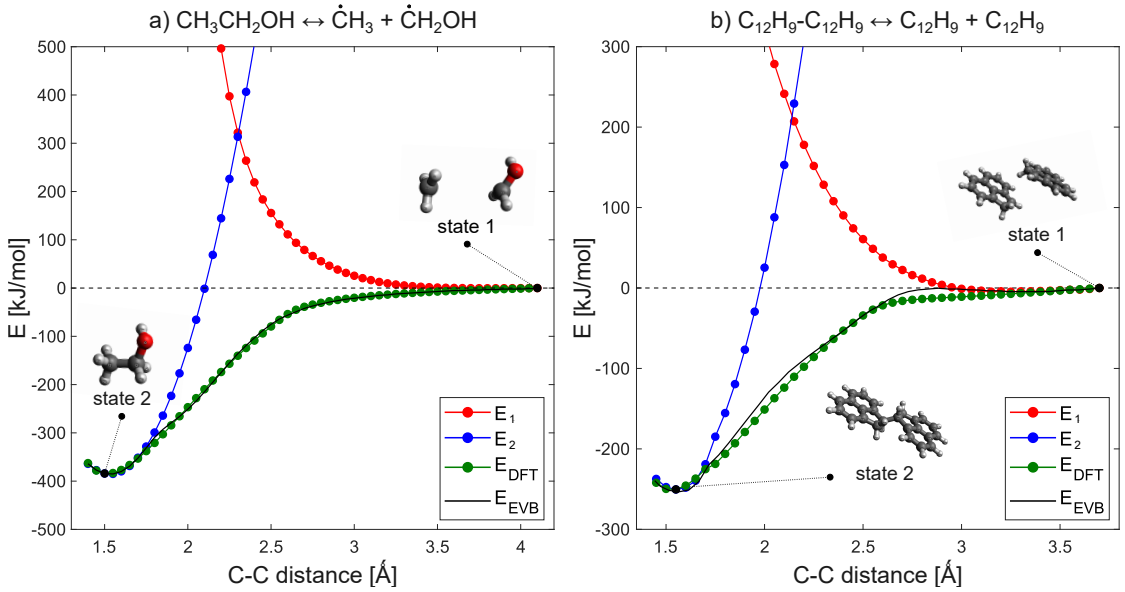


Figure 6: *Energy profiles along the C-C bond distance in ethanol (a) and along the two π-radical sites in the 1,2-dihydroacenaphthylen-1-yl dimer formation (b). Green lines show the DFT energies used as reference. The zero of energy was chosen for clarity as the C-C maximum distance (state 1). Red and blue curves show the energies calculated with the classic force field for state 1 and 2 respectively. Black curves show the obtained EVB potential energies.*

This methodology can be used to fit force fields able to describe any system in which reactions of interest can be previously identified and the obtained force field can be easily extended by adding new reactions and chemical states. The fitting parameters can then be used for a given MD application to describe the dynamics of chemical reactions as a valid alternative to more complex methods such as reactive force field and QM/MM methods. MD simulations are not reported here as they are outside of the scope of this work, but they will be part of future work.

5 Conclusions and Outlook

In this work, a new ontology for the representation of exploration of potential energy surfaces, OntoPESScan has been developed. This ontology adds further support for the representation of computational chemistry concepts in The World Avatar. The OntoPESScan ontology makes use of linked data principles by containing relations that link to existing chemistry ontologies in OntoSpecies and OntoCompChem. This enables potential energy surfaces to be queried by species information in OntoSpecies, and points along a potential energy surface to have their energies and geometries stored and described using concepts in OntoCompChem in a semantic way. Additionally, a force field fitting agent has been developed to make use of the linked data in the OntoPESScan ontology and showcase the advantages of the knowledge graph. This agent shows how force fields for reactive systems can be parametrised on-the-fly by applying the empirical valence bond

coupling method to potential energy surfaces by utilizing the description of the surface in OntoPESScan in conjunction with the linked computational chemistry data in OntoCompChem. The agent was demonstrated for two potential energy surfaces. The first is a well-known surface corresponding to carbon-carbon bond scission in ethanol. The second corresponds to a newer case of reactions between localised π-radical polyaromatic hydrocarbons. In both cases, a force field was fitted for the reactions potential energy surfaces, which can then be used in performing molecular dynamics simulations.

