JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Cite This: J. Chem. Inf. Model. 2018, 58, 1742−1754

Article

pubs.acs.org/jcim

Redesigning the Materials and Catalysts Database Construction Process Using Ontologies

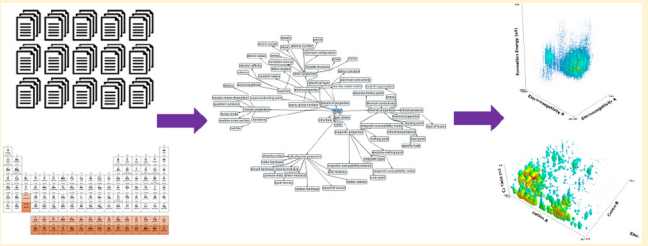
Lauren Takahashi,[†] Itsuki Miyazato,[†,‡] and Keisuke Takahashi*[,†,§]

[†]Center for Materials Research by Information Integration (CMI[2]), National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

[‡]Graduate School of Engineering, Hokkaido University, N-13, W-8, Sapporo 060-8628, Japan

[§]Institute for Catalysis, Hokkaido University, N-21, W-10, Kita-ku, Sapporo 001-0021, Japan

ABSTRACT: Materials and catalyst informatics are emerging fields that are a result of shifts in terms of how materials and catalysts are discovered in the fields of materials science and catalysis. However, these fields are not reaching their full potential due to issues related to database creation and curation. Issues such as lack of uniformity, data selectivity, and the presence of bias affect the quality and usefuless of materials databases, especially when attempting to search for materials descriptors. Without uniform rules and frameworks, databases are limited in use outside of the intent of the creators of the database. Ontologies are therefore investigated as a means of redesigning the way materials and catalysts databases are designed and created. In particular, an ontology consisting of information found within the periodic table as well as commonly used related data is constructed and applied toward the search for materials descriptors. Additional ontologies are also developed for two databases—a database consisting of computational data related to perovskites and a database consisting of experimental data related to oxidative coupling of methane (OCM) catalysts—in order to investigate the impact of merging ontologies.



INTRODUCTION

The fields of materials science and catalysis have begun to undergo a paradigm shift in terms of how data is utilized for analysis and information discovery, leading toward the establishment of materials informatics and catalyst informatics where data science is used as a primary method of investigation.[1−5] Materials informatics and catalyst informatics apply machine learning to data in order to uncover hidden trends and periodicities within data. In particular, materials informatics and catalyst informatics are powerful when applied to the material design process as there is a vast number of databases and other forms of data available for use when attempting to design a new material or catalyst using data science. However, given the number of subfields that fall under the field "materials science", it becomes difficult for researchers to cross-access available databases and collections of data due to barriers related to differences in subject terminology, methods of data collection and organization, and methods of communicating thought processes and decisions behind the data collection process More specifically, descriptors for determining materials properties and catalytic activities are generally sought for when a data set is linked to materials and catalysts properties.[6−9] However, databases are often not uniform in design and creation, making it very difficult to take advantage of the tools that materials informatics provides. It becomes imperative, therefore, to reconsider how data is collected given its central role within materials and catalyst informatics.[1,3,10]

Data collection within the materials science and catalysis fields has been active over the years where data acquisition is carried out through various approaches such as literature collection, high-throughput experiments, and calculations.[1,10−15] However, concern centered on data organization arises for applications in data science where issues such as uniformity of data, the degree of data included within a particular database, and the presence of biases can affect the quality and usefulness of data when applied to data science. Data sets are often biased toward the needs of a particular research group and are impacted by the methods of creating and curating data chosen by each individual of the research group, leading to friction when attempting to collaborate or share data with other groups.[16] Furthermore, failed data and negative data are often not reported within materials science; this has an adverse effect when data science is applied, as the inclusion of failed or negative data greatly contributes toward the discovery of novel materials.[8,17] These issues thus show that the information included within a database is strongly influenced by the specialty, interests, and preferences of the researchers that build it and is therefore limited in terms of usability.

Without providing a framework that would allow accessibility to information beyond the officially recorded data, progress toward transforming data to knowledge will be limited and time consuming. Attempts have been made to utilize metadata to help

Received: March 20, 2018
Published: August 2, 2018

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

clarify background information pertaining to the officially recorded data, yet issues such as informality of the metadata creation process and differing levels of importance placed on communication issues play a role in hindering the establishment of materials informatics.[16] Without a system to allow for defining meaning of data, applications of data science will face increasing difficulty when attempting to transform data into knowledge. Here, ontology has the potential to be a useful alternative to current methods of assigning meaning to data. By designing and implementing ontologies, databases can become more uniform and also include additional data such as relationships between data, metadata, and other pieces of information that may prove to be necessary to a particular data set and therefore become more globally accessible.

SETBACKS RELATED TO DATABASE CREATION AND CURATION

Technological advancements made over the past several decades have led to an unprecedented level of data creation and curation for all scientific disciplines. Access to such levels of data has led to the birth of new research disciplines that incorporate data science such as the establishment of bioinformatics within the field of biology.[18,19] However, the multidisciplinary nature of the materials science and catalysis fields has held these fields back from enjoying similar success. Barriers such as individualization of data, differences in research disciplines, and so-called science "friction" impact the accessibility and usefulness of a database and therefore should be examined.[16]

Barriers to accessibility arise when researchers attempt to use a database that was developed or maintained by researchers from a different discipline.[8,17] Databases are often individualized to fit the needs of a particular research group or project within a particular research field. Research disciplines often operate under terminology that changes per field as well as have diverse research goals, place various levels of importance on different types of information, and employ different methods and tools to generate and record data. In particular, some research fields may operate under the same "parent" field and use the same words and values, but the meaning of these words and values change according to the field in question. For example, great care must be made when discussing atomic clusters. While the term "atomic cluster" is generally described as "an aggregate of atoms", the distinction between atomic clusters and nanoparticles is ambiguous.[20] Atomic clusters vary in size and composition, and although they eventually can grow into a nanoparticle, there is no single defined line that distinguishes an atomic cluster from a nanoparticle. This leads to many different interpretations of the same piece of information and results in confusion from the community. Methods of collecting data also differ between disciplines, which lead to separate needs from the database. For example, experimentalists often use a variety of lab equipment and techniques in order to create samples, capture images, and take measurements for recording purposes. Computational scientists, on the other hand, rely on different programming skills, software, and hardware in order to create models and calculate values for a database. Both cases offer a substantial amount of multimedial information that may become lost when one attempts to translate that data into a database-friendly format. In addition, only the desired data is placed into a database, and information that is not needed is withheld; the withheld data, meanwhile, may prove to be very useful to other researchers. These differences in data have the potential to lead to misinterpretations and possible errors when attempting to translate the original data that is easier to understand for a researcher that specializes in another discipline.

Science "friction" is another large part of why databases are becoming limited in usefulness.[16] Databases are often developed by multiple contributors, who each have different methods of collecting data and place different levels of importance on data which can affect how thoroughly data is recorded. Collecting data published within literature is also affected due to differences in how various research groups conduct research and report findings.[11] Attempts to communicate may be tedious or time consuming, as some researchers can no longer be reached, and methods such as electronic communication can be unreliable. In addition, the true total amount of data produced is not often reflected in the officially reported database; it is not uncommon for potentially useful information and observations to be hidden within notebooks or casual discussions with colleagues. Factors such as negative results, publishing restrictions, and confidentiality restrictions put in place by investors or research grants can also impact the availability of data.

Database accessibility is thus currently limited due to ambiguously defined data, the exclusion of negative data and other data restricted by confidentiality agreements or methods of data recording, and unreliable methods of communicating with other researchers. Ontology, however, offers a possible solution to these issues. Traditionally a philosophical concept, ontology has been adopted by computer science and information science as a new way of defining meaning and relationships within data.[21,22] It has been successfully adopted within the field of bioinformatics where biological roles are globally defined.[23] Such concepts should also naturally fit when developing databases within materials science and catalysis. Ontology has the potential to make a database more uniform and allow for the semantics of data to be defined in a more meaningful manner, thereby making the data more globally accessible. Therefore, this manuscript explores the impact an ontology can have on developed databases.

APPLICATION OF ONTOLOGY IN MATERIALS AND CATALYSTS DATA

Ontology has the potential to help streamline how data is organized within materials and catalysts databases. Ontologies are composed of three parts: a set of vocabulary representing various concepts, definitions for the vocabulary set, and defined relationships between the concepts. These parts establish the existence of various concepts within a system as well as define how various parts of the system relate to each other. An ontology for a materials science or catalysis database could be categorized and applied toward the following areas: the database itself, materials and catalysts, and functionality. For the database itself, an ontology can be developed in order to define the structure and framework of the database. Additionally, an ontology can be developed in order to organize definitions and relationships between various materials and catalysts. Lastly, ontology can be used for functionality by defining the meaning and relationships between different properties of materials and catalysts.