Going forward, further development of OntoPESScan and other ontologies for chemistry are planned to extend the capabilities of The World Avatar. A potential extension to this would be the development of an ontology and semantic representation of molecular dynamics simulations and results, as this is a key area of computational chemistry. Such an ontology for molecular dynamics would naturally link well with OntoPESScan, as the information on how force field fitting parameters for a given molecular dynamics simulations are derived would be found in OntoPESScan and OntoCompChem. Additionally, agents could use the information in OntoPESScan and OntoCompChem to fit new force fields on the fly for different molecular dynamics simulations, which could then be represented in the knowledge graph as well. As with molecular dynamics, potential energy surfaces are also key for computation of rate coefficients during mechanism development. Development and coupling of OntoKin or other ontologies for rate coefficients with OntoPESScan would also work towards agents being able to continuously calculate and improve rate coefficient estimates for chemical kinetic mechanism development. This will all help work towards an open, continuous, self-growing knowledge graph for chemistry.

Research data

Research data supporting this publication is available in the University of Cambridge data repository (doi:10.17863/CAM.82487).

Acknowledgements

This project is funded by the National Research Foundation (NRF), 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 & The Alan Turing Institute. M.K. 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.

A Appendix

A.1 Empirical Valence Bond Approach

The EVB approach describes chemical reactivity by taking into account the diabatic states corresponding to the classical valence-bond structures describing the reactant and product states of a reaction, and any intermediate states, if they exist. A simple schema for a two-state reaction is shown in Fig. 7, where the energy of each diabatic state is represented by a non-reactive force field (FF), one non-reactive FF for the reactants (E_1) and one non-reactive FF for the product (E_2).

The EVB method defines a pseudo-Hamiltonian matrix ($H_{\mathrm{EVB}}(R)$) whose matrix representation (Eq. A.1) has the potential energies of the reactant and product diabatic states at a given structure obtained from standard non-reactive force fields as diagonal components ($E_1(R)$ and $E_2(R)$), whereas the off-diagonal terms are given by the coupling C_{12} between the force fields in the reaction:

$$H_{\mathrm{EVB}} = \begin{pmatrix} E_1(R)+\varepsilon_1 & C_{12}(R) \\ C_{12}(R) & E_2(R)+\varepsilon_2 \end{pmatrix} \tag{A.1}$$

where all terms depend on the set of atomic coordinates R. The ε_1 and ε_2 values in Eq. A.1 are constant diagonal energy shifts, usually chosen so as to reproduce the known exo or endo-thermicity of the reaction in question.

Following diagonalization of $H_{\mathrm{EVB}}(R)$, we obtain two possible eigenvalues, $\lambda^{\pm}$,

$$\lambda^{\pm} = \frac{1}{2}\left[E_1(R)+E_2(R) \pm \sqrt{(E_1(R)+E_2(R))^2 - 4\left[E_1(R)E_2(R) - C_{12}^2(R)\right]}\right]. \tag{A.2}$$

The EVB energy (E_{EVB}) is defined as the lowest eigenvalue:

$$E_{\mathrm{EVB}} \equiv min(\lambda^+, \lambda^-) \tag{A.3}$$

The off-diagonal coupling elements C_{12} is typically a Gaussian function of the set of atomic coordinates R. For the implementation of the EVB method in DL_POLY, C_{12} has the following functional form:

$$C_{12} = A_1 \exp - \left(\frac{E_1(R) - E_2(R) - A_2}{A_3}\right)^2 + A_4 \tag{A.4}$$

To obtain an EVB potential, the diagonal force fields are first compared to ab initio results (or experiment) ((E_{ref}) in Fig. 7) for the corresponding stable species, and if necessary, small adjustments are made in order to reproduce the structures. Then, the behavior of the diagonal force fields (shifted appropriately to take reaction energy into account) along the reaction coordinate is assessed, and compared to ab initio data.