One particular case where this can impact research is the case of searching for material descriptors. Material descriptors are understood to be the variables responsible for a particular material property.[7] When searching for descriptors for a material property, it is common to access and use available databases. However, when a database is implemented, elements and other related data are often not included. In addition, the periodic table is often ignored or is selectively used when constructing

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

databases. This has a negative impact when searching for descriptors, as many potentially vital factors are not being considered when attempting to find descriptors from a database, thereby restricting the descriptor search. This leads to an increase in time spent reconstructing databases and searching for additional data that may or may not help the search for particular descriptors.

An ontology can help fill gaps in a database by providing data that may be necessary but is often overlooked. For example, it has been demonstrated that descriptors for determining the structure, magnetic moment, lattice constant, and formation energy of materials contains information from the periodic table such as electronegativity and atomic density.[7,24−26] Traditionally, basic information about the periodic elements can be found within the periodic table. However, there is information related to the periodic table that is not explicitly shown such as trends in electronegativity and ionization energy. An ontology can help define these relationships by providing the space to not only define these properties but also to provide space to define how these properties relate to each other. In addition to the periodic table, ontologies can be used for materials and catalysts databases since the relationships and meaning between the data found within these databases are often not clear or easy to understand for researchers who were not involved in creating such databases. If these databases can be defined semantically, then it becomes possible, in principle, to integrate the databases. This manuscript thus attempts to demonstrate the use for ontologies for the periodic table and related properties as well as develop and integrate ontologies for two published databases: a computational database consisting of data related to perovskites and an experimental database consisting of data related to oxidative coupling of methane (OCM) reactions.

METHODOLOGY AND ONTOLOGY CREATION

In this manuscript, three ontologies are created: an ontology representing the periodic table and related properties, an ontology for a computational database for perovskite data, and an ontology for an experimental database for OCM reaction data. An ontology for the periodic table and common related properties is first designed. The ontology is developed consisting of the periodic elements, data found within the periodic table, and additional related data gathered from various sources. Having access to the periodic table in this manner is useful when integrating databases. Such a framework can potentially assist the data mining process, which is often used to reveal hidden information behind the elements within a database. The primary goal in constructing such an ontology is to provide a core framework that allows for proper navigation and inclusion of properties related to the periodic elements.

Figure 1 depicts the workflow taken in this manuscript. There are five main steps taken within this manuscript: ontology conception and planning; collection of data, related literature, and tools; ontology construction; ontology testing; and applications of the ontology. Before constructing the ontology, the purpose and goals of the ontology must be decided. In this particular case, the ontology needs to be developed for easy access to core elemental data in order to use in descriptor searches and other materials-related research. Therefore, the ontology must be designed where not only can data for the elements be easily accessed but classes and object properties must also be defined and allow for future edits. By doing this, it becomes possible to define property relationships that come to light when searching for descriptors. For the purposes of this

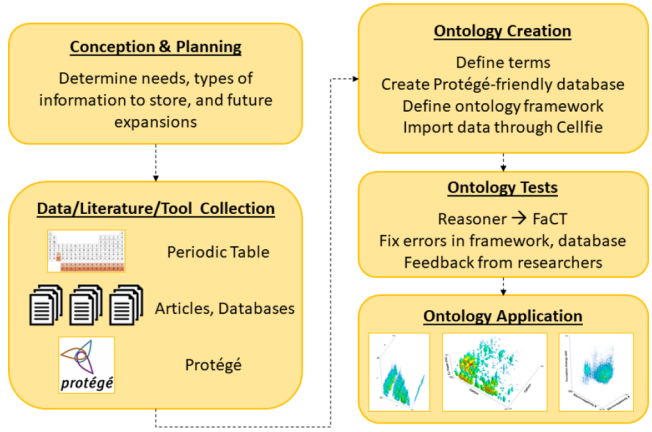


Figure 1. Workflow taken for developing the ontology.

manuscript, an ontology is designed for storing the elemental data of the periodic elements and for organizing the commonly accessed properties of these elements according to type. Classes, semantic relations, and definitions for individuals are created with these points in mind.

Once the goals of the ontology are determined, the necessary data and tools are gathered. The ontology consists of information found within the periodic table as well as commonly used properties and values for all elements. Wolfram Mathematica's "ElementData" function acts as the source for the data, which includes data published by NIST, the UK National Physical Laboratory, and large collections of published experimental data.[27−36] The "November 28 1016" version of the periodic table published by the International Union of Pure and Applied Chemistry is the version of the periodic table used in this manuscript.[37] Scientific data is collected into a single database and edited in order to import the data into the ontology itself. The ontology is written using the W3C Web Ontology Language (OWL) and uses the ontology editor Protege, which provides a graphical user interface for ontology management as well as tools for importing data, visualizing class relations, and accessing different types of ontology syntax.[38] The plug-in Cellfie and Manchester syntax are utilized in order to import the collected data into the ontology. The reasoner FaCT++ 1.6.5 is used in order to check for inconsistencies within the semantics of the ontologies.[39] Reasoners are particularly important due to their ability to check the ontology for logical inconsistencies as well show the results of rule inferences throughout the ontology. Once corrections to errors and inconsistencies listed by the reasoner are made, the ontology is then tested by researchers who specialize in materials science to ensure that the ontology is usable and that the definitions and properties are written correctly. Feedback from the researchers is then applied to the ontology, and the testing process is repeated as necessary.

The ontology is constructed by defining the classes and subclasses of the ontology as well as define the relationships between these classes through the use of object and data properties. Classes can be viewed as categories with unique properties, while object properties are the rules for how classes relate to each other. Meanwhile, data properties can be seen as properties where raw data can be stored. Thus, these three types of properties allow the ontologies to organize and delegate relationships between data.

The ontology for the periodic table generates a total of 4496 axioms. Additionally, the ontology has a class count of 170, an

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

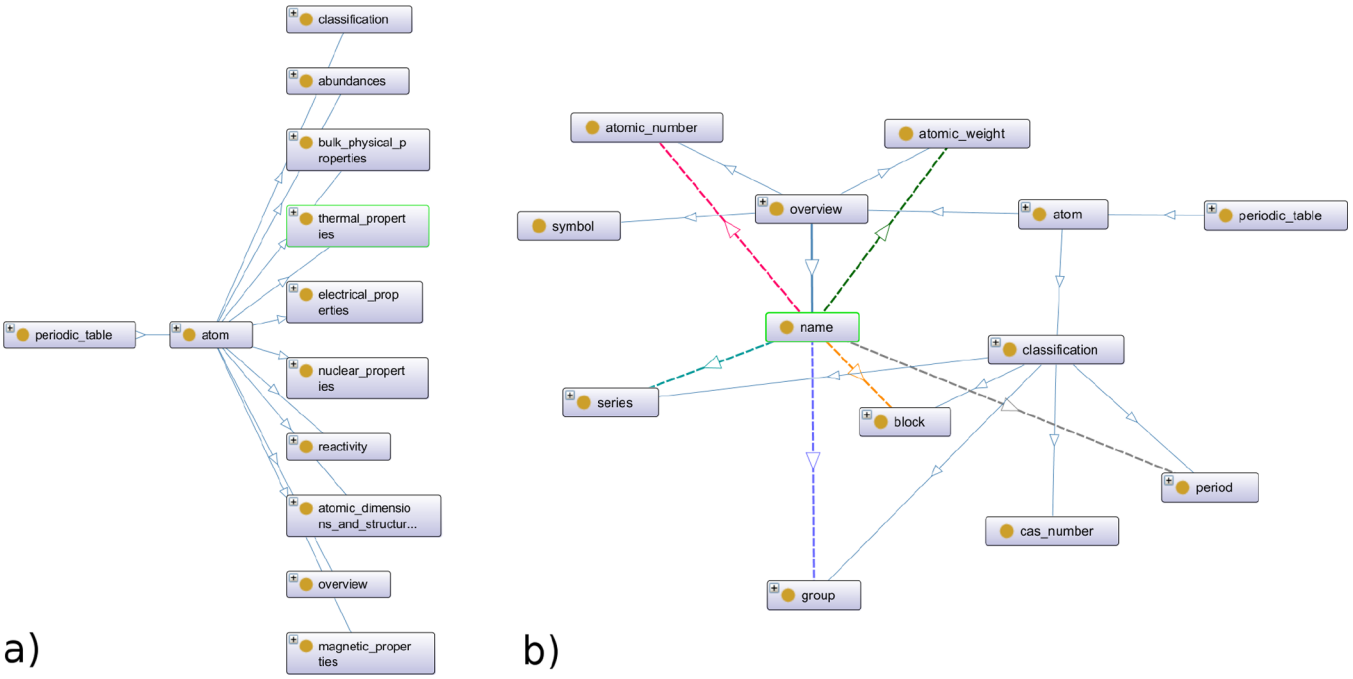


Figure 2. (a) Basic class tree hierarchy of the ontology developed for the periodic table. (b) Example of how the various subclasses relate to each other. Solid blue lines represent the hierarchy, while the colored dotted lines represent the object properties defining relations between classes.

object property count of 9, and a data property count of 66. There are also 118 individuals, which represent the individual atomic elements and are designed to contain the numerical data and other so-called raw data of the elements. The naming convention of the ontology was set to Snake Case as it is not only found to be easier to process by nonspecialists and leaves less room for error but it also is more accommodating for multilingual development that does not consider capitals or does not use a Latin alphabet.[40]

Figure 2 showcases the class hierarchy of the ontology "periodic_table". As can be seen in Figure 2a, major subclasses of the class "atom" include subclasses "overview", "classification", "bulk_physical_properties", and "atomic_dimensions_and_structure". These subclasses are further developed and act as the parent classes for classes that represent particular properties. For examples, the class "overview" has several subclasses such as "name", "symbol", "atomic_number", "density", "atomic_weight", "boiling_point", and "melting_point", while the class "atomic_dimensions_and_structure" has subclasses such as "bulk_modulus" and "young_modulus". These classes and subclasses are organized by property type, making it easier to navigate the data. Object properties also connect these subclasses and define how the classes relate to each other, which can be seen in Figure 2b. Additionally, the 118 atomic elements are defined as instances of the class "atom". These instances represent each element of the periodic table where the raw data of various commonly used properties are defined, as seen in Figure 3 with the example of "Hydrogen".