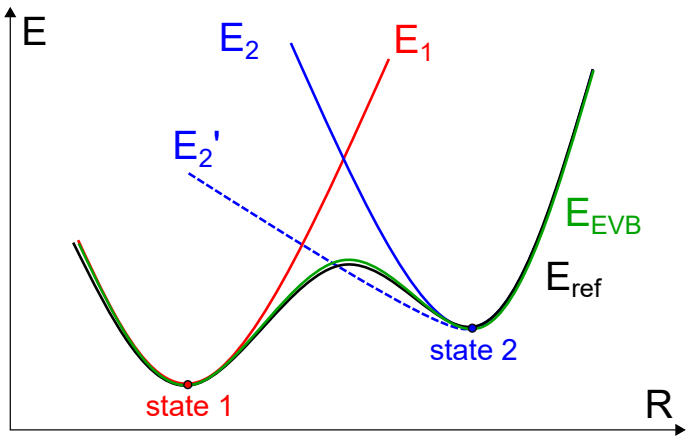


Figure 7: *Schematic representation of an EVB reactive potential energy surface, E_{EVB} (solid green line), obtained from two diabatic states E_1 (solid red line) and E_2 (solid blue line) corresponding to reactant and product states. Reference energy landscape is shown in black. An example of an inappropriate product diabatic state function, E_2' (dashed blue line), that lies lower than the target potential energy for some values of R, is also shown.*

As shown in Fig. 7, if we used the force field built for state 1 (E_1), state 2 will never be sampled, since the values of E_1 in the region of state 2 will be exceedingly large. Analogously, state 1 will never be sampled if we used E_2 as the force field to describe the interactions. The effect of diagonalization of the EVB pseudo-Hamiltonian is and can only be to lower the energy relative to the lowest of the diagonal state energies. Hence if one or more of the diagonal states is significantly lower in relative energy than the ab initio data (E_{ref}) in one section of the reaction coordinate, as for the case of E_2' shown in the schematic Fig. 7, then it will be impossible to get a good fit of the potential energy surface.

A.2 Calibration of EVB coupling term

Once it has been established that the diagonal force fields E_1 and E_2 produce an acceptable description, the off-diagonal terms (C_{12}) can be fitted to the target potential energy surface E_{ref}. The set of parameters in Eq. A.4

$$\theta = (A_1, A_2, A_3, A_4)$$

is estimated to minimize the least-squares objective function given by:

$$\Phi(\theta) = \sum_{n=1}^{N} [E_{\text{EVB}}(R^{(n)}, \theta) - E_{\text{ref}}(R^{(n)})]^2 \tag{A.5}$$

where E_{EVB} is given by Eq. A.3, E_{ref} is the the reference energy landscape (usually ab initio data), N is the number of scan points and R is the set of atomic coordinates. The calibration process initially employs low-discrepancy quasi-random global sampling through

a Sobol sequence generator [63]. This provide intial points for a SolvOpt optimization algorithm [61], selected for the non-linearity of the least-squares objective function.

References

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

[2] 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(4):1701–1717, 2021.

[3] 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*, 2022.

[4] P. Ballone. Modeling potential energy surfaces: From first-principle approaches to empirical force fields. *Entropy*, 16(1):322–349, 2014.