When reconsidering how to reorganize the periodic table, it becomes apparent that raw data for an element must be stored in a way that is easily accessible. Defining data properties and assigning them to individuals allows for this. In this manuscript, 66 data properties are defined. These data properties include data from the periodic table as well as commonly used data associated with the atomic elements. In addition, data properties allow for names to be given to raw data as well as define the type of data the information is stored at. These properties can then be stored within "individuals", which are unique instances of a set of data.

For example, the instance "Hydrogen" has 31 data property assertions, and as can be seen in Figure 3, various properties such as density, atomic weight, and electron configuration are defined with their appropriate values. These property assertions can be used to help define the element's relationship with other classes defined within the ontology. For example, the data property "property_element_series" is defined to equal "nonmetal" for "Hydrogen". This property relates to the class "series", which represents the element series that an atomic element belongs to. The class "series" has 11 subclasses that represent each possible series group an element can belong to. These subclasses have rules that define which instances belong to each subclass. For example, the subclass "nonmetal" is defined where the data property "elementseries" is equal to the value "nonmetal" or "Nonmetal". When the reasoner FaCT++ is started, the ontology successfully categorizes the instance "Hydrogen" to belong to the class "nonmetal". This type of rule is important for ontology inferences, as the ontology will place elements into their respective groups according to their data, thereby organizing information based on rules rather than relying on manual input for each element. As with the instance "Hydrogen", all elements are stored within the ontology in a similar manner.

Object properties are also created in order to define how different classes relate to each other. Object properties are of particular interest because they define how classes and data properties can relate to each other. The periodic table ontology has a total of 10 object properties, which are defined based on the information found within the periodic table. Object properties define relations between classes by defining how a class interacts with another class, which can be seen in Figure 2b. For example, the object property "is_part_of_block" is defined where the class "name" shares a relation with the class "block". In

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

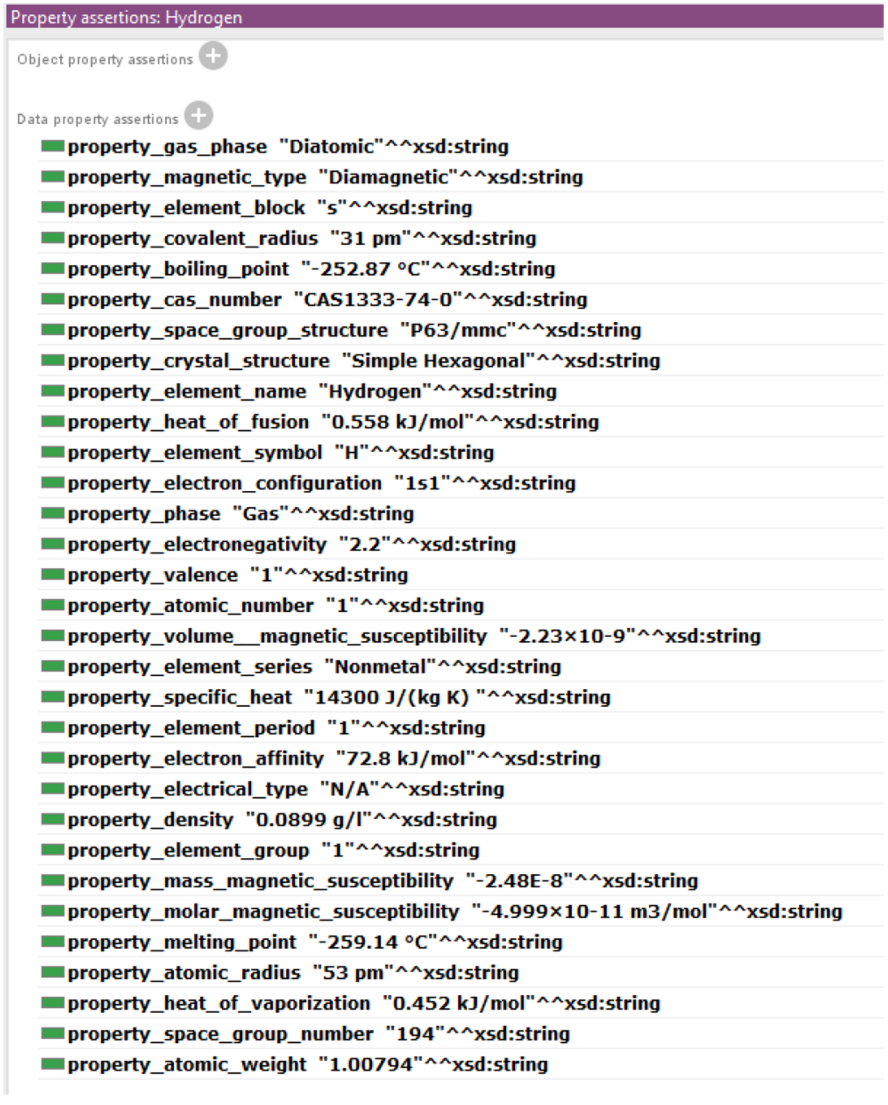


Figure 3. Data properties of the individual "Hydrogen".

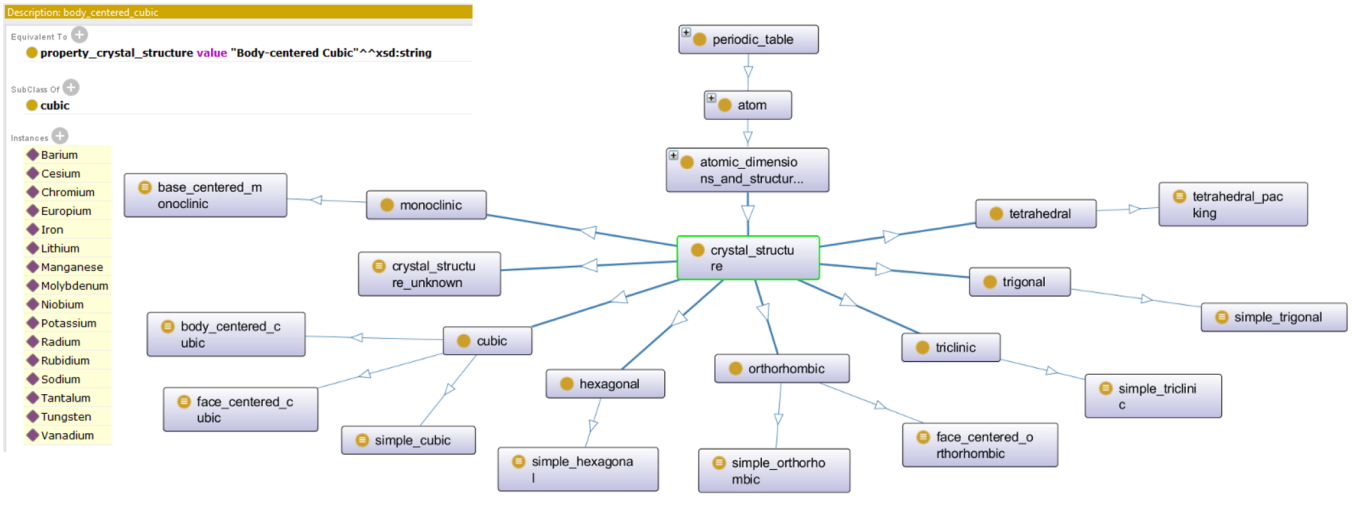


Figure 4. Subclass "crystal_structure" is expanded in order to show additional subclasses. An example subclass "body_centered_cubic" is shown to be defined as being equivalent to the rule "property_crystal_structure value" "Body−centered Cubic". The yellow box under "Instances" designates the instances that are inferred by the syntax of the ontology. The following instances of the periodic table are inferred to the "body_centered_cubic" subclass: Ba, Cs, Cr, Eu, Fe, Li, Mn, Mo, Nb, K, Ra, Rb, Na, Ta, W, and V.

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

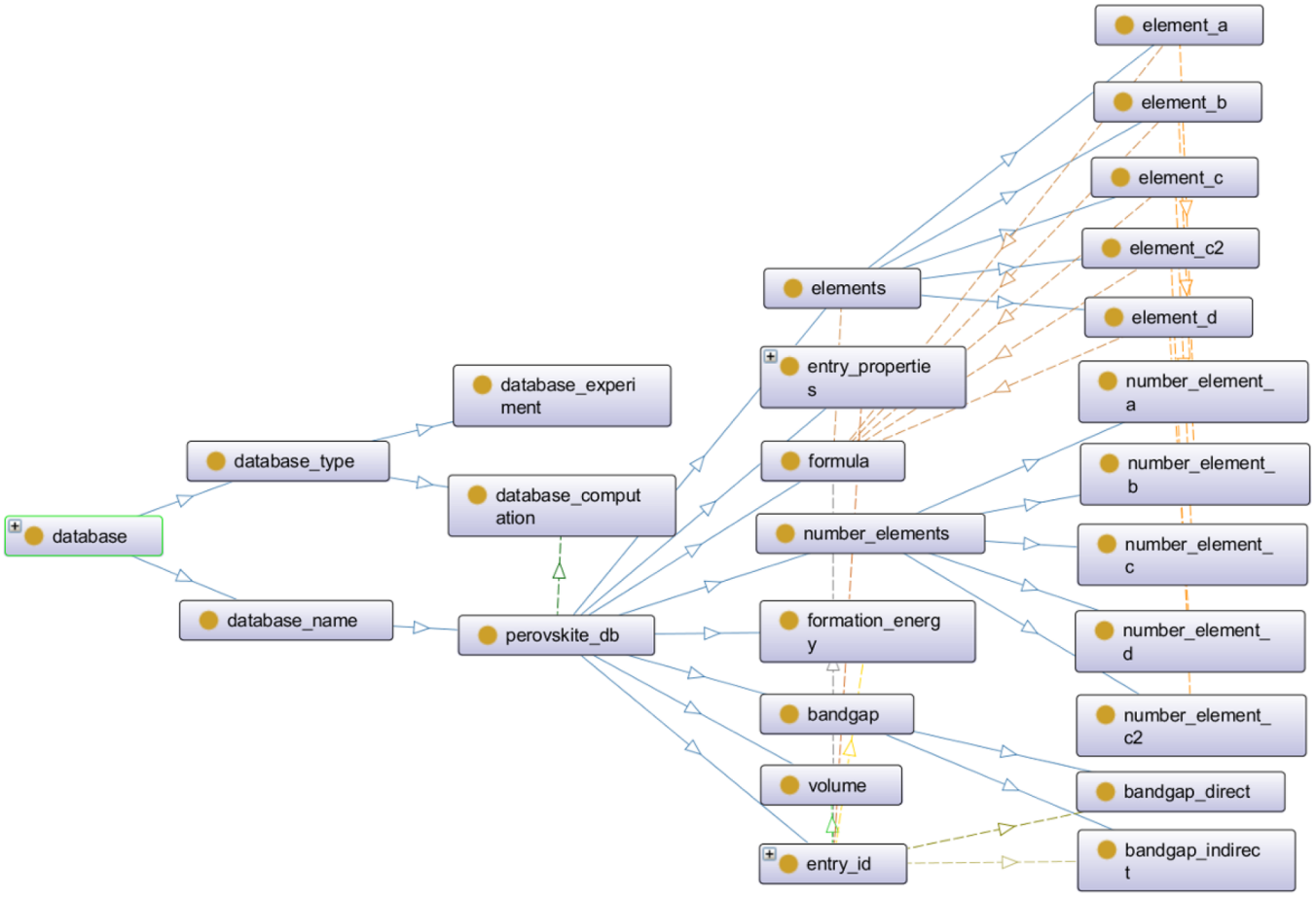


Figure 5. Class tree hierarchy of the ontology developed for a database focused on perovskites. Solid blue lines represent class hierarchy, while colored dotted lines represent the object properties defining the relationships between classes.

particular, it is defined where a member of the class "name" accesses the class "block" or of one of its subclasses when the object property is invoked. Object properties such as these help define relations that may not otherwise be defined within a published database.

Figure 4 explores how a property such as "crystal structure" can be organized and used by an ontology. Please note, however, that the class "crystal_structure" only refers to the crystal structures of the periodic elements. The class "crystal_structure" is composed of eight subclasses that represent the various types of crystal structures an atomic element could possess. This class is a subclass of the parent "atomic_dimensions_and_structure" and has several sibling classes that are similarly related. As can be seen in Figure 4, there are relationships shared between the various types of crystal structures that may not be easily understood by looking at the data alone. These classes are also defined to include the membership of atomic elements with the proper crystal structure value. This becomes useful, as can be seen by the inference results of the reasoner, as the elements are properly organized according to their proper crystal structure groups. This is very useful, as now one can see how the elements relate to each other according to their crystal structure. Through inferences, the possibility of uncovering new relationships between elements and other similar types of data arises and can help hasten the development of materials.

This ontology is designed to not only organize the properties related to the periodic table but also to be merged with ontologies of other databases for use. This method of organization not only helps organize the raw data into categories that are easier to access cross-database but also can connect data between ontologies that are separately written. This connection can be made because of the ability to define data relationships using object properties and descriptive logic that is made more accessible to the everyday user by how Protege is developed. This approach is further explored when additional ontologies are constructed and attempted to be merged.

TRANSLATING DATABASES AND MERGING ONTOLOGIES

In order to see how these ontologies can relate to ontologies developed from published databases, ontologies of two different types of databases are written and merged with the periodic table and property ontologies. In particular, two databases are used: a database consisting of computational data related to perovskites and a database consisting of experimental data related to oxidative coupling of methane (also known as OCM).[11,41]

An ontology for a computational database composed of data of various perovskite materials is first investigated. Perovskite materials possess bandgaps that are desirable for solar cell applications as they are found to be ideal when attempting to capture solar light. This database is therefore composed of information relating to these materials such as the elements of the material, volume, band gap, and formation energy. Figure 5 represents a basic class tree hierarchy for the ontology created for the perovskite database. In total, the developed ontology has a total of 1136 axioms with a class count of 76, object property

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

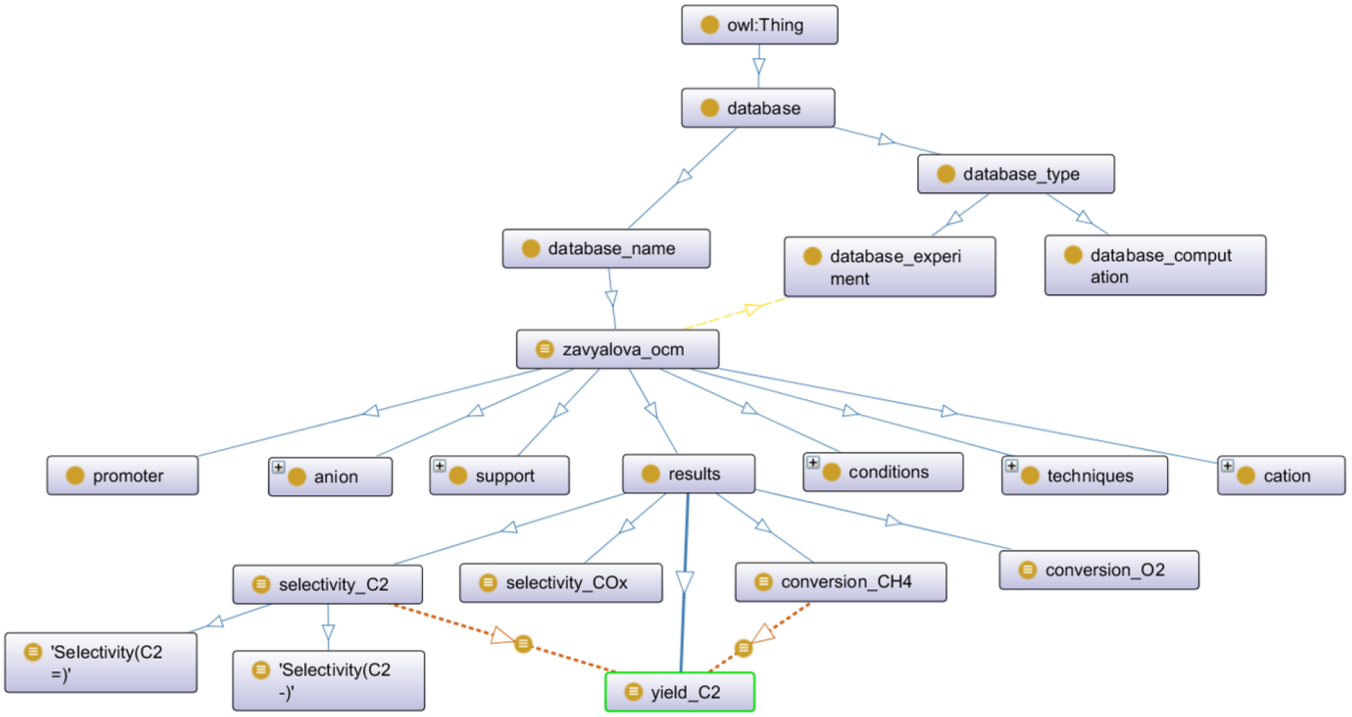


Figure 6. Partial class tree hierarchy of the ontology developed for a database centered on OCM reactions. Solid blue lines represent class hierarchy, while colored dotted lines represent the object properties defining the relationships between classes. In the case of "yield_C2", the orange dotted lines demonstrate that the classes "selectivity_C2" and "conversion_CH4" are directly related to "yield_C2".