[5] J. L. Banks, H. S. Beard, Y. Cao, A. E. Cho, W. Damm, R. Farid, A. K. Felts, T. A. Halgren, D. T. Mainz, J. R. Maple, R. Murphy, D. M. Philipp, M. P. Repasky, L. Y. Zhang, B. J. Berne, R. A. Friesner, E. Gallicchio, and L. R. M. Integrated modeling program, applied chemical theory (IMPACT). *Journal of Computational Chemistry*, 26:1752–1780, 2005.

[6] J. R. Barker. Multiple-well, multiple-path unimolecular reaction systems. i. multiwell computer program suite. *International Journal of Chemical Kinetics*, 33(4): 232–245, 2001.

[7] M. Batty. Digital twins. *Environment and Planning B: Urban Analytics and City Science*, 45(5):817–820, 2018.

[8] I. Bush, I. Todorov, and W. Smith. A DAFT DL_POLY distributed memory adaptation of the smoothed particle mesh ewald method. *Computer Physics Communications*, 175:323–329, 2006.

[9] A. Chadzynski, N. Krdzavac, F. Farazi, M. Q. Lim, S. Li, A. Grisiute, P. Herthogs, A. von Richthofen, S. Cairns, and M. Kraft. Semantic 3D city database – an enabler for a dynamic geospatial knowledge graph. *Energy and AI*, 6:100106, 2021.

[10] CMCL Innovations. MoDS: Model development suite (version 2020.2), 2020. URL https://cmclinnovations.com/solutions/products/mods/.

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

[12] A. Devanand, G. Karmakar, N. Krdzavac, R. Rigo-Mariani, Y. 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.

[13] A. Devanand, G. Karmakar, N. Krdzavac, F. Farazi, M. Q. Lim, Y. 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, 2022.

[14] F. Duarte and S. C. L. Kamerlin. *Theory and Applications of the Empirical Valence Bond Approach: From Physical Chemistry to Chemical Biology*. John Wiley & Sons, 2017.

[15] 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, 2019. doi:10.1016/j.compchemeng.2019.106586.

[16] A. Eibeck, D. Nurkowski, A. Menon, J. Bai, J. Wu, L. Zhou, S. Mosbach, J. Akroyd, and M. Kraft. Predicting power conversion efficiency of organic photovoltaics: models and data analysis. *ACS omega*, 6(37):23764–23775, 2021.

[17] A. El Saddik. Digital twins: The convergence of multimedia technologies. *IEEE multimedia*, 25(2):87–92, 2018.

[18] 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(1):108–120, 2019.

[19] F. Farazi, N. B. Krdzavac, J. Akroyd, S. Mosbach, A. Menon, D. Nurkowski, and M. Kraft. Linking reaction mechanisms and quantum chemistry: An ontological approach. *Computers & Chemical Engineering*, 137:106813, 2020.

[20] M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, and B. A. Grzybowski. Architecture and evolution of organic chemistry. *Angewandte Chemie International Edition*, 44(44):7263–7269, 2005. doi:10.1002/anie.200502272.

[21] M. Frenklach. Transforming data into knowledge—process informatics for combustion chemistry. *Proceedings of the Combustion Institute*, 31(1):125 – 140, 2007. doi:https://doi.org/10.1016/j.proci.2006.08.121.

[22] M. Frisch. Gaussian 03 rev. e. 01. *http://www.gaussian.com/*, 2004.

[23] M. Frisch, G. Trucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, G. Petersson, H. Nakatsuji, et al. Gaussian 16, 2016.

[24] M. J. Frisch. Gaussian09. *http://www.gaussian.com/*, 2009.

[25] 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):1–15, 2015.

[26] 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.

[27] Y. Georgievskii and S. Klippenstein. Varecof. *Sandia National Laboratories and Argonne National Laboratory*, 2006.

[28] M. Ghahremanpour, P. J. Van Maaren, and D. Van Der Spoel. The Alexandria library, a quantum-chemical database of molecular properties for force field development. *Scientific data*, 5:180062, 2018.

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

[30] 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):1–13, 10 2011. doi:10.1371/journal.pone.0025513.