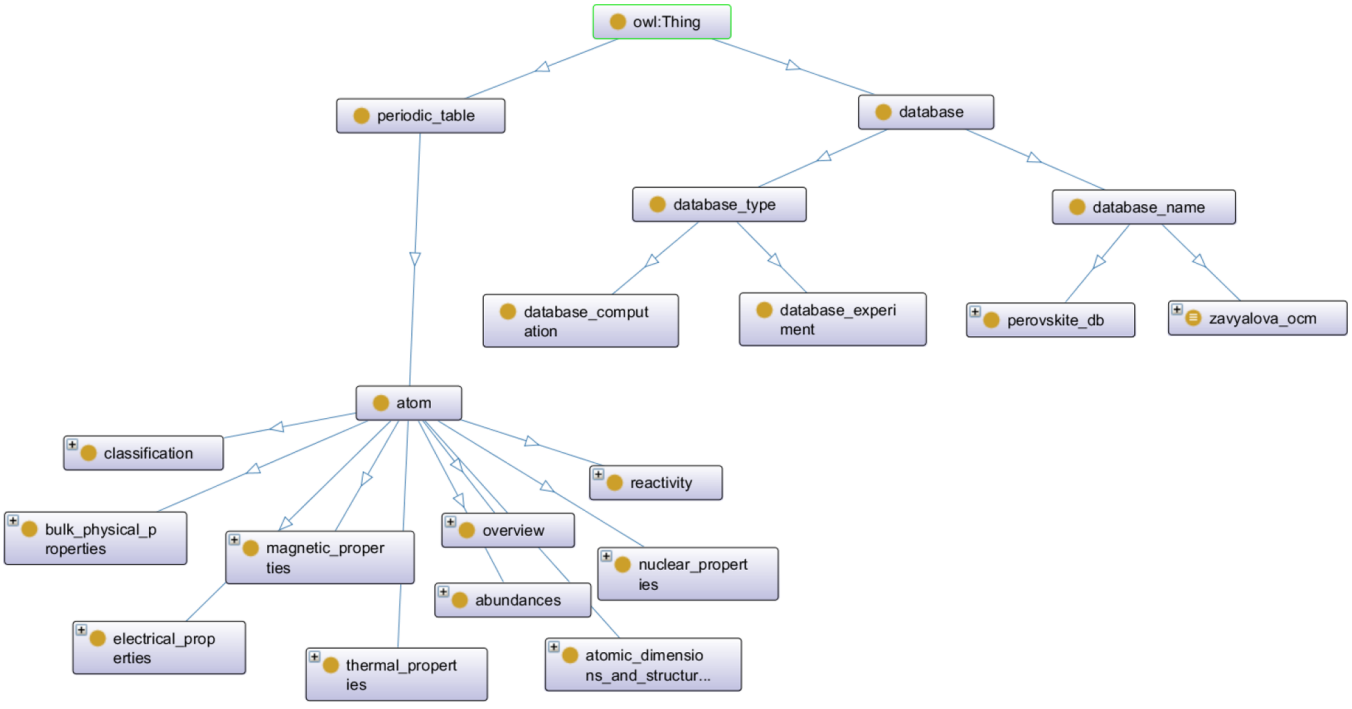


Figure 7. Partial class tree hierarchy of all three ontologies merged into a single ontology. Solid blue lines represent class hierarchy, while colored dotted lines represent the object properties defining the relationships between classes.

count of 13, and data property count of 21. Given the number of entries within the original database, 50 entries were imported as instances in order to test the ontology. Subclasses "bandgap", "elements", "entry_id", "entry_properties", "formation_energy", "formula", "number_elements", and "volume" are found underneath the parent class "perovskite_db" and represent the different types of information found within the database. As can be seen in Figure 5, several classes are related to each other. For instance, the class "number_elements" is related to "elements" in that the numerical information found within "number_elements" is only relevant to the appropriate elements listed under "elements" (e.g., "Number of Element A" is directly

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

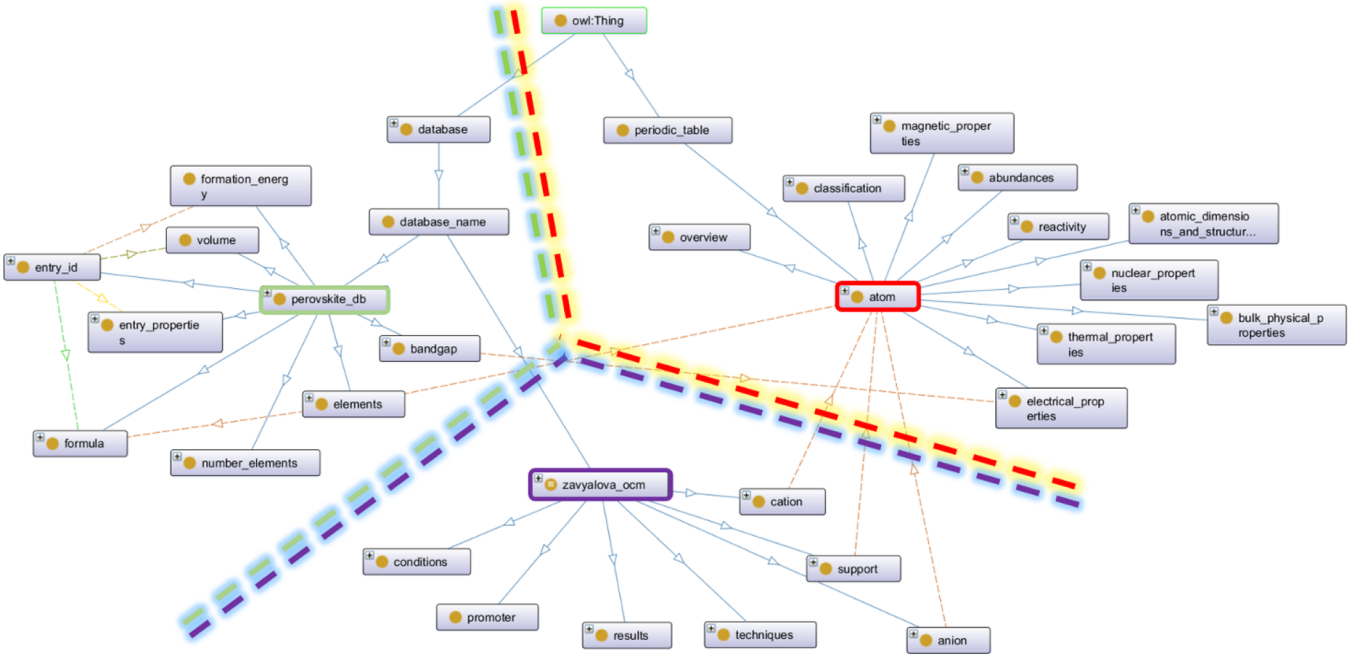


Figure 8. Alternative partial class tree hierarchy where the three different ontologies are blocked. Solid blue lines represent class hierarchy, while colored dotted lines represent the object properties defining the relationships between classes.

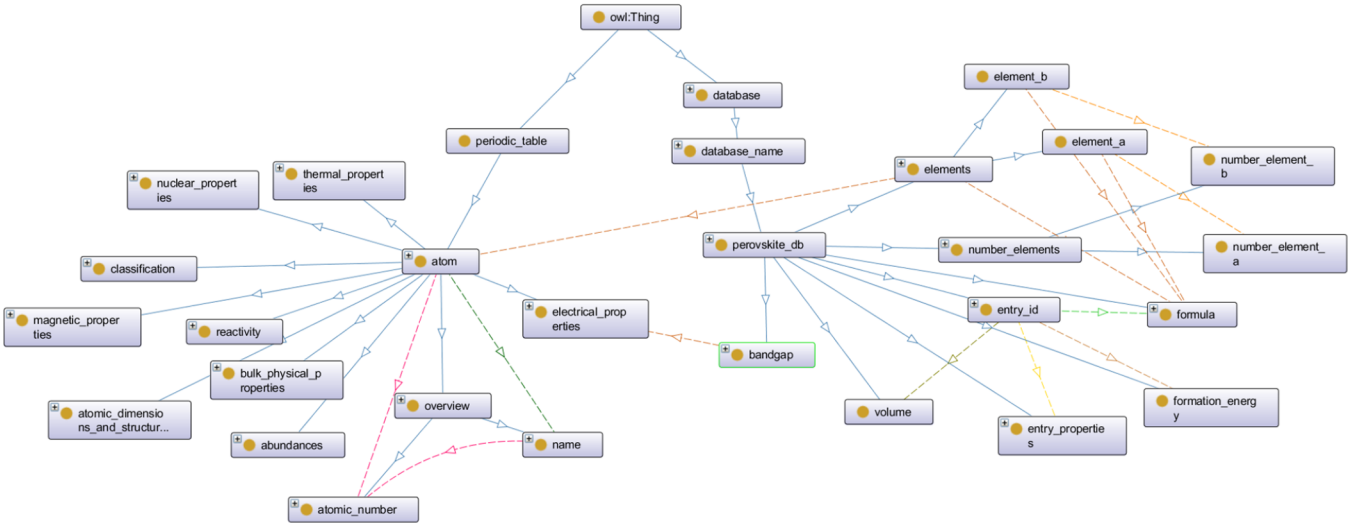


Figure 9. Example of how classes from the database ontology "perovskite_db" can access information from the subclass "atom" found under the class "periodic_table". In particular, relationships between "elements" and "atom" as well as "bandgap" and "electrical_properties" are shown. Solid blue lines represent class hierarchy, while colored dotted lines represent the object properties defining the relationships between classes.

related to "Element A"). Additionally, subclasses "elements" and "number_elements" are related to "formula" since "formula" is composed of different compositions of different elements. These relationships can be defined using object properties and are visualized as dashed lines in Figure 5, thereby demonstrating that the relationships between different pieces of information can be successfully defined for better understanding of the data.

An ontology for an experimental database composed of data related to catalysts used for the OCM reaction is also investigated. The OCM reaction garners interest in the catalyst community as it is used in attempts to convert methane (CH_4) to ethylene (C_2H_4) and ethane (C_2H_6). As such, this sort of database is therefore composed of data pertaining to the catalysts used, the experimental conditions involved, and the resulting yields and products of the experiment. Figure 6 represents a partial class tree hierarchy of the ontology developed for the OCM database. In total, the developed ontology has a total of 225 axioms, a class count of 52, an object property count of 5, and a data property count of 30. It also has an individual count of 9, which represent various techniques that are used in experiments. The ontology separates the data types into several categories; for example, all subclasses related to experimental conditions fall under the parent class "conditions". Anions, cations, and supports are broken down separately, where each possible anion, cation, and support have subclasses representing the element and percentage (mol) of the element found within the material. The benefits of using an ontology becomes apparent when attempting to define the relationship

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

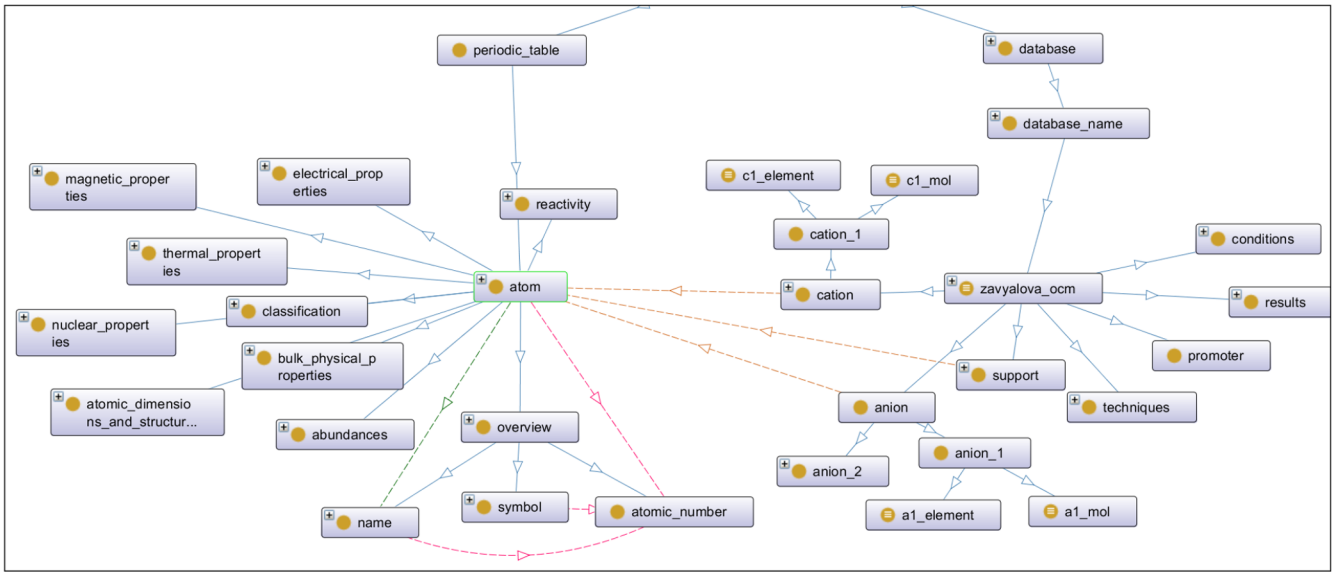


Figure 10. Closer look at how classes from the database ontology "zavylova_ocm" accesses information from the subclass "atom" from within the class "periodic_table". Solid blue lines represent class hierarchy, while colored dotted lines represent the object properties defining the relationships between classes.

between data pertaining to the experimental results. Results such as yield and selectivity are listed linearly within the database with no clear marker about how these results relate to each other. However, further investigation reveals that one cannot have "yield_C2" without also understanding the "selectivity_C2" and "conversion_CH4". Object properties allows for this relationship to be defined, making it much easier for other researchers and nonspecialists to understand the data that is being presented. Again, object properties allow for a more meaningful presentation of data and provide a more multidimensional way of viewing and interacting with data, thereby allowing others to enter into materials and catalyst informatics in an easier fashion.

The potential of ontology in terms of databases within the materials science and catalysis fields can be seen when all three ontologies are merged together into a single larger ontology. Figure 7 shows how the various ontologies are organized together. As can be seen in Figure 7, there are two overarching classes: "database" and "periodic_table". The ontologies are initially set up in this manner in order to better organize the types of data that the various ontologies incorporate when they are merged together. The class "database" is further broken down into subclasses "database_type" and "database_name", where the perovskite and OCM ontologies are listed under the class "database_name". The periodic table ontology is placed with the intention of being applied toward other ontologies. Figure 8 shows an alternative view of how the ontologies relate to each other once merged, with each ontology roughly blocked off from the other.

Various properties within the database ontologies are connected to classes within the "periodic_table" subclasses through the use of object classes. This is particularly useful in cases such as the perovskite ontology because the materials found within the original database are composed of different atomic elements. Figures 9 and 10 show how classes belonging to different ontologies can relate to each other. By creating the object property "content_refers_to", classes that are found within the database ontologies can be set as domains and subclasses within the "periodic_table" class can be set as the range. For example, in the case of Figure 9, the subclass "elements", which belongs to the perovskite ontology, is linked to the subclass "atom" within the periodic table ontology, leading to the rule "elements content_refers_to atom". This rule establishes the relationship between the elements listed within the class "elements" to the atomic elements listed within the class "atoms" thereby linking the two ontologies together. A similar relationship is established between the subclass "bandgap" found within the perovskite ontology and the subclass "electrical_properties" within the periodic table ontology, which are linked together by the object property "content_refers_to". In the case of the OCM ontology, which is depicted in Figure 10, subclasses within classes "cation", "anion", and "support" are linked to the class "atom" in the periodic table ontology, as these subclasses related by the periodic elements. By utilizing object properties, the data originally listed within their respective databases are now not only better understood in terms of how they relate to data within its own database but are also now better understood in regards to how it relates to information found outside of the original database (in this case, how the data found in the databases relate to and interact with the periodic table).

These results show the potential for not only understanding the relationships between different data within a particular database but also for defining and connecting data between databases in a smoother more coherent fashion. This is important for research within the materials science and catalysis communities as it allows for a more thorough understanding of what the data means. In addition, this concept also allows for the combination of data outside of its intended users; while the periodic table ontology was merged with ontologies developed for databases focused on catalysis and materials, it can be included with ontologies of other research fields that would require this sort of information. This helps address difficulties in interdisciplinary research by providing a system that allows for a more meaningful way of accessing data.