[31] J. Hastings, P. De Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, 41(D1):D456–D463, 2012.

[32] 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(D1):D1214–D1219, 2016.

[33] 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.

[34] R. Hodgson. The NASA QUDT handbook, 2014. URL `http://ontolog.cim3.net/file/work/OntologyBasedStandards/2013-10-10_Case-for-QUOMOS/NASA-QUDT-Handbook-v10--RalphHodgson_20131010.pdf`.

[35] E. Hong, G.and Rosta and A. Warshel. Using the constrained DFT approach in generating diabatic surfaces and off diagonal empirical valence bond terms for modeling reactions in condensed phases. *The Journal of Physical Chemistry B*, 110: 19570–19574, 2006.

[36] P. Jacob and A. Lapkin. Statistics of the network of organic chemistry. *React. Chem. Eng.*, 3:102–118, 2018. doi:10.1039/C7RE00129K.

[37] R. Johnson III. CCCBDB computational chemistry comparison and benchmark database. *NIST Standard Reference Database Number*, 101, 1999.

[38] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118:11225–11236, 1996.

[39] S. C. Kamerlin and A. Warshel. The empirical valence bond model: theory and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1 (1):30–45, 2011.

[40] C. A. Kastner, A. Braumann, P. L. Man, S. Mosbach, G. P. Brownbridge, J. Akroyd, M. Kraft, and C. Himawan. Bayesian parameter estimation for a jet-milling model using metropolis-hastings and wang-landau sampling. *Chemical Engineering Science*, 89:244–257, 2013.

[41] R. J. Kee, J. A. Miller, and T. H. Jefferson. CHEMKIN: A general-purpose, problem-independent, transportable, FORTRAN chemical kinetics code package. Technical report, Sandia Labs., 1980.

[42] S. Kim, P. A. Thiessen, T. Cheng, B. Yu, B. A. Shoemaker, J. Wang, E. E. Bolton, Y. Wang, and S. H. Bryant. Literature information in PubChem: associations between PubChem records and scientific articles. *Journal of Cheminformatics*, 8(1): 32, 2016.

[43] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109, 2019.

[44] M. J. Kleinelanghorst, L. Zhou, J. J. Sikorski, Y. S. E. Foo, K. Aditya, S. Mosbach, I. A. Karimi, R. Lau, and M. Kraft. J-Park Simulator: roadmap to smart eco-industrial parks. In *ICC*, pages 107–1, 2017.

[45] N. Krdzavac, S. Mosbach, D. Nurkowski, P. Buerger, J. Akroyd, J. Martin, A. Menon, and M. Kraft. An ontology and semantic web service for quantum chemistry calculations. *Journal of Chemical Information and Modeling*, 59(7):3154–3165, 2019.

[46] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann. Industry 4.0. *Business & information systems engineering*, 6(4):239–242, 2014.

[47] A. J. Lawson. The making of Reaxys – towards unobstructed access to relevant chemistry information. *The Future of the History of Chemical Information*, 1164: 127–48, 2014.

[48] E. Lewars. Computational chemistry. *Introduction to the theory and applications of molecular and quantum mechanics*, page 318, 2011.

[49] M. F. Lopez, A. Gomez-Perez, J. P. Sierra, and A. P. Sierra. Building a chemical ontology using Methontology and the ontology design environment. *IEEE Intelligent Systems and their Applications*, 14(1):37–46, Jan 1999. doi:10.1109/5254.747904.

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

[51] J. W. Martin, L. Pascazio, A. Menon, J. Akroyd, K. Kaiser, F. Schulz, M. Commodo, A. D'Anna, L. Gross, and M. Kraft. π–diradical aromatic soot precursors in flames. *J. Am. Chem. Soc.*, 143(31):12212–12219, 2021.

[52] A. Menon, N. B. Krdzavac, and M. Kraft. From database to knowledge graph—using data in chemistry. *Current Opinion in Chemical Engineering*, 26: 33–37, 2019.