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

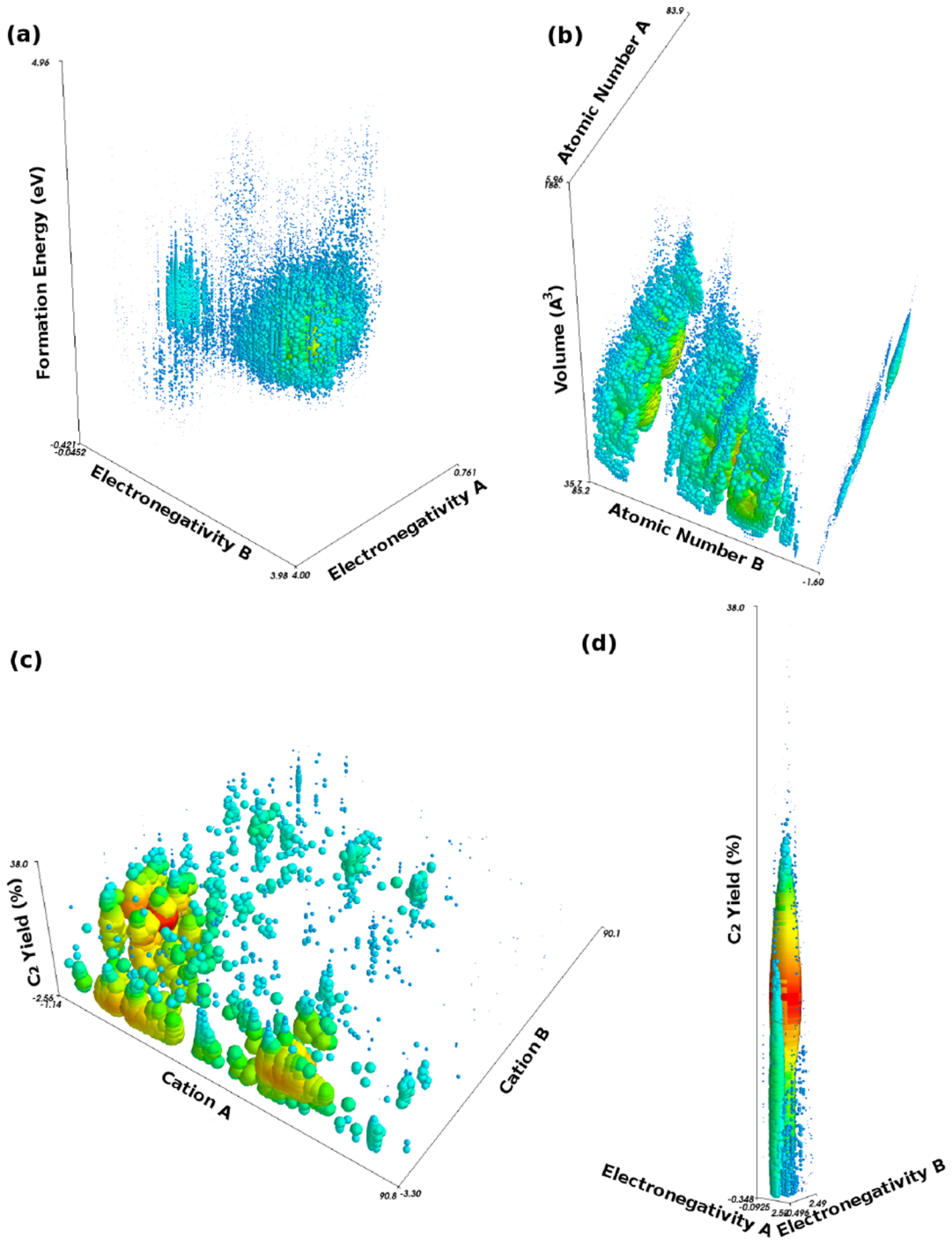


Figure 11. Heatmap of (a) electronegativity of A, electronegativity of B, and formation energy, (b) atomic number A, atomic number B, and volume of unit cell in Å^3 in perovskite solar cell materials, ABC_2D, database. Heatmap of (c) C_2 yield in %, cation A and cation B, (d) C_2 yield in %, electronegativity of A, and electronegativity of B in oxidative coupling of methane database. Note that A and B represent the atomic element in database. Colors indicate intensity of data (blue, cold; red, hot).

■ ADVANTAGES OF ONTOLOGY

There are several advantages in using ontologies during the data curation process. In particular, the following are strong advantages for the materials science and catalysis communities: establishment of semantic relationships between data, database integration, and the ability to define ambiguous data and include metadata within a database.

One particularly advantage of ontologies is that relationships between data can be defined, thereby giving meaning to the data. This feature is useful as it allows for smoother navigation of data and eases the process of transforming data into knowledge. Ontologies help expand our understanding of information by mapping out how different concepts relate to each other. Additionally, defining the relations between data allows users to query specific sets of data using query languages such as SPARQL. Time spent searching through data manually for

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

information decreases, while the ability to query an ontology allows nonspecialists to navigate information more efficiently, thereby expanding the accessibility of the data.

Another advantage is its ability to allow for database integration. Researchers are often left to collect databases and reconstruct them in order to fit their needs, leading to an increase in time and resources spent toward data interpretation and reconstruction. By using ontologies, databases can be used in tandem with other databases more smoothly. Additionally, ontologies can be used for databases that involve multimedia such as imagery, which often provide information that cannot be translated into an Excel file, for example. Therefore, ontologies allow for a more centralized and multidimensional database integration process.

Ontologies also enable the ability to define terminology, thereby reducing ambiguity. This ability is particularly important for applications in the materials science field because of the field's multidisciplinary nature. It is not uncommon for terms to be "recycled" and used throughout several disciplines; however, field-specific definitions are often lost as databases are shared, which can lead to misinterpretations and errors. Ontologies help this problem by allowing users to define relationships ("object properties" in Protege) as well as allow annotations to be defined with the appropriate properties, giving researchers the ability to leave comments, observations, and clarifications about ambiguous or controversial data. This helps reduce the ambiguity of data.

These factors can greatly impact the degree of time spent on searching data for research as well as the quality of research. By using ontologies, data accessibility becomes greater, data ambiguity is reduced, and data is enriched by the inclusion of multimedial data. This is particularly important for the materials science and catalysis fields, which are composed of many different fields with field-specific techniques, terminology, and requirements and expectations of data.

APPLYING ONTOLOGY TOWARD MATERIALS AND CATALYSTS DATABASES

The usefulness of an ontology is then explored by re-examining the perovskite and catalysts databases. In particular, the computational perovskite solar cell materials database and the experimental oxidative coupling of methane catalysts database are used.[11,41] Element and composition information in these databases are expanded based on the ontology. The effects of the developed ontology upon the databases are visualized using heatmaps as shown Figure 11.

The database composed of perovskite data is first revisited. A database consisting of 15,000 perovskite solar cell materials and constructed by density functional theory calculations is investigated.[41,42] The element information on the perovskite materials, ABC_2D, is expanded based on the developed ontology. Electronegativity of A and electronegativity of B against formation energy in eV is shown Figure 11(a). Figure 11(a) shows that the data is populated when the electronegativity of A and B is small. Similarly, atomic number A and atomic number B of perovskite materials ABC_2D are plotted against the volume of the perovskite materials and displayed as a heatmap as shown in Figure 11(b). There are several highly populated areas in Figure 11(b) depending on the atomic combination of A and B. Furthermore, the volume of perovskite materials differs from the atomic combinations of A and B. These results show that there is a hidden trend within the data that would have otherwise been missed due to the lack of information provided by the original database. Without the provision of an ontology with the database, other researchers that are not as experienced with this type of data may otherwise miss these relationships and have a more difficult time with research. These results show that using the developed ontology can help in understanding the data structure within the original materials database by providing information that was otherwise not included within the database.

The database compose of OCM reaction data is also revisited. A database consisting of 1866 experimental oxidative coupling of methane catalysts collected from various literature is revisited.[11,43] Within the 1866 catalysts, cation A and cation B are visualized against C_2 yield in oxidative coupling of the methane reaction as shown in Figure 11(c) where A and B represent the atomic element. Figure 11(c) demonstrates that low cation numbers of A and B are densily populated with a C_2 yield of approximately 20%. In other words, one can consider that researchers have been focusing on specific catalysts in experimental oxidative coupling of methane (OCM) database. This level of focus placed on one particular type of catalyst can impede catalyst research as not only other catalysts are not investigated thoroughly but factors that could affect all catalysts may be missed due to this sort of tunnel vision. Thus, additional data that would be considered related to the data concerning the compounds is also added. C_2 yield against electronegativity of A and B are also visualized as shown in Figure 11(d). Similar to Figure 11(c), the data is densily populated at specific electronegativities of A and B against the C_2 yield. This can hint toward further understanding of the catalytic activities listed from the database that would have otherwise been missed if only the originally published database was investigated. Thus, the data structure of the catalysts database can also be well understood from using the developed ontology.

The developed ontology can be an aid toward revealing descriptors for material and catalytic properties. By expanding existing databases with information collected within an ontology, new methods of examining data and new insights into collected data can be discovered, thereby enhancing the usefulness of materials and catalysts databases. While the developed ontologies are focused on the information found within particular databases, they can be integrated into new ontologies dealing with other types of catalysis and streamlined through the application of object properties. Inclusion of such information has been shown to positively affect research regarding experimental and computational data.[43] Ontologies can also be changed and updated to reflect new information and discoveries made experimentally and computationally. This allows the information stored within the ontology to remain up-to-date, which helps reduce the time and resources required to maintain databases that are already established. In whole, adopting ontologies within the materials and catalyst informatics fields can help aid in research centered on the use of materials and catalyst databases.

CONCLUSION

Overall, an ontology for the periodic table was developed where data from the periodic table and commonly used data related to the periodic elements are collected and imported into an ontology. The ontology consists of 4496 axioms, a class count of 170, an object property count of 9, and a data property count of 66 where the periodic elements are stored as 118 individuals that represent their respective elements and the raw data that belongs with those elements. Of the 170 classes, many are stand-in

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

classes for common properties that can be updated to reflect knowledge that appears from searching for descriptors of properties as well as other materials-related research. The ontology is also applied toward ontologies developed for two materials and catalyst databases, where the data structures are better understood with the inclusion of data from the ontology. Overall, this ontology offers an alternative, more globalized method of designing and creating materials databases, which can help address current issues of materials and catalyst database creation and maintenance in relation to materials and catalyst discovery as well as the descriptor search.

AUTHOR INFORMATION

Corresponding Author

*E-mail: TAKAHASHI.Keisuke@nims.go.jp.

ORCID

Lauren Takahashi: 0000-0001-9922-8889

Itsuki Miyazato: 0000-0002-1533-9790

Keisuke Takahashi: 0000-0002-9328-1694

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is funded by Japan Science and Technology Agency (JST) CREST Grant JPMJCR17P2, JSPS KAKENHI Grant-in-Aid for Young Scientists (B) Grant JP17K14803, and the Materials research by Information Integration (MI^2I) Initiative project of the Support Program for Starting Up Innovation Hub from JST. Additionally, this work was conducted using the Protege resource, which is supported by Grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

REFERENCES

(1) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.

(2) Nørskov, J. K.; Bligaard, T. The Catalyst Genome. *Angew. Chem., Int. Ed.* **2013**, *52*, 776−777.

(3) Walsh, A. Inorganic Materials: The Quest for New Functionality. *Nat. Chem.* **2015**, *7*, 274.

(4) Rajan, K. Materials Informatics: The Materials "Gene" and Big Data. *Annu. Rev. Mater. Res.* **2015**, *45*, 153−169.

(5) Takahashi, K.; Tanaka, Y. Materials Informatics: A Journey Towards Material Design and Synthesis. *Dalton Trans.* **2016**, *45*, 10497−10499.

(6) Curtarolo, S.; Hart, G. L.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway To Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191.

(7) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.

(8) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73.

(9) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations. *Nat. Commun.* **2017**, *8*, 14621.

(10) Pfeif, E. A.; Kroenlein, K. Perspective: Data Infrastructure for High Throughput Materials Discovery. *APL Mater.* **2016**, *4*, 053203.

(11) Zavyalova, U.; Holena, M.; Schlögl, R.; Baerns, M. Statistical Analysis of Past Catalytic Data On Oxidative Methane Coupling For New Insights Into The Composition of High-Performance Catalysts. *ChemCatChem* **2011**, *3*, 1935−1947.

(12) Breuer, H.-P.; Petruccione, F. *The Theory of Open Quantum Systems*; Oxford University Press on Demand, 2002.

(13) Xu, Y.; Yamazaki, M.; Villars, P. Inorganic Materials Database For Exploring the Nature of Material. *Jpn. J. Appl. Phys.* **2011**, *50*, 11RH02.

(14) Hummelshøj, J. S.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nørskov, J. K. CatApp: A Web Application For Surface Chemistry and Heterogeneous Catalysis. *Angew. Chem.* **2012**, *124*, 278−280.

(15) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *NPJ. Comput. Mater.* **2015**, *1*, 15010.

(16) Edwards, P. N.; Mayernik, M. S.; Batcheller, A. L.; Bowker, G. C.; Borgman, C. L. Science Friction: Data, Metadata, and Collaboration. *Soc. Stud. Sci.* **2011**, *41*, 667−690.

(17) Hosono, H.; Tanabe, K.; Takayama-Muromachi, E.; Kageyama, H.; Yamanaka, S.; Kumakura, H.; Nohara, M.; Hiramatsu, H.; Fujitsu, S. Exploration of New Superconductors and Functional Materials, and Fabrication of Superconducting Tapes and Wires of Iron Pnictides. *Sci. Technol. Adv. Mater.* **2015**, *16*, 033503.

(18) International Human Genome Sequencing Consortium. .Initial Sequencing and Analysis of the Human Genome. *Nature* **2001**, *409*, 860.

(19) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; et al. The Sequence of the Human Genome. *Science* **2001**, *291*, 1304−1351.

(20) Jena, P.; Castleman, A. W., Jr. *Nanoclusters: a Bridge Across Disciplines*; Elsevier, 2010; Vol. *1*.

(21) Gruber, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing? *Int. J. Hum. Comput. Stud.* **1995**, *43*, 907−928.

(22) Feilmayr, C.; Wöß, W. An Analysis of Ontologies and Their Success Factors for Application to Business. *Data Knowl. Eng.* **2016**, *101*, 1−23.

(23) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000**, *25*, 25.

(24) Takahashi, K.; Takahashi, L.; Baran, J. D.; Tanaka, Y. Descriptors for Predicting the Lattice Constant of Body Centered Cubic Crystal. *J. Chem. Phys.* **2017**, *146*, 204104.

(25) Takahashi, K.; Tanaka, Y. Role of Descriptors in Predicting the Dissolution Energy of Embedded Oxides and the Bulk Modulus of Oxide-Embedded Iron. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 014101.

(26) Miyazato, I.; Tanaka, Y.; Takahashi, K. Accelerating the Discovery of Hidden Two-Dimensional Magnets Using Machine Learning and First Principle Calculations. *J. Phys.: Condens. Matter* **2018**, *30*, 06LT01.

(27) Audi, G.; Bersillon, O.; Blachot, J.; Wapstra, A. The NUBASE Evaluation of Nuclear and Decay Properties. *Nucl. Phys. A* **2003**, *729*, 3−128.

(28) Arblaster, W. Densities of Osmium and Iridium. Recalculations Based Upon a Review of the Latest Crystallographic Data. *Platin. Met. Rev.* **1989**, *33*, 14.

(29) Barbalace, K. *Periodic Table of Elements: Aluminum - Al*, 1995. https://environmentalchemistry.com/yogi/periodic/Al.html (accessed August 2018).

(30) Cardarelli, F. *Materials Handbook: A Concise Desktop Reference*; Springer Science & Business Media, 2008.

(31) Coursey, J.; Schwab, D.; Tsai, J.; Dragoset, R. Atomic Weights and Isotopic Compositions (version 4.1). NIST, 2015. http://physics.nist.gov/Comp (accessed August 2018).

(32) Gray, T.; Mann, N.; Whitby, M. PeriodicTable.com. Element Collection, Inc,, 2017. http://periodictable.com/ (accessed August 2018).

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754

(33) Kelly, T.; Matos, G.; DiFrancesco, C.; Porter, K.; Berry, C.; Crane, M.; Goonan, T.; Sznopek, J. *Historical Statistics for Mineral and Material Commodities in the United States*; U.S. Geological Survey, 2005.
(34) Lide, D., Ed.; *CRC Handbook of Chemistry and Physics*, 87th ed.; CRC Press, 2006.
(35) Speigh, J., Ed.; *Lange's Handbook of Chemistry*; McGraw-Hill, 2004.
(36) *NIST Standard Reference Database Number 69*; NIST Chemistry WebBook; U.S. Department of Commerce, 2017. https://webbook.nist.gov/chemistry/ (accessed August 2018).
(37) *Periodic Table of Elements*, IUPAC, 2016. https://iupac.org/ (accessed August 2018).
(38) Musen, M. A. The Protégé Project: A Look Back and a Look Forward. *AI matters* **2015**, *1*, 4−12.
(39) Tsarkov, D.; Horrocks, I.*FaCT++ Description Logic Reasoner: System Description*. International Joint Conference on Automated Reasoning, 2006; pp 292−297.
(40) Sharif, B.; Maletic, J. *An Eye Tracking Study on Camelcase and Under_score Identifier Styles*. IEEE 18th International Conference on Program Comprehension, 2010; pp 196−205.
(41) Castelli, I. E.; Olsen, T.; Datta, S.; Landis, D. D.; Dahl, S.; Thygesen, K. S.; Jacobsen, K. W. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy Environ. Sci.* **2012**, *5*, 5814−5819.
(42) Takahashi, K.; Takahashi, L.; Miyazato, I.; Tanaka, Y. Searching for Hidden Perovskite Materials for Photovoltaic Systems by Combining Data Science and First Principle Calculations. *ACS Photonics* **2018**, *5*, 771.
(43) Takahashi, K.; Miyazato, I.; Nishimura, S.; Ohyama, J. Unveiling Hidden Catalysts for the Oxidative Coupling of Methane Based on Combining Machine Learning and Literature Data. *ChemCatChem* **2018**, na DOI: 10.1002/cctc.201800310.

DOI: 10.1021/acs.jcim.8b00165
J. Chem. Inf. Model. 2018, 58, 1742−1754