[53] J. A. Miller, R. Sivaramakrishnan, Y. Tao, C. F. Goldsmith, M. P. Burke, A. W. Jasper, N. Hansen, N. J. Labbe, P. Glarborg, and J. Zádor. Combustion chemistry in the twenty-first century: Developing theory-informed chemical kinetics models. *Progress in Energy and Combustion Science*, 83:100886, 2021.

[54] S. Mosbach, A. Braumann, P. L. Man, C. A. Kastner, G. P. Brownbridge, and M. Kraft. Iterative improvement of bayesian parameter estimates for an engine model by means of experimental design. *Combustion and Flame*, 159(3):1303–1313, 2012.

[55] 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(12):6155–6166, 2020.

[56] 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 (1):1–14, 2011.

[57] N. S. Ostlund and M. Sopek. GNVC: Gainesville core ontology - standard for publishing results of computational chemistry, ver. 0.7, 2015. URL http://ontologies.makolab.com/gc/gc07.owl. Accessed October 24th, 2018.

[58] J. Pérez, M. Arenas, and C. Gutierrez. Semantics and complexity of SPARQL. *ACM Transactions on Database Systems (TODS)*, 34(3):1–45, 2009.

[59] G. Schmitz, I. H. Godtliebsen, and O. Christiansen. Machine learning for potential energy surfaces: An extensive database and assessment of methods. *The Journal of Chemical Physics*, 150(24):244113, 2019.

[60] T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, et al. The ReaxFF reactive force-field: development, applications and future directions. *npj Computational Materials*, 2(1):1–14, 2016.

[61] N. Z. Shor. *Gradient-type Methods with Space Dilation, Minimization Methods for Non-Differentiable Functions*. Springer-Verlag, New York, 1985.

[62] G. Smith and D. Golden. Application of RRKM theory to the reactions $OH+NO_2+N_2 \rightarrow HONO_2+N_2$ (1) and $ClO+NO_2+N_2 \rightarrow ClONO_2+N_2$ (2); a modified Gorin model transition state. *International Journal of Chemical Kinetics*, 10(5): 489–501, 1978.

[63] I. Sobol. On the systematic search in a hypercube. *SIAM Journal on Numerical Analysis*, 16(5):790–793, 1979.

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

[65] I. T. Todorov, W. Smith, K. Trachenko, and M. T. Dove. DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism. *J. Mater. Chem.*, 16:1911–1918, 2006.

[66] T. S. Totton, A. J. Misquitta, and M. Kraft. A quantitative study of the clustering of polycyclic aromatic hydrocarbons at high temperatures. *Phys. Chem. Chem. Phys.*, 14:4081–4094, 2012.

[67] 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.

[68] D. Weininger, A. Weininger, and J. L. Weininger. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, 1989.

[69] C. W. Yong. Descriptions and implementations of DL_F notation: A natural chemical expression system of atom types for molecular simulations. *Journal of Chemical Information and Modeling*, 56:1405–1409, 2016.

[70] C. Yu, M. Seslija, G. P. Brownbridge, S. Mosbach, M. Kraft, M. Parsi, M. Davis, V. Page, and A. Bhave. Deep kernel learning approach to engine emissions modeling. *Data-Centric Engineering*, 1:e4, 2020.

[71] 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, 2017.

[72] 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.

[73] 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.

[74] X. Zhou, A. Eibeck, M. Q. Lim, N. B. Krdzavac, and M. Kraft. An agent composition framework for the J-Park Simulator – a knowledge graph for the process industry. *Computers & Chemical Engineering*, 130:106577, 2019.

[75] X. Zhou, M. Q. Lim, and M. Kraft. A smart contract-based agent marketplace for the J-Park Simulator – a knowledge graph for the process industry. *Computers & Chemical Engineering*, 139:106896, 2020.